BODY and SOUL
MATHEMATICAL SIMULATION TECHNOLOGY
Draft

(To be downloaded and then opened to get links to work)

Johan Jansson and Claes Johnson 2010
with contributions by Kenneth Eriksson, Don Estep, Peter Hansbo and Johan Hoffman

Contents

To KTH Students Fall 2010

Yes, you can! (Zlatan Ibrahimovic)

Almost always, the creative dedicated minority has made the world better. (Martin Luther King, Jr.)

Do not worry about your problems with mathematics, I assure you mine are far greater. (Einstein)

There is nothing so easy but that it becomes difficult when you do it reluctantly. (Pythagoras)

Dear KTH Student:

The course you are now about to start gives you a chance to acquaint yourself with a *new mathematics education* motivated by the revolutionary changes of science and technology brought by the computer, IT, Internet and Google.

The Success of Google: How?

Google is founded on a mathematical search algorithm using the singular value decomposition SVD of a matrix. Is Google a success?

Did you get the message? If not, see The Anatomy of Search Engines and Latent Semantic Indexing.

FIGURE 1. Body and Soul.

New Mathematics Education: BodyandSoul

The new math education is called BodyandSoul with the following meaning:

- Soul: brains, thinking, analytical mathematics, programming,
- Body: number-crunching, computation by computer.

You are now in your second year of university engineering education and have not so far met much of the revolution, right? But now it comes! I hope you are ready! The starting point is:

- IT is based on math
- BodyandSoul is IT math.

The new math education gives you new skills and tools which open to a new role as student:

- Constructive!
- Do yourself!
- Instruct the computer!
- Model the World!
- Analyze!

- Understand!!
- Design-Invent!! Develop your own Apple Apps.

Why Should I Care about My Math Education?

The mathematics education you have met so far at KTH follows a tradition going back more than 100 years, in fact 300 years back to the Calculus and Linear Algebra of Newton, Leibniz and Descartes forming the basis of the scientific revolution. This is a mathematics without computer with simple computational tools such as tables, slide rule and mechanical calculator.

The computer is now changing the use of mathematics in science, engineering and society: Google founders Larry Page and Sergey Brin understood that basic tools of mathematics such as SVD could be used to construct a search engine...

Computational mathematics can thus be used to index and search information, but computational mathematics is also the tool for creating new information in the form of pictures, movies, sound, science, technology, medicin, entertainment, computer games, simulators,...,in short for simulation of real and imagined worlds.

After 300 years a new scientific revolution is now changing life and work, an information revolution based on computational mathematics, but the educational system is slow to react because tradition dominates.

How then to react as a student? There are two possibilities:

- Follow the tradition without asking questions.
- Think for yourself, look around, listen and ask questions:
- Why/What Mathematics for Engineers?

If you go for the second option, BodyandSoul can give you a platform to construct-simulate-understand-control-design as, for example:

- scientist,
- engineer,
- generalist with specialist competence in many fields,
- manager.

What Is the Role of Mathematics?

To get started we want to confront you a couple of questions: First about mathematics in general:

- What is the role of mathematics in science and engineering?
- What is the connection between mathematics and computer?
- What is the role of mathematics in the IT age?

What Did I Learn from Mathematics Education?

Next ask yourself about your experience from mathematics education:

- What is the role of mathematics in your education?
- What did I learn during 12 years of school math education?
- What did I learn during 1 year of KTH math education?
- Why is $2 \times 3 = 3 \times 2$?
- What is $\sqrt{2}$ and how is it computed?
- What is meant by saying that $f(t)$ is a *function* of t?
- What is the connection between *integral* and *derivative*?
- How is the *exponential function* $\exp(x) = e^x$ defined?
- How can you find the value $\exp(x)$ for a given x, with and without a computer or table of values?
- How is the *trigonometric function* $\sin(x)$ defined and how can it be computed?
- How are *Bessel functions* defined and computed?
- Why are $\exp(x)$ and $\sin(x)$ called *Elementary Functions EF*?
- What is the role of *differential equations* in science and technology?

Short concise answers, please!

Tradition vs Motivation

The objective of the above questions is to make you aware of the fact that engineering education is based on a tradition without the computer, which still dominates basic courses in mathematics, mechanics and physics and thereby sets the frame for the whole education. Tradition!

FIGURE 2. Important to be motivated! But motivated by what?

Do you dare to ask yourself if the education you meet really prepares you in a good way for a professional life in the IT-age? Is there really a need for a new math education? Yes or No?

If you answer Yes, then you are motivated to learn something new in this course, something useful which you can carry with you in your mind and in your computer as you go on to cope with the World and make it better. Then your are motivated to read the text, reflect about what you read, start to ask questions, look around and develop skills and understanding.

If you answer No, then I ask you to motivate: Is this because you have studied the question yourself or because someone has told you so? If it is not your own conviction after careful study, but an idea taken from somebody, for example a math teacher, ask that person to motivate the No, and check if it is convincing.

New Paradigm: Computational Mathematics

Mathematics has two forms:

- symbolic: formulas on paper: analytical
- constructive: computer follows instructions of computer program: computational

New Paradigm in Nutshell

- space coordinate x
- time coordinate t
- state of a system: function $u(t)$ (assuming only dependence on t)
- rate of change of state: time derivative $\dot{u} = \frac{\partial u}{\partial t}$
- connection between $\dot{u}(t)$ and $u(t)$: $\dot{u}(t) = f(u(t))$ with $f(u)$ given
- *Differential Equation DE*:
- $\dot{u}(t) = f(u(t))$ for $t > 0$ with $u(0)$ given *initial value*
- solve DE by *time stepping*:
- $u((n+1)dt) = u(ndt) + f(u(ndt))dt$, $n = 0, 1, 2, 3, ...,$ dt time step
- present state plus update $f(u(ndt)dt$ gives next state
- human gives f, computer does the job of time stepping.

Elementary Functions by Time Stepping

- $u(t) = \exp(t)$ solves $\dot{u}(t) = u(t)$ for $t > 0$, $u(0) = 1$
- time stepping $\exp(t + dt) \approx exp(t) + exp(t)dt = (1 + dt)\exp(t)$
- Elementary functions EF solve elementary DEs
- Values of EF computed by solving DE
- DE for $\sin(t)$?

Read! Reflect! Question! Look Around! Express!

Read the text:

- Preface
- Part I: Icarus and Daedalus!
- Part II: Newton's World of Mechanics
- Part IV: Leibniz' World of Calculus.

FIGURE 3. Read the Text!! Question the Text!!

Enter into

- Part III: World of Games.

Reflect on

- What is the essence?
- What is new?
- What is of interest to you?

Look around for connecting ideas:

- In your math books?
- In your physics books?
- On Internet?

Express

- Summary of most essential aspects!
- Questions?

Simplicity, Generality, Functionality

Computational mathematics combines

- simplicity of basic principles,

FIGURE 4. Think! And Model the World.

- generality of application,
- work of the mind on principles, planning, organization, goal, meaning,
- work of the computer for routine computation

into

- general purpose tool,
- with large variety of special applications,
- automation of mathematical modeling.

Remember that understanding of physical phenomena comes from mathematical modeling and understanding of the mathematical model using analytical mathematics (formulas).

Combining analytical and computational mathematics, you can fly:

- (simple) analytical math necessary for understanding
- analytical computation: tricky, difficult, special,
- digital computation: simple, effcient with computer as work horse,
- analytical computation: walk by foot from one village to another,
- computational math: helicopter anywhere.

For some perspective, take a look at

- Scientists and Science in Cartoons

FIGURE 5. The Cloud: What does it mean to you and your education?

- Did Einstein Not Understand Math?
- My Book of Knols

0.1 Formulate Your Goal: Find Motivation

I suggest that after reading this preface and browsing through Part I, you identify and write down your own motivation to engage in this course and what you expect to get out of it.

Why should you do this? Because it can be helpful to be motivated, and if your are not motivated, to understand why not. After the course you can then sum up and compare with your initial value.

It is about you and your education: If you like math, technology and computers, BodyandSoul can help you to develop your interest. If you are not a fan of traditional courses in mathematics (or mechanics and physics), BodyandSoul offers you an alternative new approach which you may appreciate better.

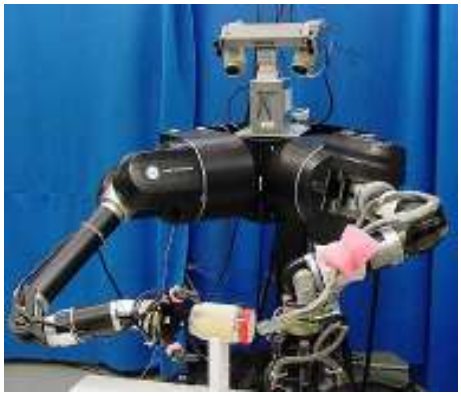


FIGURE 6. KTH Vision?

Please send your formulated motivation and expectation (short concise) by email to cgjoh@csc.kth.se. This can be very helpful for the further development of the program, and also in your own development.

0.2 KTH Vision? Your Vision?

Compare with KTH in the service of people, for the society of tomorrow:

- *Renewal and dynamics will be the key terms for the next few years.*
- *Knowledge is, and will always be, the most decisive of human beings assets and consequently must be managed and renewed with great care.*

Dynamical renewal for tomorrow is the leading principle of BodyandSoul. The statement *Knowledge must be renewed with great care*, is less clear: Does it express that KTH waits to renew until the new has become old? As a student at KTH you are part of the KTH vision and thus you may want to know what it is. So what is it? Anybody you can ask?

0.3 Discussion Forum

You are invited to take part in the BodyandSoul Discussion Forum, which you find on my blog, or by a link from my home page.

Please express your reactions to BodyandSoul, positive and negative! Your input is important to make it better! Your questions are welcome!

Stockholm October 24 2010

Claes Johnson

Preface

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

Perhaps it is better to be irresponsible and right, than to be responsible and wrong. (Winston Churchill)

You are not only responsible for what you say, but also for what you do not say. (Martin Luther)

Great bodies of people are never responsible for what they do. (Virginia Woolf)

Crisis without Responsibles

Nobody claims that todays mathematics education is functional. A steady flow of increasingly alarming reports tell that student achievements steadily decrease under the instruction by mathematics teachers with steadily decreasing competence. The decline seems to be bounded below only by zero, and the hope that the process can be reversed is vanishing. Google gives 70.000 hits on "crisis in mathematics education": 1, 2, 3, 4, 5, 6,...

But nobody is responsible for the resultless education: On each level the responsibility is shifted to the previous level. The reason that university mathematics fails is that highschool mathematics does not deliver the required prerequisites, and the reason highschool mathematics is a hopeless project, are missing skills from basic school mathematics, and so on to missed opportunities in the cradle (new research shows that

chicks can count to three...). In particular, mathematics professors at the top of the pyramid at the universities are not responsible for the dysfunctional mathematics teaching and training of mathematics teachers. And of course politicians cannot be held responsible because they have to rely on the high priests of the mathematics education church.

Stalemate at High Cost

Nobody is responsible, and nobody is allowed to take on the responsibility because that would upset the established order. It is a complete stalemate where no move is possible. But the stalemate is costly because lots of human and financial resources are lost. Mathematics eduaction is a big operation involving many students, many teachers and many hours. The total cost of a public school student per year may be 20.000 US dollars/year, out of which say 20 precent is spent on mathematics, thus 4000 dollars/year.

In Sweden this would amount to about 4 billion dollars/year or about 4 percent of the total state budget.

The Reason for the Crisis

To come out of the mathematics education crisis, it is necessary to understand the true reason for the crisis. Why did mathematics education work in the 1950s (more or less) but does not today (for sure). Yes, you are right, it is the computer!

The computer is changing our lives and the computer is changing mathematics in particular, because the computer is based on mathematics, on computational mathematics.

Traditional mathematics education teaches elementary arithmetics of integer and rational numbers used in solution of simple (linear and quadratic) algebraic equations and elementary analytic geometry.

But computers use computational mathematics or computer mathematics, which is not part of traditional mathematics education. There is thus a mismatch today between the use of mathematics in science, technology and society, and mathematics education, and this mismatch is the root of the crisis. No education can thrive under such conditions.

Bridging the Gap and Resolving the Crisis

It is thus necessary to reform mathematics education to make it conform with the use of mathematics in the booming computer age. The questions

and answers are new, because computational mathematical technology offers new capabilities. The computational technology is new, but the basic mathematics of Calculus and analytical geometry is much the same. Altogether we experience today a realization of the vision of Descartes and Leibniz of automation of mathematical modeling in the form of a Digital Calculus.

Computational mathematics allows mathematics education to expand the scope from simple algebraic equations to general differential equations, which can transform mathematics education from a meaningless excercise meaningful only to a few selected, into a meaningful activity meaningful to many.

From Tricks to Principles

The basic idea of information technology is to free the human spirit to creative work by letting the computer do tedious routine computation, such as searching for which the computer is both able and willing.

Instead of memorizing and practicing algorithms for long division of natural numbers in an endless number of meaningless exercises, students are encouraged to themselves write computer code for the algorithm of long division and then let the computer do the tedious routine work. Similarly, students may code the differential equations modeling the World and then let the computer solve the equations as a tool to discover the World.

A physical process is then seen as an analog computational chain process converting input to output according to certain physical laws, which is simulated by computation transforming digital input to output according to computer codes expressing the laws. This is the basic principle carrying the BodyandSoul program presented in this book.

Constructive Mathematics, God and $\sqrt{2}$

Mathematical modeling has two basic dual aspects: one symbolic-analytical and the other constructive-numerical, which reflect a 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 represented by the formalist/logicist school by Hilbert/Russell and the intuitionist school by Brouwer. A constructivist would argue that the real number $\sqrt{2}$ as the positive solution of the equation $x^2 = 2$ exists, because any given finite number of digits of its decimal expansion can be computed in a finite number of arithmetical operations, e.g. by Newton's method. A formalist/logicist could argue that $\sqrt{2}$ exists because non-existence would contradict the axiom of existence of a least upper bound of a set of real numbers bounded above. A constructivist does not believe that existence can come from contradiction, only from construction.

The formalist/logiscist proof of the existence of the real number $\sqrt{2}$ is similar to a proof of the existence of God from a contradiction of non-existence with Allmightyness, because Allmightyness must include existence (certainly Allmightyness cannot lack anything, in particular not existence). The constructvist would instead point to something marvellous like a new-born child, as a concrete proof of existence of something God-like.

A constructivist would say that finite decimal approximations of $\sqrt{2}$ of arbitrary finite precision can be computed and thus do exist, but that an infinite decimal expansion of infinite precision does not exist, at least not in the same sense. A formalist/logicist does not make this distinction.

This argument extends in BodyandSoul from the simple algebraic equation $x^2 = 2$ to general differential equations, for which approxinate solutions of finite precision can be (and are) computed and thus exist by construction, while the question of existence of an infinitely precise exact solution in any generality neither can nor has to be answered.

Split and Fusion

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 possibilities are opened by creating from the start a synthesis of constructive and symbolic mathematics representing a synthesis of BodyandSoul: 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 out of its present deep crisis. This is the purpose of the BodyandSoul program.

Proofs

Students often find matematical proofs difficult to appreciate, in particular non-constructive proofs, while step-by-step constructions, like the assembly of IKEA furniture according to a list of instructions, is understood by most people, even professors.

In BodyandSoul elementary functions, like the logarithm, exponential and trigonometric functions, are constructed by the students as solutions of elementary differential equations. Students are then encouraged to verify (prove) basic properties of elementary functions such as $\exp(a+b) = \exp(a)\exp(b)$, as consequences of the defining differential equations (with support by direct observation from computation).

Why Spend Time on Computer Games?

BodyandSoul students are encouraged to construct computer games based on mathematical models of real and imagined phenomena. The purpose is both to activate the student and to bring out the essential input-output aspect of mathematical modeling. The player of a computer game reacts to the output of the model, and gives feedback input to the model. To construct a game it is necessary to understand the input requirements of the model, and also what outputs or quantities of interest are relevant. Thus constructing computer games can be both entertaining and illuminating.

Summary of Main Features of BodyandSoul

- 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 software using Python.
- The student develops skills of mathematical modeling and programming by constructing computer games.
- 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 multi-physics problems.
- The program increases the motivation of the student by applying mathematical methods to interesting and important concrete problems already from the start. Emphasis may be put on problem solving, project work and presentation.
- The program gives theoretical and computational tools and builds confidence.
- The program contains the essential 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 Lipschitz continuity, differentiability, integration, constructive method for solving algebraic/differential equations, together with an ability to utilize these tools in advanced applications.
- The program may be run at different levels of ambition concerning both mathematical analysis and computation, while keeping a common basic core.

Acknowledgement

The BodyandSoul program has been developed over a long period in a sequence of books

1. Computational Solution of Partial Differential Equations by the Finite Element Method, 1987,
2. Computational Differential Equations, coauthored with Don Estep and Peter Hansbo, 1995.
3. BodyandSoul: Applied Mathematics Vol I-III, coauthored with Kenneth Eriksson and Don Estep, 2003,
4. Computational Turbulent Incompressible Flow, coauthored with Johan Hoffman, 2008.

Parts of the material of these books is reused in the present new e-version. I thank my coauthors for most constructive cooperation over the years. Without them the BodyandSoul program would not have been born.

Stockholm in November 2010

Claes Johnson

Ron Thomas/Giles Communication, via Associated Press

Elmo and his "Sesame Street" co-star Big Bird will be among those trying to show students the value of science and math.

FIGURE 7. Obama's math surge.

PS: Obama and Math

NYT reports in White House Begins Campaign to Promote Math Education (November 23 2009):

- *To improve science and mathematics education for American children, the White House is recruiting Elmo and Big Bird, video game programmers and thousands of scientists.*
- *President Obama renewed a commitment that would move the United States from the middle to the top of the pack in science and math over the next decade.*

Obama cannot understand the reason of the mathematics (education) crisis, because Obama is a lawyer, and nor can Elmo, but the idea to recruit video game programmers is not stupid...and we have adopted it.as you will see...

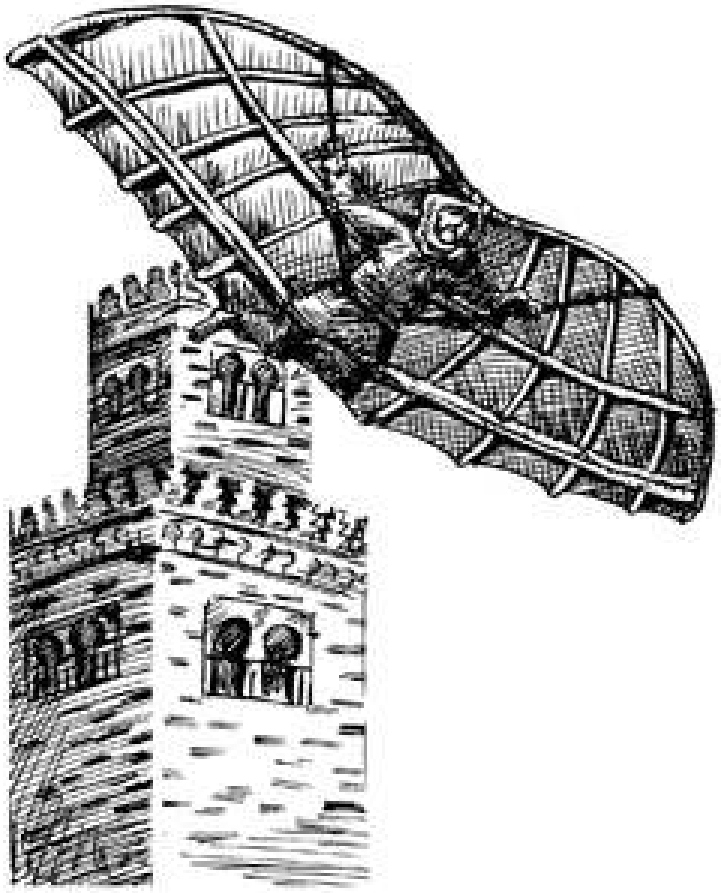


FIGURE 8. Abbas Ibn Firnas taking off from the Mosque tower in Cordoba in year 852. Are you ready for take-off?
.

Part I

Icarus and Daedalus

FIGURE 9. Icarus taking off. And you?

1
Start

> Education is an admirable thing, but is is well to remember from time to time that nothing that is worth knowing can be taught. (Oscar Wilde)
>
> By denying scientific principles, one may maintain any paradox. (Galileo)

1.1 An Important Decision

Congratulations: You have decided to follow a Bachelors/Masters/PhD program in mathematics, computer science, science, engineering, economics or some other field using mathematics. This is a very important decision which will form your coming professional life as a scientist, engineer, innovator, teacher, writer, scholar, businessman, manager, administrator...

The role of education is to give you a platform and mental tool-bag allowing you to access and analyze new information and solve problems, in short to make the World better by better understanding the World, by using mathematics. Remember that "understanding" in science and technology effectively means "mathematical understanding" in the sense that different relations are expressed in quantitative form in terms of mathematical formulas and numbers.

FIGURE 1.1. To discover a principle is a great thing. There are many out there yet to be discovered.

1.2 Welcome to BodyandSoul

Welcome to the *BodyandSoul Mathematical Simulation Technology* program. The basic idea is to combine the brain/soul with the computational power of the computer, with the objective of simulating real and imagined phenomena.

Symbolic mathematics (formulas) is soul and computational mathematics (digital computation by computer) is body, and together they form a very powerful tool for understanding and controling the real World, and to construct virtual worlds.

The World is complex, but seemingly governed by simple principles expressed in formulas such as Newton's 2nd Law $F = ma$ with F force, m mass and a accelleration. In principle, everything there is can be seen as material particles interacting by forces. Newton's 2nd Law connects particle motion to force by expressing that particle acceleration is proportional to force, and from accelleration you get velocity and from velocity you get position, by summation over many time steps, and from position you get force. And so the World goes around from one time step to the next...

So by letting the computer do the summation, you can simulate particles interacting by forces, that is the World, if you can formulate the basic principles as formulas. Elementary but profound! According to Leibniz principle

$$\delta \int_{t_1}^{t_2} L(q_i, \dot{q}_i) dt = 0$$

FIGURE 1.2. The Principle of Least Action describing all of mechanics.

of the Best World as the most complex world governed by the most simple principles!

I hope you are now ready to explore the real and virtual World. Remember that

- Soul is principle formulated as formula.
- Body is digital computation according to the formula.

Let's now start our work!

1.3 The Secret Behind the Surface

Here is how it works: Behind the skin there are particles connected by elastic springs interacting by Newtons 2nd Law:

- Flying Circus Cow: Exterior
- Flying Circus Cow: Interior
- Human Heart Simulator
- What It's All About 1 and 2.

1.4 Simulators

Simulation technology is developing exponentially in medicine, research, technology, sports and entertainment, and will revolutionize education. Here are some glimpses:

- Golf Tennis
- Boeing 757 Take-Off
- Heart Surgery
- Fluids Siggraph 2009 Preview

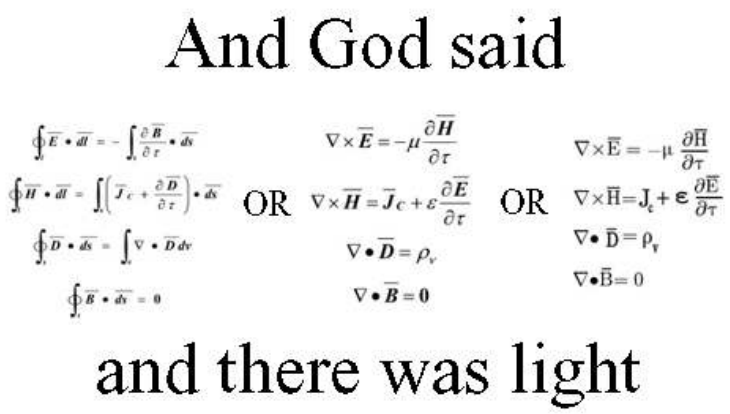


FIGURE 1.3. Maxwell's equations describing all of electromagnetics.

FIGURE 1.4. Taylor's Principle: The principal object of management should be to secure the maximum prosperity for the employer, coupled with the maximum prosperity for each employé.

2
From Pythagoras to Google

> Google In Quotes allows you to find quotes from stories linked to from Google News. Google News compiles these quotations from online news stories and sorts them into browsable groups based on who is being quoted.
>
> It is unworthy of excellent men to lose hours like slaves in the labor of calculation which could be relegated to anyone else if machines were used.(Gottfried Wilhelm von Leibniz (1646-1716))

2.1 The Singularity

Listen to Ray Kurzweil describing what is now happening as we are approaching the Singularity. This gives you a perspective on your studies, from the very start.

2.2 The Digital World

You are lucky to have been born into the computer age offering you a new digital world through the Internet. It is a wonderful world of words, texts, books, sound, music, images, films, computer games, software and software tools. In this program you will learn skills and which can help you

FIGURE 2.1. Surfing on the IT-wave.

to understand, organize, control to some degree and enjoy, both the real world and digital virtual world.

In the digital world everything is represented by numbers, strings of numbers, in the computer strings of the two digits 0 and 1. Information Technology or IT concerns processing digitized information and the processing means performing arithmetic operations on numbers including addition, subtraction, multiplication and division, and sorting and searching.

The digital world is built from text, pictures, sound and videos in formats such as

- pdf
- jpg, eps
- avi, mpg, mp4
- mp3

used in

- texts, books
- pictures, movies
- computer games
- Google
- Youtube.

FIGURE 2.2. Alan Turing: Creator of the Turing Machine = Computer.

The new IT-society is a *read-write-execute* society as a development of the traditional *read society*, where we can develop from passive consumers to active consumers-producers in the blogosphere and media such as

- Facebook
- Myspace
- Twitter.

IT offers new tools to humanity such as

- mobile telephone
- gps
- tomography, scanner, ultrasound, medical imaging
- robotics, control
- synthetic speech
- weather forecast

all based on computational mathematics.

The real world can be viewed as a form of *analog computation* with physical objets interacting by certain forces, as the World evolves from one time instant to the next. For example, the planets in our Solar System move according to Newton's 2nd Law $F = Ma$ connecting force F to mass M and accelleration a, combined with Newton's law connecting the force F to position. Or the atoms in the air you are breathing, moving around according to certain forces of attraction and repulsion.

The real world can be simulated by representing physical objects by numbers and replacing the analog computation by *digital computation*. In short:

- If you can compute with numbers, then you can simulate the world.

2.3 From Formal to Real Knowledge

Education is now changing from formalities to realities: Yesterday your employer would not ask what specifics you learned in school, as long as you could show evidence that you did well; education offered formal competence, which was useful for the individual in getting and keeping a position in business life. In the IT society of today, you need more than formalities; some knowledge of IT is necessary and expert knowledge can be very useful.

As individuals we can now set up our own

- blog: Blogger
- photo studio: Photoshop
- musical studio: Garageband
- movie studio:Movie Studio 9
- journal: NEWSMILL, MyMill
- publishing house: Amazon-Createspace
- TV-program: Youtube
- chair of expertize: Google-Knol, My-Book-of-Knols

The possibilities are endless: We are free to pick from the trees, eat, enjoy, and plant some new trees.. Spotify gives youu access to a World of Music, and Google will soon give you access to all books of the World..Let's go!.

2.4 Needed: Geometry and Calculus

To describe the World using mathematics you will need two basic tools

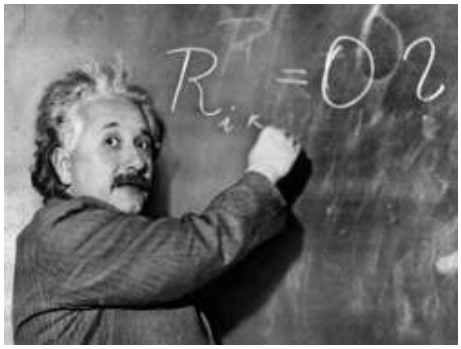


FIGURE 2.3. Perspective on Mathematics?

- analytic geometry - linear algebra
- Calculus: function, derivate, integral

which are the tools of scientific revolution initiated in the 17th century by Newton and Leibniz, leading into the industrial society and our modern information society, see BS What is Mathematics?

2.5 Analytic vs Computational Mathematics

The computer is today changing society, science and education, and mathematics. Mathematics is the science closest to the computer, and thus the revolution of the computer opens mathematics to new questions and answers. This is what is now going in science, engineering, society and education, but as all revolutions it is battle between the tradition and the new, a battle between analytical mathematics performed with symbols on paper and computational mathematics performed by computers with numbers.

But the science and education of mathematics is a rigidly built church which is not easily changed by a novelty such as the computer, which upsets the whole belief system and litania.

Analytical mathematics has its heros, the Field Medal Winners, such as Terence Tao. The results of todyas edge of professional analytical mathematics is closed to evaluation, because it is only understood by a small group of experts. The general public can only understand *that* something remarkable has been done, not *what* has been done.

Computational mathematics produces simulations of the World which can be evaluated/appreciated by a general public. Just like the performance of a piano virtuoso can be evaluated/appreciated by a large audience.

FIGURE 2.4. Raphael's perspective on Plato's School of Athens.

2.6 Perspectives

As you follow this course you will be encouraged to develop your own universe of mathematical simulation technology or Simulacra. It is always useful to have some perspective on what you are doing, because it helps you to find a direction forward. The author outlines different perspectives in My Book of Knols including

- Scientists and Science in Cartoons
- Hyperreality in Physics
- What is Science?
- Simulation Technology
- Is the World a Computation?
- Why/What Mathematics for Engineers?
- Modern Mathematics Education
- Is God Mathematician?
- Mathematics = Magics?

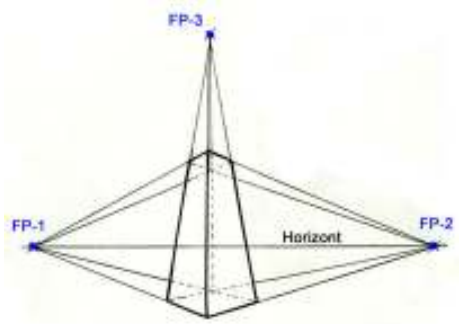


FIGURE 2.5. Three point perspective.

FIGURE 2.6. Perspective machine by Albrecht Dürer

- Did Einstein Not Understand Mathematics?.

with a parallel discussion on the blog

- Claes Johnson on Mathematics and Science.

See also

- Computational Technology Laboratory

leading the Simulation Technology program at KTH.

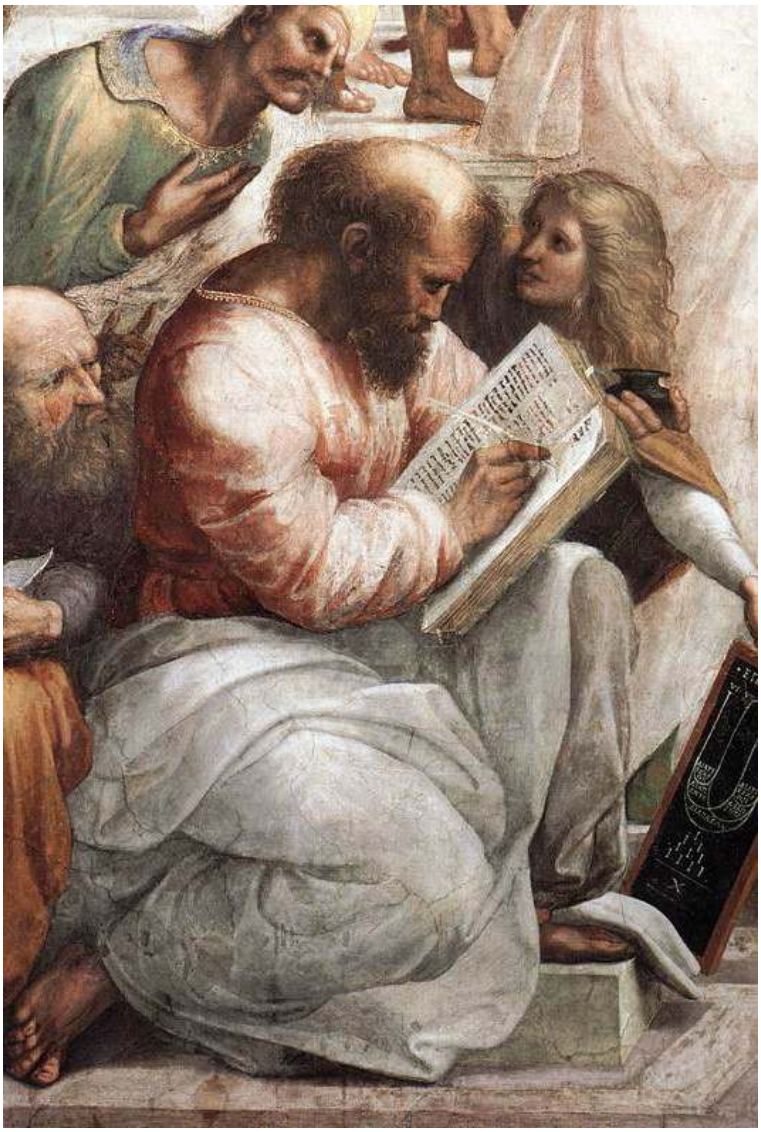


FIGURE 2.7. Pythagoras proving his theorem in Raphael's School of Athens

3
About BodyandSoul and Your Studies

In all things of nature there is something of the marvelous. (Aristotle)

It is simplicity that makes the uneducated more effective than the educated when addressing popular audiences. (Aristotle)

You cannot teach a man anything; you can only help him find it within himself. (Galileo Galilei (1564-1642))

There is no adequate defense, except stupidity, against the impact of a new idea. (Percy Williams Bridgman (1882-1961) Nobel Prize in Physics, 1946)

It requires a very unusual mind to undertake the analysis of the obvious. (Alfred North Whitehead (1861-1947))

3.1 What is BodyandSoul?

BodyandSoul is a new mathematics education program based on combining the power of the human soul with the power of body in the form of the computer as computational workhorse. In short: mathematics boosted by computer.

The computer is now changing society, science and education, and since the computer is based on mathematics, it also changes the nature of mathematics, from analytical mathematics based on symbolic computation with

FIGURE 3.1. An ATV computational mathematician able to go from any point A to any point B.

pen and paper to computational mathematics controlled by human brains and performed by computers.

Symbolic mathematics represents soul and computation represents body, with computational mathematics a synthesis of symbolic mathematics and number crunching computation, as a synthesis of body and soul.

The objective of mathematics and science is to simulate real or virtual worlds, to reach understanding, prediction and control. Computational mathematics boosted by computers allows simulation of complex phenomena, such as turbulence, which is is impossible by symbolic analytical matematics alone.

With computational mathematics almost anything thinkable is possible, more or less, while with analytical mathematics almost everything is impossible or very difficult and tricky. This is because computational mathematics is like an All Terrain Vehicle ATV allowing you get from A to B regardless of any paved road or track, while analytical mathematics seeks to find an elegant dirtless shortcut from A to B, which may not exist at all or is very difficult to find.

As you follow the BodyandSoul program you will discover that computational mathematics is based on a quite small set of principles with an amazing range of applicability. It is like having a small set of moral principles to guide you through the complexity of life, with body and soul in constructive cooperation. You will meet these principles over and over again and with each encounter understand more of their potential.

FIGURE 3.2. A classical analytical mathematician able to balance from a point A to point B, on a preset wire.

3.2 Technology For/With Simulation

You will find that Simulation Technology can be interpreted as

- Technology **For** Simulation
- Technology **With** Simulation,

in the following sense:

- **For**: How to make simulations using mathematics and computer.
- **With**: How to find out things about the World using simulations.

3.3 A Game About Constructing Games

The BodyandSoul program contains texts and software and supporting educational material, all integrated on a webbased platform. The entirely new possibilities in teaching and learning which are now opened by Internet, can broadly be described as new forms of interactive simulation with the student playing different computer games with the teacher/teaching material. It is like an arcade game with the objective of acquiring skills and tools to master (and win) the game.

Mathematical Simulation Technology can be described as the art of constructing computer games, and the BodyandSoul educational program can

itself be viewed as a form of interactive computer game, with the objcetive of learning how to construct computer games. BS as a game about games, a *game about constructing games*.

3.4 About the Text

Part I-XI gives an introduction to Mathematical Simulation Technology, with the core being Part IV Leibniz' World of Mathematics and Part V Descartes World of Analytical Geometry. The material is based on the earlier books

- BS Applied Mathematics Vol I-III
- Computational Differential Equations
- BS Applied Math Vol IV: Computational Turbulent Incompressible Flow

and connect to the following new books which can be used as source of inspiration for explorations in different directions:

- Computational Thermodynamics
- The Clock and the Arrow: A Brief Theory of Time
- The (Mathematical) Secret of Flight
- The (Mathematical) Secret of Sailing
- Many-Minds Relativity and Quantum Mechanics.

3.5 Layout

The text is organized as follows:

- I: Introduction: Cover story of Icarus and Daedalus.
- II: Newton's World of Mechanics: Newton's 2nd Law.
- III: World of Games.
- IV: Leibniz' World of Mathematics: Calculus and Linear Algebra.
- V: Descartes' World of Analytical Geometry.
- VI: Tool Bags: Summary of Calculus and Linear Algebra.
- VII: Sessions: Road Maps to Mastery.

- VIII: World of Differential Equations: Modeling
- IX: World of Finite Elements: Solving Differential Equations.
- X: Simulators: Vehicles to Drive.
- XI: Technology With Simulation: Understanding, Predicting, Controling.
- XII: 1D Calculus.
- XIII: MultiD Calculus.
- XIV: Complex Calculus.

Parts I-XI can be seen as the main parts with Parts XII-XIV as a form of supplement. After browsing the introductory Parts I-II you are invited to the World of Games giving you a direct experience of both mathematics, finite elements and programming, which prepare for a more detailed study of Technology For Simulation in IV-X leading into Technology With Simulation in Part XI. Part XII-XIV gives a more detailed presentation of various aspects of Calculus from 1D to MultiD to Complex Calculus including Fourier Analysis.

3.6 Combine Parts III, IV-V and VII

It is a good idea to do Parts III, IV-V and VII in parallel. Part III gives you computational experience of the basic concepts of geometry, position, velocity, accelleration, derivative with respect to time/space, physical laws as differential/algebraic equations, which you can connect to physical experience.

The mathematics of Part IV-V analyzes these concepts using both a microscope to see details and and a telescope to see gross patterns. Mathematics is partly a play with symbols, and to play the game it is useful to understand the meaning of the symbols. The other aspect of mathematics is number crunching, and to play this game it is useful to understand computational algorithms and to be able to implement them in computer programs.

The Sessions in Part VII help you to learn to master the tools summarized in Tool Bags in VI.

3.7 Zapping through BodyandSoul

The IT world and the Internet is based on zapping from one thing to another, by hyperlinks in texts or using search engines like Google. To read

a massive book from first to last page line after line, is practiced by a decreasing number of classically schooled people. Instead an eclectic mode of learning is encouraged, where you quickly can access information about anything from the nearest pizza place to global warming or digital photo, or Calculus and mathematical modeling.

BodyandSoul as an ebook invites the student to zapping, following links in various directions, and the text is not really intended to be read page after page. BodyandSoul can be then seen as a set of tools and material stored on shelves of a bike shop, which the student is invited to use to construct new bikes for enjoyment, learning, research or commercial use.

3.8 Constructive Mathematics as Turing Machines

You will find that constructive/computational mathematics can be expressed in computer codes. One can argue that the essence of the mathematics then is represented by the code, in the spirit of Turing and his Turing Machine or universal computer.

In constructive mathematics only constructed (or constructible) mathematical objects exist, in the form of Turing machines. In constructive mathematics the set of all real numbers does not exist (not even the set of all natural numbers), only specific numbers brought to existence by a Turing machine.

To insist that the set of real numbers does not exist may be shocking to a mathematician brought up in the ruling paradigm of the logistic/formalist school based on the dogma that it does exist. You can test the strength of a constructive approach, or weakness of a formalistic/logistic approach by questioning the dogma. You find material for debate in Do Mathematicians Quarrel?

Following BodyandSoul you will yourself construct mathematics from scratch by writing the computer codes representing mathematics, and you will thus become the master of your own mathematical bike shop.

3.9 BodyandSoul: Games

Through the following sequence of games you will be introduced to Calculus and Linear Algebra and to the construction of simulators and simulation tools:

1. 1d 2d 3d Pong (motion without forces)
2. 1d 2d 3d Ping-Pong (motion with forces: gravity and friction)
3. 1d 2d 3d Elastic Pong (elastic ball)

4. Elastic String/Membrane/Body
5. Elastic String/Membrane (transversal motion)

which prepare to design games of, for example,

- Tennis
- Table-tennis
- Golf, Minigolf
- Volleyball...

With this preparation you are then ready to simulate just about anything...In particular, remember that science is a game with the objective of simulating natural phenomena...in words, symbols, images, numbers, by mathematics plus computer...

3.10 BodyandSoul: Sessions

The BodyandSoul Sessions helps you to get in direct interactive contact with the material by connecting the mathematical theory from start to computational simulation. Going through the sessions you develop understanding of the mathematical principles and you also acquire skill of programming and implemention of mathematical algorithms. Each Session is self explantory, with references to the text, and treats a specific central topic. Sessions A gives an introcudtion to programming. Sessions B-D covers the basics of Calculus and Linear Algbra. Sessions E-F concerns mathematics and programjing of FEM.

3.11 BodyandSoul: Simulators

With the tools and skills you will acquire, you will be able to construct your simulators for a large variety of phenomena. BodyandSoul offers prototypes

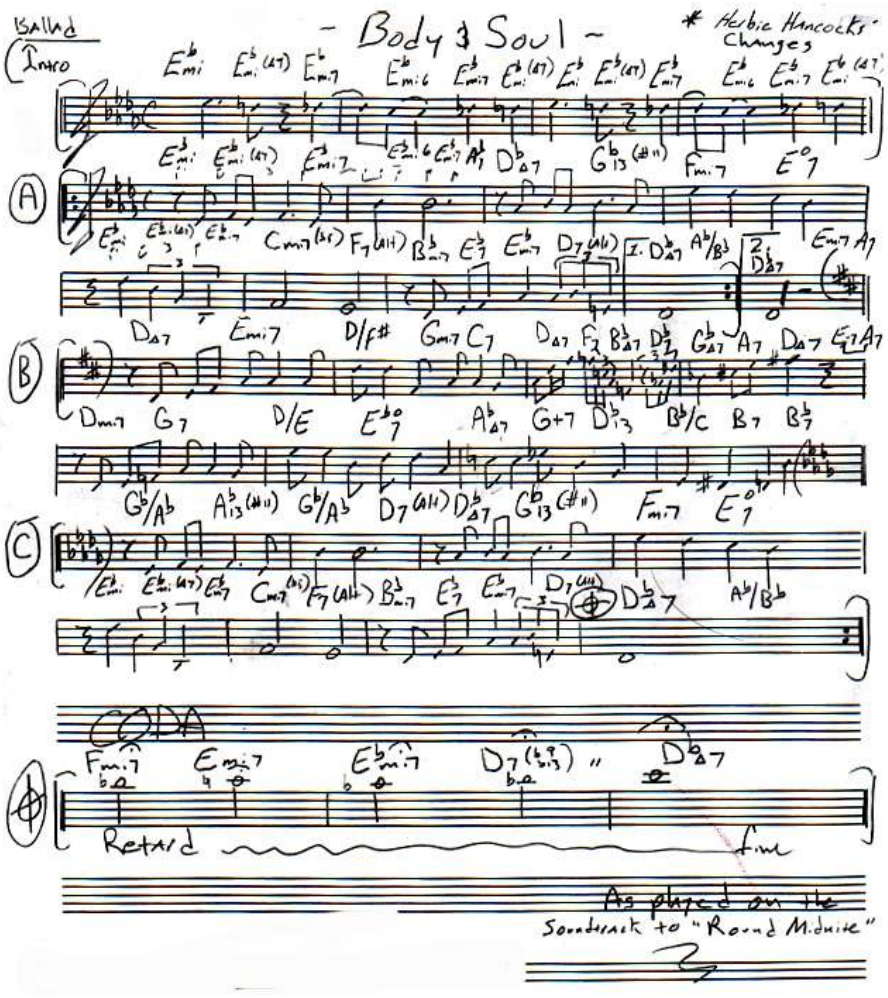


FIGURE 3.3. Herbie Hancock's chord changes on BodyandSoul in the film Round Midnight.

or templates which you can use to get a flying start in your own work. You can build games based on your simulators, or participate in competions between different simulators.

3.12 BodyandSoul: The Value of Proofs

What is the role of mathematical *proofs*, in mathematics and mathematics education? One answer could be that they are used to prove mathematical *theorems* expressing mathematical truths, of enough interest to be called theorems. An auxiliary result needed to prove a theorem is called *lemma*, a theorem of minor importance.

One could the focus on the theorem as the end result of the proof as the most important, or on the proof. The goal or the road to the goal as most important.

What could then the value be of knowing a proof of a theorem, e.g. the proof of Pythagoras Theorem? Isn't it enough to know the theorem, that $a^2 + b^2 = c^2$?

The advantage of knowing the proof is that it gives you the ability to answer the question *Why?* Why does the length c of the diagonal of a rectangle with sides of length a and b, satisfy $c^2 = a^2 + b^2$? And it is useful to know answers to the question Why?, in politics, business, science and life in general. Why is it useful?

When a child starts to ask the question *Why?* at the age of three or so, it represents an important step in the development to an adult. To answer is a parental duty which can a pleasure, if you know a good answer, but also frustrating in the many cases when you do not know an answer. In school the child quickly learns to not ask the question too often, and not in later professional life either.

In science and mathematics, the question is central Why, because this is what science and mathematics is about. That is in the ideal case. In practice science and mathematics is too often the opposite, that is to just learn by heart certain formulas and theorems, without being offered any understandable proofs.

In BodyandSoul we seek to stick to the principles and thus put emphasis on understandable proofs, as understandable answers of the question Why?

Knowing the answer gives you a strong position in arguments and also the ability to understand the meaning of the theorem. This is not simply to read what the theorem states, like a parrot, because the statement has to be properly interpreted, and if you don't know the proof it is great danger that you misunderstand. If you understand the proof, then you understand the theorem. If not, then misunderstanding is imminent.

It is well known that Einstein did not do well in mathematics in school, but it is generally believed that nevertheless he developed a mathemati-

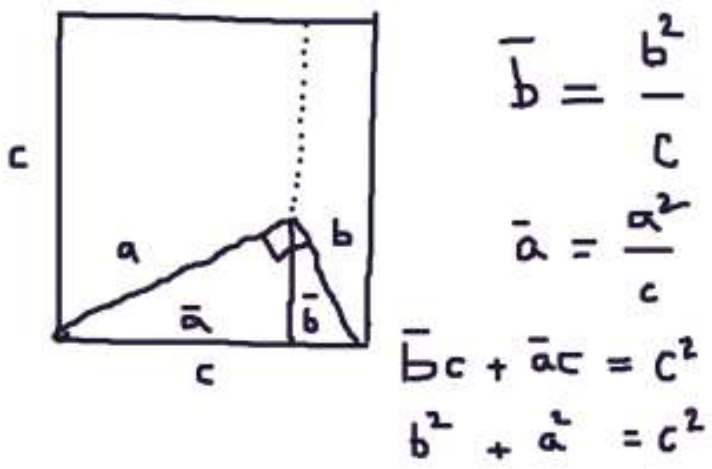


FIGURE 3.4. Proof of Pythagoras Theorem. Can you understand it? Is similarity used to show that $\bar{b} = \frac{b^2}{c}$? Compare below.

cal theory of relativity with stunning theorems about curved-space time with proofs so difficult (obscure) that nobody has ever claimed to understand them, not even Einstein himself. You can read more in the knol Did Einstein Not Understand Mathematics?.

My hope is that you will do better than Einstein in math and science, and that you will spend some time to understand the proofs you will meet. The number of proofs is kept to a minimum, based on the idea that it is better to well understand a couple of central proofs or types of arguments, than to half-understand or misunderstand a larger number. Calculus and Linear Algebra may first seem to be hopeless mess of theorems, but you will discover that it carries a long way to master a few central theorems, with proofs.

3.13 BodyandSoul: Mathematics vs Music

There are several close relations between mathematics and music, which will be illuminated as we go along. In short, music is a combination of

FIGURE 3.5. Billie Holiday singing Body and Soul

melody, *harmony* and *rythm* formed by sequences of *tones* and *chords* from different *scales* of tones. Compare Carla Bley's Ad Infinitum.

Classical music is usually performed from sheet music, written by a now dead famous composer, as interpretations by classically trained musicians capable of playing the notes according to the sheets.

Jazz music on the other is improvised without sheets to follow, only certain predetermined harmonic and rythmic patterns, like a 12 bar blues pattern $CCCC7|F7F7C7C7|G7F7C7G7$ grouped into three 4-bar patterns. A jazz musician creates a direct flow of music drawing from a toolbox of melodic, harmonic and rythmic patterns. The training of a jazz muscian consists of learning how to use certain standard tools and to develop personal tools.

We shall see that classical analytical mathematics is similar to classical notated music: It is usually very difficult and follows a preset scheme written down by a now dead famous mathematician. The role of the analytical mathematician is to interprete the mathematics of the masters, like a pianist interpreting a Beethoven Sonata by skillfully playing the right notes. An interpreter does not have to know how to compose music, only to play what is already composed.

More interestingly, we will see that a computational mathematician is like a jazz musician creating music while playing: A computational mathematician plays on the computer using tools from a toolbox, and the training

FIGURE 3.6. Doug McKenzie using tools from his toolbox

consists in learning how to use certain standard tools and to develop personal tools.

As a good example of webbased jazz piano instruction, take a look at Doug McKenzie's Youtube jazz2511's Channel:

- Doug playing Body and Soul

You see the notes being played by Doug on the keyboard coming up as sheet music automatically, and you can follow how Doug uses standard tools and his owns tools to create a new version of BodyandSoul, everytime he is playing the song. Doug helps you go peek behind the curtains and understand *how* the music is put together, and *why* certain notes are played rather than others.

The idea of BodyandSoul is the make something similar in mathematics.

3.14 The Power of Language

We do not learn to walk in school, nor to speak our mothers tongue. Children generally speak grammatically correct at the age of three without being taught any grammar at all, in some form of intuitive self-learning process based on some innate capacity for this complicated task.

Speaking, like musical improvization, means to construct new (more or less) meaningful sentences or phrases from little elements of sound.

To tell something is to describe some real or imagined phenomenon using words, as a simulation in words. You can also in words order someone to

do something according to some specification, which may give the desired result if your message is understood and has a proper form.

Mathematics offers you a special language allowing you to construct precise instructions describing real or imagined worlds, which can be understood and executed by computers.

3.15 BodyandSoul: Lyrics

My heart is sad and lonely
For you I sigh, for you dear only
Why haven't you seen it
I'm all for you body and soul
I spend my days in longing
And wondering why it's me you're (ogling)
I tell you I mean it
I'm all for you body and soul I can't believe it
It's hard to conceive it
That you turn away romance Are you pretending
It looks like the ending
And less I could have one more chance to prove, dear
My life a wreck you're making You know
I'm yours for just the taking
I'd gladly surrender myself to you body and soul
My life a wreck you're making
You know I'm yours for just the taking
I would gladly surrender myself to you body and soul.

3.16 BodyandSoul: Philosophy

- Lacan on the Unconscious
- On Bergson on Body-Soul
- Ghost in the Machine?
- Existence of Soul?
- Descartes in 3 minutes
- The Soul Paradox

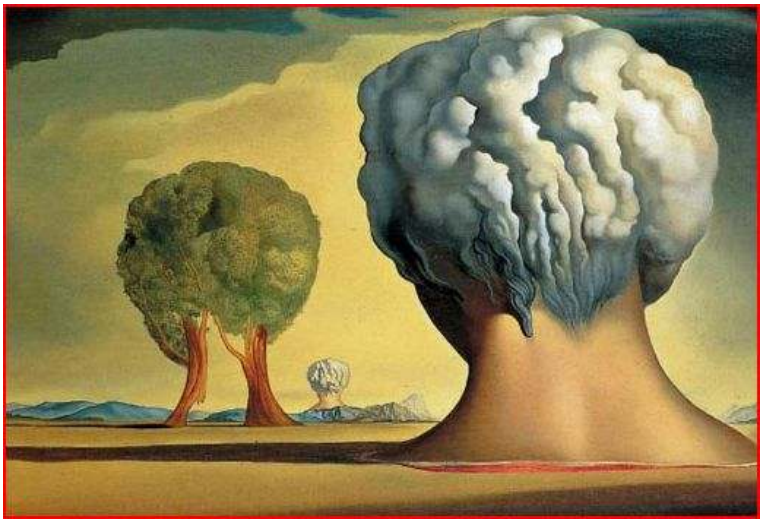


FIGURE 3.7. The Body and Soul of Salvator Dali.

3.17 BodyandSoul: FEniCS

BodyandSoul is closely related to the FEniCS open source software project aimed at setting a new standard of *Automated Computational Mathematical Modeling* by combining generality, efficiency and simplicity of mathematical methodology, implementation, and application.

In short, FEniCS offers sofware for automation of (i) formulating mathematical equations (modeling) and (ii) solving equations (computation), with the equations usually taking the form of differential/integral equations. See

- The FeniCS Project
- The Vision of FEniCS 2003
- Will FEniCS Fly?
- Talks at FEniCS09
- FEniCS10 May 10-12 KTH

FEniCS is based on the same mathematics as you will learn to master in BodyandSoul. You can use FEniCS as a model for your own software development as you develop into an expert user capable of contributing to the further development of FEniCS. In particular, you will be able to see yourself that FEniCS is concrete evidence that the methodology of BodyandSoul is functional.

FIGURE 3.8. FEniCS was born on October 30 2003

FIGURE 3.9. Header for FEniCS 10

3.18 PS: On Mathematics and Music

Musical form is close to mathematics – not perhaps to mathematics itself, but certainly to something like mathematical thinking and relationship. (Igor Stravinsky)

The most distinct and beautiful statement of any truth (as of music) must take at last the mathematical form. (Henry David Thoreau)

We do not listen with the best regard to the verses of a man who is only a poet, nor to his problems if he is only an algebraist; but if a man is at once acquainted with the geometric foundation of things and with their festal splendor, his poetry is exact and his arithmetic music. (Ralph Waldo Emerson)

It is harmony which restores unity to the contrasting parts and which moulds them into a cosmos. Harmony is divine, it consists of numerical ratios. Whosoever acquires full understanding of this number harmony, he becomes himself divine and immortal. (B. L. van der Waerden)

In the future, we can expect that not much difference will exist between education and entertainment. We just have to put intelligence behind the entertainment. (North Carolina State University's James Lester, quoted at the 12th International Conference on College Teaching and Learning)

Musical training is a more potent instrument than any other, because rhythm and harmony find their way into the inward places of soul, on which they mightily fasten, imparting grace, and making the soul of him who is rightly educated graceful. (Plato)

Music is the pleasure the human soul experiences from counting without being aware that it is counting. (Leibniz)

Mathematics and music, the most sharply contrasted fields of scientific activity which can be found, and yet related, supporting each other, as if to show forth the secret connection which ties together all the activities of our mind, and which leads us to surmise that the manifestations of the artist's genius are but the unconscious expressions of a mysteriously acting rationality. (Hermann von Helmholtz)

May not music be described as mathematics of the sense, mathematics as music of the reason? (James Joseph Sylvester)

4
Discrete-Continuum-Discrete

I see no hope for the future of our people if they are dependent on the frivolous youth of today, for certainly all youth are reckless beyond words. When I was a boy, we were taught to be discrete and respectful of elders, but the present youth are exceedingly wise and impatient of restraint. (Hesiod)

Life defies our phrases, it is infinitely continuous and subtle and shaded, whilst our verbal terms are discrete, rude and few. (William James)

And the continuity of our science has not been affected by all these turbulent happenings, as the older theories have always been included as limiting cases in the new ones. (Max Born)

At the point when continuity was interrupted by the first nuclear explosion, it would have been too easy to recover the formal sediment which linked us with an age of poetic decorum, of a preoccupation with poetic sounds. (Salvatore Quasimodo)

4.1 Life at 24 Frames/Second

You will find that computational simulations are performed by computing a sequence of frames or pictures which make up a film as a sequence of frames following upon each other with a certain time step (e.g 24 frames/second).

Each frame is computed by updating a finite number of variables, and the simulation is thus performed by using a model which is *discrete in both time and space*, that is, the model involves the values of a finite number of variables representing values at certain points in space at a discrete sequence time instants.

You will first meet such *discrete models* in the form of mass-spring systems modeling the motion of elastic bodies. By increasing the number of discrete points in time and space, that is increasing the resolution in space (increasing the number of pixels of a picture) and decreasing the time step, we will be led to *continuous models* in the form of differential equations.

For elastic solids, we thus start with discrete models and arrive at continuous models as idealizations with infinitely fine resolution in time and space.

For fluids it is more rational to start with continuous models and then perform the discretization into discrete models by the *Finite Element Method FEM* based on

- variational formulation or Galerkin's method,
- piecewise polynomial approximation.

We will recover certain discrete models for elastic solids discretizing continuous models by FEM and we will thus become familiar with the full circle discrete-continuous-discrete, where the computer only accepts discrete models and continuous models are useful to prepare for discretization by FEM.

The main tools for formulating continuous models as differential equations is

- *Calculus*: functions, derivatives and integrals.

The main tool for discretizing by FEM and solving discrete systems by computers, is

- *Linear Algebra*: vectors, matrices, linear transformations.

You collect experience of the interplay between discrete and continuous and understand that the continuous models are fictional in the sense that their solutions are "untouchable" or "unknowable" in complete detail. Nevertheless the continuos models are useful by their extreme economy of expression, which is helpful for both computation and imagination through the ingenious Calculus by Leibniz.

4.2 Watch

- 3d Facial Animator

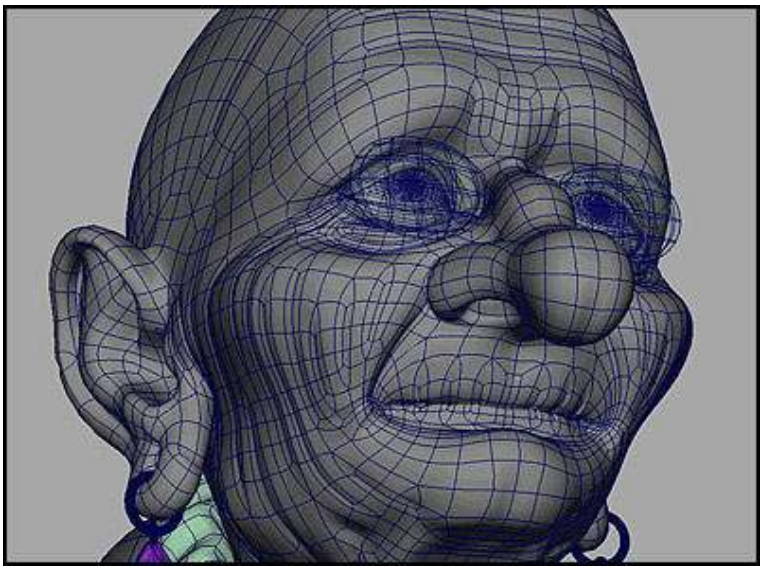


FIGURE 4.1. Piecewise polynomial face lift.

FIGURE 4.2. Untouchable continuous solution.

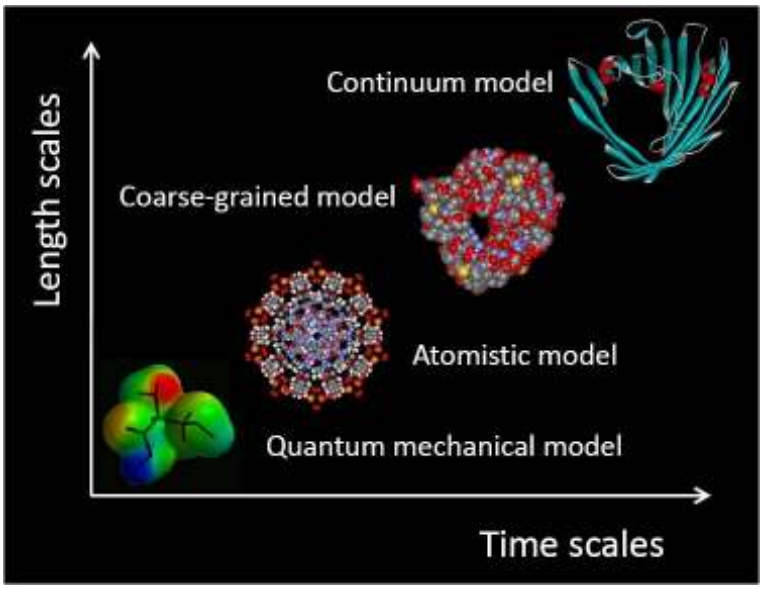


FIGURE 4.3. Multiscale modeling.

- Discrete Modeling of Facial Expressions of Emotions
- Piecewise Polynomial Approximation of Face
- 3d Head in 8 minutes

4.3 The Illusionist

We shall discover that analytical mathematics is a form of illusion playing with symbols like $\sqrt{2}$, π, $sin(1)$, and $\exp(1)$, which represent numbers with neverending non-repeating decimals expansions, which can only be specified or made known to a finite number of decimals. Analytical mathematics is thus similar to a novel as a play with words the exact meaning of which cannot be specified.

Computational mathematics on the other hand plays directly with digits and decimals and in this sense is more concrete and knowable. On the other hand it is impossible to follow all the steps of a long digital computation performed by a computer. We can thus inspect in clear light the output of a long computation, but not follow the individual steps leading to the result.

On the other hand, in an analytical argument, or derivation of an analytical formula, we are supposed to follow all the steps, but the precise nature of the result remains hidden to us.

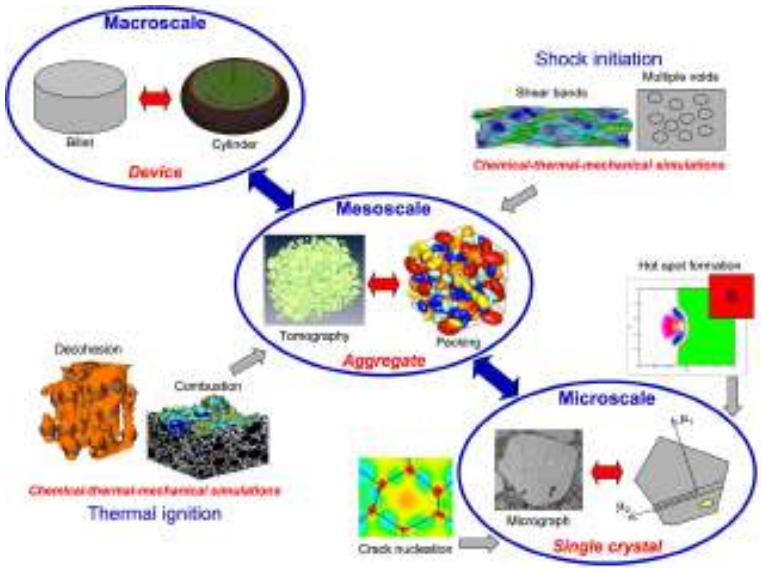


FIGURE 4.4. More Multiscale modeling.

We thus can chose bewteen the following possible ways of using mathematics:

- analytical: following the steps in detail to a partly hidden result,
- computational: not following the steps in detail to a fully visible result.

But we cannot, except in simple cases, follow all the steps in complete detail to a fully visible result.

FIGURE 4.5. Analytical Mathematician = Illusionist.

FIGURE 4.6. Fractal image of broccoli generated by one-line code.

4.4 The Genetic Code and Emergence

The genetic code is the set of rules by which information encoded in genetic material or genome (DNA or mRNA sequences) is translated into proteins (amino acid sequences) by living cells.

The human genome is the genome of Homo sapiens, which is stored on 23 chromosome pairs of 3 billion DNA base pairs, contains ca. 23,000 protein-coding genes (about 1.5 percent of the genome) while the rest consists of non-coding RNA genes, regulatory sequences, introns, and (controversially named) "junk" DNA.

The genom can be seen as a computer code which generates the life of an individual upon execution in interaction with the environment. The human genome project has listed the code, but the task to understand the code remains.

Life is an example of emergence with complex systems and patterns arising from repeated simple interactions: A short computer code, like the genome, which generates complex output, is an example of emergence: The code itself does not display the complexity which comes ou upon execution.

Fractal images are complex emergent patterns generated by one-line code loops. Turbulence is a prime non-organic example of emergence, which you will meet below.

5
Wilhelm von Humboldt and Education

> Whatever does not spring from a man's free choice, or is only the result of instruction and guidance, does not enter into his very being, but still remains alien to his true nature; he does not perform it with truly human energies, but merely with mechanical exactness. (WvH)
>
> If it were possible to make an accurate calculation of the evils which police regulations occasion, and of those which they prevent, the number of the former would, in all cases, exceed that of the latter. (WvH)

5.1 What and Why in Education?

What is education good for? What are you supposed to get out from your university studies? Is it

- *liberal arts education* forming you into a good knowledgable democratic citizen?
- science/technology/economics forming you to a production wheel in society's economical growth?
- a road to free your creative spirits and intellectual capacity?

FIGURE 5.1. Wilhelm von Humboldt: *True enjoyment comes from activity of the mind and exercise of the body; the two are ever united.*

FIGURE 5.2. Wilhelm von Humboldt and his University

Liberal arts education has its roots in the Greek *paideia* as an education for free citizens to develop ethics and logics insights to become a good human being and member of society, with emphasis on skill of rethorics in political discussion and argumentation towards good goals. In the same way as physical training can make your Body more able, studies in *philosophy* were believed to strengthen the abilities and moral of your Soul.

With his two years younger brother Alexander, Wilhelm von Humboldt (1767-1835) belonged to a generation which witnessed the collapse of absolute monarchies in the wake of the French Revolution and helped to shape the construction of a new Europe and Prussian State. The two brothers were both educated in the spirit of Rousseau and of the philanthropic school; in their youth, they adopted the ideas of the enlightenment, lived through the Sturm und drang period and went on to join the Weimar circle of poets where they enjoyed the friendship of Schiller and Goethe. While Alexander travelled the world and guided natural science into new paths, Wilhelm paved the way for the development of the modern moral sciences.

Wilhelm lay the foundations of a new education system in Prussia leading into the liberal arts education of the 20th century. In his *Theory of Human Education* from 1793, Wilhelm states the

- *the ultimate task of our existence is to give the fullest possible content to the concept of humanity in our own person [...] through the impact of actions in our own lives.*

The *paideia* and philosophy of liberal arts education is largely missing in science/technology/economics of today, but Wilhelm's ideas on a modern

FIGURE 5.3. The travels by Alexander von Humboldt.

educational theory have been attracting increasing attention in recent years and may meet a renaissance in our new information age.

In all modesty BodyandSoul seeks to follow maxims of Wilhelm von Humboldt such as

- *How a person masters his fate is more important than what his fate is.*
- *The government is best which makes itself unnecessary.*
- *True enjoyment comes from activity of the mind and exercise of the body; the two are ever united.*
- *If we glance at the most important revolutions in history, we see at once that the greatest number of these originated in the periodical revolutions on the human mind.*
- *Coercion may prevent many transgressions; but it robs even actions which are legal of a part of their beauty. Freedom may lead to many transgressions, but it lends even to vices a less ignoble form.*
- *However great an evil immorality may be, we must not forget that it is not without its beneficial consequences. It is only through extremes that men can arrive at the middle path of wisdom and virtue.*

Compare with Measuring the World about to giants in mathematics/science Carl Friedrich Gauss and Alexander von Humboldt:

- *One travels, one stays at home. One is liberal, the other conservative. One is a lover of women, the other forms suspicious attachments to men. They are bound by genius and nationality.And, finally, at the 1828 Scientific Congress in Berlin, they meet...*

6
Simulated Hyperreality: Disney World

> What you have to do is enter the fiction of America, enter America as fiction. It is, indeed, on this fictive basis that it dominates the world. (Baudrillard)
>
> The very definition of the real becomes: that of which it is possible to give an equivalent reproduction. The real is not only what can be reproduced, but that which is always already reproduced. The hyper real. (Baudrillard)
>
> Deep down, the US, with its space, its technological refinement, its bluff good conscience, even in those spaces which it opens up for simulation, is the only remaining primitive society. (Baudrillard)

According to the French post-modern philosopher Jean Baudrillard, Disney World is a simulation of a fictitiuos real world which does not exist. In other words, Disney World is an example of hyperreality. Baudrillard makes a distinction between reality and hyperreality according to the following characteristics:

- real: what can be reproduced
- hyperreal: what is already reproduced
- hyperreal: model of real without real origin
- hyperreal: masks non-existence of real origin

FIGURE 6.1. Hyperreal characters from a hyperreal world.

as expressed in his treatise Simulacra and Simulation:

- *The simulacrum is never that which conceals the truth–it is the truth which conceals that there is none. The simulacrum is true.*

Baudrillard identifies the following forms of simulation:

1st Order Simulation:

- map of territory
- simulation with real origin
- clear difference between simulation and origin.

2nd Order Simulation:

- map covers territory (Borges On Exactitude in Science)
- simulation cannot be distinguished from origin
- including reproductions of original art, clothes,...

3rd Order Simulation:

- map replaces territory
- simulation without origin
- outside realm of good and evil only performance counts computer game.

Baudrillard gives the following examples:

- Disneyland: simulation of non-existing idyllic America
- Barbie doll: simulation of non-existing female physics
- Watergate process: mask of non-existing true judiciary process.

Once you start to think about it, you will find that matematical simulation of real phenomena have qualities of hyperreality, as exposed in Hyperreality in Physics. See also Simulations by Wittgenstein

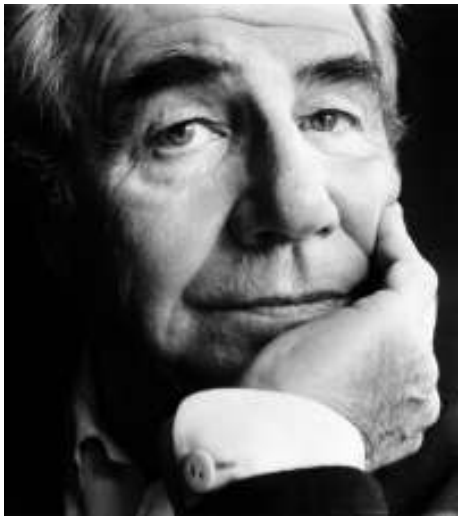


FIGURE 6.2. Photographic simulation of Baudrillard.

FIGURE 6.3. Jacking in to the Matrix.

7
Avatar Simulation Techniques

> Everything is backwards now, like out there is the true world, and in here is the dream. (Jake Sully)

The movie Avatar is based on performance image capture which creates computerized images from real human action including human emotions through facial of body expression.

A number of new revolutionary visual effects techniques were used in the production of Avatar: To achieve the face capturing, actors wore individually made skull caps fitted with a tiny camera positioned in front of the actors' faces; the information collected about their facial expressions and eyes is then transmitted to computers.[94] According to Cameron, the method allows the filmmakers to transfer 100 percent of the actors' physical performances to their digital counterparts. Besides the performance capture data which were transferred directly to the computers, numerous reference cameras gave the digital artists multiple angles of each performance.

Unlike past methods that captured dots placed on human faces to trace movements that are reconstructed digitally, each frame was analyzed for facial details such as pores and wrinkles that help re-create a moving computerized image.

7.1 Watch

- Avatar Animation Technique 1
- Avatar Animation Technique 2

FIGURE 7.1. Human Jake Sully as Avatar and Neytei Zoe Saldana.

FIGURE 7.2. Avatar emotion capture.

8

The Secret Pythagorean Society

> There is geometry in the humming of the strings, there is music in the spacing of the spheres... There is nothing so easy but that it becomes difficult when you do it reluctantly. (Pythagoras)

The secret Pythagorean Society in Greece 400 BC led by Pythagoras was based on the belief that everything in the World can be represented as relations between the natural numbers 1,2,3,.....But one day somebody discovered that $\sqrt{2}$, the length of the diagonal of a square with side 1, cannot be expressed as a *rational number* as the quotient of two natural numbers, e.g. as $\frac{22}{7} = 3\frac{1}{7}$.

It was thus discovered that $\sqrt{2}$ is not a rational number, that is, that $\sqrt{2}$ is an *irrational number*. This was first kept as a secret, but like in Climategate a whistleblower revealed the secret public. This was so devastating to the basic belief of the Pythagoreans that their society collapsed, and was replaced by the geometric School of Euclide, which resolve the difficulty of the irrationality of $\sqrt{2}$, by simply defining $\sqrt{2}$ geometrically as the length of the diagonal of a square with side 1.

The geometric school of Euclide propagated by Aristotle ruled science for almost 2000 years until Descartes in the 17th century replaced geometry by analytic geometry based on numbers thus returning to Pythagoras, and initiating the scientific revolution transforming medieval society into the industrial society of the 19th century leading up to the information society of the late 20th and 21st century, which you are lucky to have been born into. A Pythagorean society based on numbers!

You may recall 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 $1.414^2 = 1.999386$. No matter how many decimals of x we add, x^2 will not become exactly equal 2. For,example, with 415 decimals and

$$\begin{aligned} x = 1.&41421356237309504880168872420969807856967187537694807317667973799073247846210703885038753432\\ &76415727350138462309122970249248360558507372126441214970999358314132226659275055927557999505\\ &01152782060571470109559971605970274534596862014728517418640889198609552329230484308714321450\\ &83976260362799525140798968725339654633180882964062061525835239505474575028775996172983557522\\ &03375318570113543746034084988471603868999706991936436722392248627179189098787015809960232\\ &64059726131264076040569129995030929574783188859695007088740560583365016522715738094455933206900458172642221739359695332425151587602336042729948891418035989710382049561848123333216251601609728313712306449949794365347969862977651683340665770240318513306002427232125175273043547767486608089987807935797774759645877082503170068870585486010 \end{aligned}$$

we have that

$$\begin{aligned} x^2 = 1.&99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999998638103700279039354754492148156752071936433672239224862717918909878701580996023264059726131264076040569129995030929574783188859695007088740560583365016522715738094455933206900458172642221739359695332425151587602336042729948891418035989710382049561848123333216251601609728313712306449949794365347969862977651683340665770240318513306002427232125175273043547767486608089987807935797774759645877082503170068870585486010 \end{aligned}$$

No matter how many decimals we take in a guess x of $\sqrt{2}$, we never get a number which squared gives exactly 2.

Can you prove that? Can you reveal the secret of the Pythagorean society, the knowledge that ended the reign of the Pythagoraens?

Hint: Assume that $\sqrt{2} = \frac{p}{q}$ with all common factors of 2 in the natural numbers p and q. Then consider the equation $2q^2 = p^2$ obtained by squaring and multiplying by $2q^2$. Conclude that p must contain the factor 2 and thus p^2 the factor $4 = 2 \times 2 = 2^2$. Conclude that q must contain a factor 2, which contradicts the assumption that p and q have no common factor 2.

A rational number has a finite or periodic (repeating) decimal expansion, and an irrational number has a neverending non-periodic (non-repeating) decimal expansion. Since it is impossible to compute all decimals of an irrational number, we must acknowledge that an irrational number really is "irrational" in the sense that its exact value cannot be pinned down in the same precise sense as for a natural or rational number. Irrational numbers satisfy the same computational rules as rational numbers, but their exact values are hidden to our inspection: there is always another decimal to be computed/discovered somehow. In particular, given two irrational numbers, it may be impossible to decide if they are exactly equal (all decimals being equal) or not. For example, the statement $0.99999999... = 1$ is correct only if the dots indicate a (periodic) never-ending squence of the digit 9.

Another irrational number is $\pi = 3,14159265...$, which has been computed to 2 billion digits, but that is not the whole truth...

8.1 $\sqrt{2}$-gate, Climategate and Watergate

This was the argument which was kept secret by the Pythagoraen society, and in a form of $\sqrt{2}$-*gate* led to the collapse of the society, when it leaked. Recall that Watergate was the political scandal in the US in the 1970s caused by the break-in into the Democratic National Committee headquarters at the Watergate office complex in Washington, D.C, which led to the resignation of President Richard Nixon, and indictment and conviction of several Nixon administration officials.

Climategate unfolded in November 2009 when a whistleblower uploaded thousands of emails by scientists connected to the UN Intergovernmental Panel for Climate Change IPCC formed to study *Anthropogenic Global Warming AGW* by CO_2 emission from burning of fossil fuels. The emails revealed questionable scientific methods and thus undermined the scientific basis of AGW.

Below you will meet mathematical models of weather and climate of different complexity , and you will discover that understanding mathematics helps to uncover some of the mysteries of weather and climate. There is plenty of experimental data because the physical experiment is going all time all over the globe.

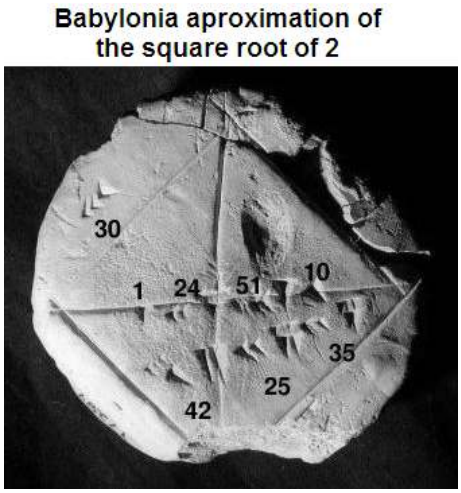


FIGURE 8.1. Babylonian approximation of $\sqrt{2}$.

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \ddots}}}}.$$

FIGURE 8.2. A formula for computing $\sqrt{2}$. From where?

8.2 Pythagoras and Music

Pythagoras belief that the World is based on numbers, was supported by his discovery that the ratio of frequencies of musical scales are simple rational numbers such as $\frac{3}{2}$ for a fifth (G in a scale of C), $\frac{9}{8}$ for a second (D), $\frac{27}{16}$ for a sixth (A), and $\frac{4}{3}$ for a fourth (F).

8.3 Read More

- The Squareroot of Two.

9

Aristotle and Hypatia: Mathematicians

> Reserve your right to think, for even to think wrongly is better than not to think at all. (Hypathia)
>
> Humor is the only test of gravity, and gravity of humor; for a subject which will not bear raillery is suspicious, and a jest which will not bear serious examination is false wit. (Aristotle)

As a formidable student, researcher, teacher, and philosopher in virtually all scientific disciplines, Aristotle (384 BC -322 BC) had a profound impact on the way science and mathematics is practiced and investigated today. His analytical method, now known as Aristotelian logic, is the backbone of not only mathematics, but of all the natural sciences.

Hypatia of Alexandria (born between 350 and 370; died 415) was a Greek scholar from Alexandria in Egypt, considered the first notable woman in mathematics, who also taught philosophy and astronomy. She lived in Roman Egypt, and was killed by a Christian mob who falsely blamed her for religious turmoil, see the recent film Agora.

A Neoplatonist philosopher, she belonged to the mathematical tradition of the Academy of Athens represented by Eudoxus of Cnidus; she followed the school of the 3rd century thinker Plotinus, discouraging empirical enquiry and encouraging logical and mathematical studies.

John of Nikiu (7th century) writes;

- *And in those days there appeared in Alexandria a female philosopher, a pagan named Hypatia, and she was devoted at all times to magic, astrolabes and instruments of music, and she beguiled many people*

FIGURE 9.1. Aristotle: Philosopher and Mathematician

FIGURE 9.2. Destruction of Library of Alexandria: In 391, Christian Emperor Theodosius I ordered the destruction of all "pagan" (non-Christian) temples, and the Christian Patriarch Theophilus of Alexandria complied with this request.

FIGURE 9.3. Hypatia from Alexandria 350-415: Mathematician and Scientist.

through Satanic wiles...A multitude of believers in God arose under the guidance of Peter the magistrate...and they proceeded to seek for the pagan woman who had beguiled the people of the city and the prefect through her enchantments. And when they learnt the place where she was, they proceeded to her and found her...they dragged her along till they brought her to the great church, named Caesareum. Now this was in the days of the fast. And they tore off her clothing and dragged her...through the streets of the city till she died. And they carried her to a place named Cinaron, and they burned her body with fire.

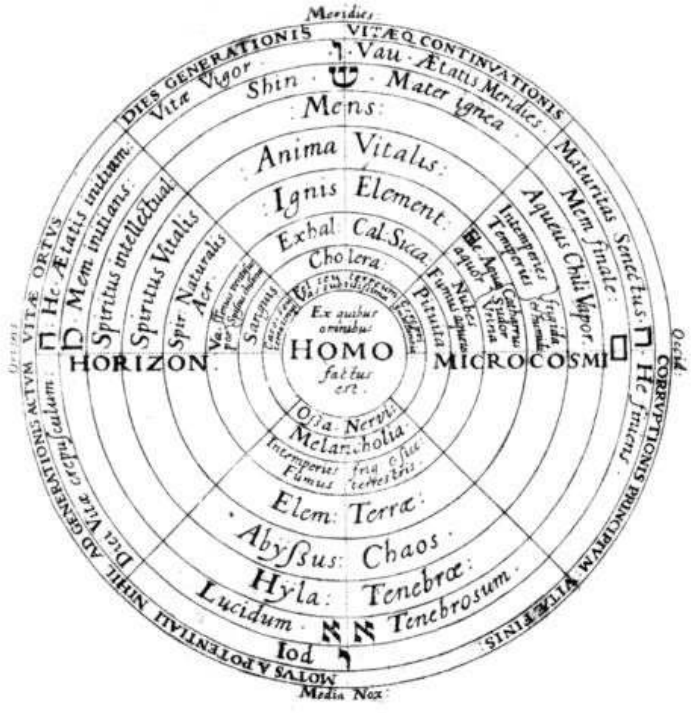


FIGURE 9.4. Man's Constitution and the World's Constitution.

9.1 Integers and Rational Numbers

Of course you know about natural numbers $0, 1, 2, 3, ...$, and integers $0, \pm 1, \pm 2, \pm 3, ...$ and rational numbers as quotients $\frac{p}{q}$ of integers with $q \neq 0$. You may also know that rational numbers have ending (finite) or periodic (non-ending) decimal expansions. You can refresh your conception of rational numbers by reading

- Integers.
- Rational numbers.

Remember that numbers is the basis of the digital world or IT-world.

9.2 Read More

- Hans Kayser, 20th Century Pythagorean Master.

10
Cover Story: Icarus

Icarus: All limits are self-imposed.

Daedalus: If you fly too low, the waves will soak your dragging feathers, and make them too heavy. If you fly too high, the sun will scorch your feathers and make them too heavy.

If you want you can choose as your companions through your studies, the young Icarus and his father Daedalus, a craftsman from Athens. Icarus and Daedalus were imprisoned by King Minos in the Labyrinth of Knossos built for the Minotaur, half-man half-bull. Daedalus was exiled because he gave Minos' daughter, Ariadne, a clew of string, ball of yarn, in order to help Theseus, the enemy of Minos, survive the Labyrinth and defeat the Minotaur.

Similarly, you can view yourself imprisoned in a Labyrinth of Ignorance, with the challenge to get out to liberate your potential.

To get into mood, watch:

- Icarus Cup
- Icarus Flying Machine
- Da Vinics Decoded
- Testing Da Vinci's Flying Machine
- Simulating Flight.

Your studies may now start with the following dialog:

FIGURE 10.1. Daedalus teaching Icarus the basics of aerodynamics.

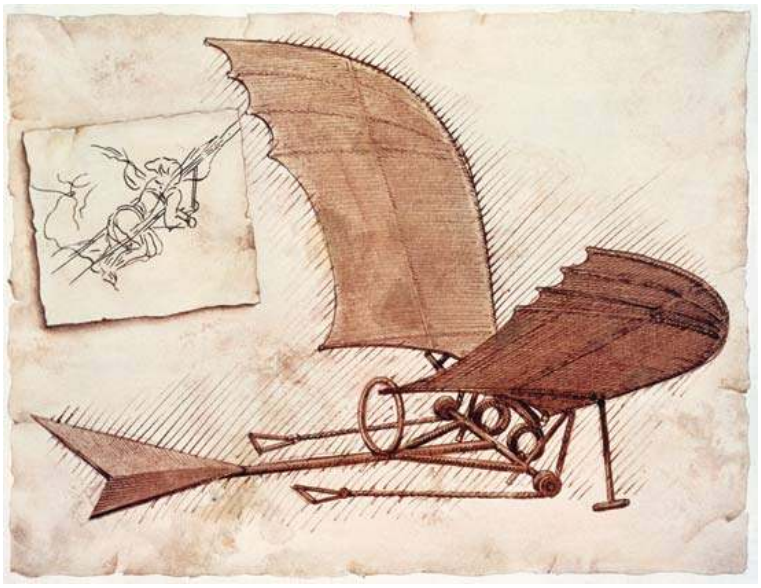


FIGURE 10.2. Dream of Icarus I

10.1 The Story

Daedalus: We are in real trouble. I fear we will have to stay imprisoned for ever. I see no way out.

Icarus: Yes, it seems hopeless. But this morning something surprising happened: A bird flew in to the Labyrinth and dropped a piece of paper in front of me. Let me show it to you:

$$\begin{aligned} \dot{\rho} + \nabla \cdot (\rho u) &= 0 \\ \dot{m} + \nabla \cdot (mu + p) &= f \\ \dot{\epsilon} + \nabla \cdot (\epsilon u + pu) &= 0 \end{aligned} \tag{10.1}$$

Deadalus: Let me see. Strange. Is it some kind rebus to be resolved?

Icarus: Can it be a message telling us how to get out? A message from a bird telling us how fly as a bird? I have some friends out there who may help us to understand.

Deadalus: Can you get the bird to bring the paper to your friends for them to dechiffer the message?

Icarus: I'll try. Here the bird comes again...I except to get some feedback...

Daedalus That is our only hope...

FIGURE 10.3. Dream of Icarus II.

(A couple of days later:)

Icarus: See, there same bird again. It is dropping a message. I have it, let see what it says:

- *The rebus is a set of mathematical equations describing the flow of air. The equations are called the Navier-Stokes equations. The solution hides the secret of flight, which the birds have discovered but keep for themselves. The trouble is that it seems impossible to find a formula for the solution. People have tried for centuries all possible formulas but no-one works. Some say that it is because the solution is turbulent and there is no formula for turbulence. The rich merchant Cassius Clay has offered one million Drachmas together with his beautiful daugther Madonna to the person who can find the formula, but it has not helped. But there are some rumours that it is possible to find a solution using a computer, and we are now pursuing this upshot. We will inform you about any progress. Don't give up your hopes. Since the birds can fly, it should be possible also for humans...*

Daedalus: Can it really be possible for us to fly? I know that mathematicians have proved long ago that it is impossible, but since birds anyway fly there must be something wrong with the mathematics, if I rely on my engineering intuition.

10.2 Your Role

So there you are: You are one of the friends of Icarus and you want to help Icarus out of the Labyrinth. In doing so you will help yourself out of your prison of ignorance. To your disposal you have the Internet and a computer, plus pen and paper. I suggest you start with the question: Is it possible to fly?

You can view your studies as a form of computer game with the objective of liberating Icarus and Daedalus from imprisonment, that is yourself from Ignorance.

You will find that you can see the objective of your studies in science and engineering to be

- *to simulate the world*,

which can be described as a

- *how to construct computer games*

because simulation is a form of game performed with a computer.

10.3 Do Not Read: Secret of Flight

- why it is possible to fly

10.4 Do Not Read: Secret of Turbulence

- computational turbulent flow

10.5 To Watch

- Early Flight Attempts
- Orthithopter
- Wright's Flyer
- Wright Brothers
- Flyer Replica
- Otto Lilienthal
- Modern Lilienthal

FIGURE 10.4. Otto Lilienthal before breaking his neck by stalling at 15 m height.

10.6 Human Powered Aircraft HPA

The dream of Icarus was realized only recently: The first Human Powered Aircraft HPA was the Gossamer Condor, see slideshow and video, with a wingspan of 30 meters, wing area of 60 squaremeters and weight 32 kg, which in 1977 managed to cruise for 7 minutes at speed of 5 meters per second powered by human legs delivering 0.3 horse powers. The Gossamer Condor was the winner of the Kremer Prize of 50.000 pounds Sterling, and there are more Kremer prizes to win. See also MIT Daedalus.

Following BodyandSoul you will be able to understand how this was possible. The key ingredients are: speed V, weight W, power P connected by the following magic formula

$$P = \frac{W}{F}V \tag{10.2}$$

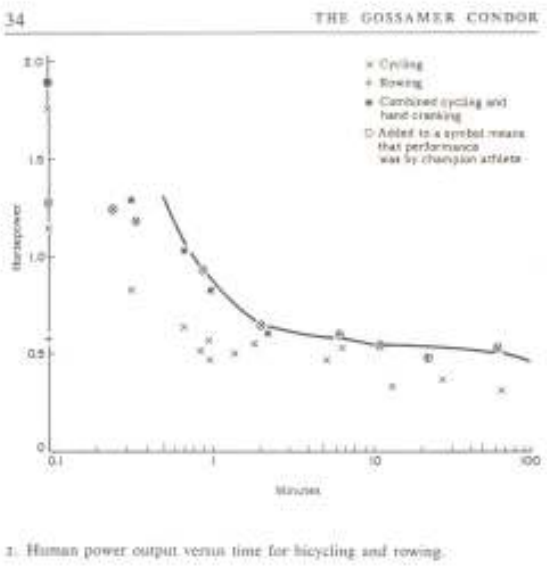
where F, called the *finesse,* is a magical factor. With $F = 20$, $W = 1000$ Newton and $V = 5$ meter/second, we get

$$\frac{W}{F}V = \frac{100 \times 5}{20} = 250 \quad \text{Newtonmeter/second} \ = 0.3 \quad \text{horsepowers}, \tag{10.3}$$

thus satisfying the requirement of (10.2). There is a second magic formula for the required wing area S:

$$W = c_L V^2 S = 0.7 V^2 S, \ \text{that is} \ S = \frac{1000}{0.7 \times 25} \approx 60 \ \text{square meters}. \tag{10.4}$$

FIGURE 10.5. The first Human Powered Aircraft HPA: The Gossamer Condor, August 23 1977



1. Human power output versus time for bicycling and rowing.

FIGURE 10.6. Icarus tests of human power output.

FIGURE 10.7. MIT Daedalus: A later HPA project.

with $c_L = 0.7$ as a second magical factor.

It remains for you to understand what the magical finesse factor $F = 20$ represents as well as the factor 0.7 in the wing area formula. You will find that the answers are hidden in the equations (10.1). Are you ready for lift-off?

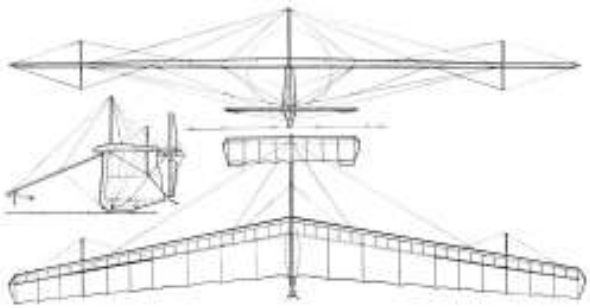


FIGURE 10.8. Construction drawings of Gossamer Condor.

11
Ask Why? Be Scientist!

> What's the go of that? What's the particular go of that? (James Clerk Maxwell (1831-1879) Scottish physicist. Comments made as a child expressing his curiousity about mechanical things and physical phenomena)
>
> Why are things as they are and not otherwise? (Kepler (1571-1630))
>
> To raise new questions, new possibilities, to regard old problems from a new angle, requires creative imagination and marks real advance in science. (Einstein)

11.1 From Questioning to Understanding

BodyandSoul encourages you to be critical, to ask questions, and only accept what you can understand on rational grounds. You will find that the nature of mathematics invites to such a critical approach, because in mathematics you draw conclusions from certain assumptions using logic and symbolic or numerical computation. If the assumptions are clearly stated, and each logical and computational step is open to inspection, then it is possible to objectively check if mathematical conclusion or result is correct or not, up to the correctness of the assumptions.

In other words, you will be able to work very much like a scientist, like a critical scientist who constantly ask ther questions Why? and Why Not?

You will yourself discover some of the power of this approach (and also some of its limitations).

As a child you asked many questions, but then later in school you learned not to ask too much. In a way you should now try to recover from this effect of your schooling and return to the questioning of your childhood. It is not always so easy but it can be very rewarding. The Internet and the computer are at your disposal, and do not get tired by too many questions (like maybe your teachers, friends and family) or much work, and thus can give you good answers if you can only discriminate. To learn to do so is part of the critical training you can get through BodyandSoul.

You will discover that to say that you understand something of a some physical process, typically means that there is an underlying mathematical model with certain properties. For example, if you say that you understand the motion of pendulum swinging back and forth, as a repeated exchange between potential and kinetic energies, it means that you know the equations of motion of the pendulum and you can prove e.g. that the sum of potential and kinetic energies remains constant.

Or if you say that you understand how an ice skater can increase the spin faster by pulling the arms tight into the body, it means that you know the equations of motion and the connection between spin and moment of inertia.

11.2 Some Questions

As a mathematical scientist you should be ready ask for example, **WHY** is it so that

$$\begin{aligned}
1+1 &= 2,\\
(-1)(-1) &= 1,\\
2+3 &= 3+2,\\
\exp(a)\exp(b) &= \exp(a+b),\\
\log(ab) &= \log(a)+\log(b),\\
\exp(\log(a)) &= a,\\
\sin(t)^2+\cos(t)^2 &= 1,\\
\text{length of the perimeter of a circle of unit radius} &= 2\pi,\\
\text{area of a circular disc of unit radius} &= \pi,\\
\text{volume of a sphere of unit radius} &= \frac{4}{3}\pi.
\end{aligned} \tag{11.1}$$

Maybe you already know good answers, but if you don't know, don't worry; you will naturally discover the answers as you go along, and answers to many more questions...

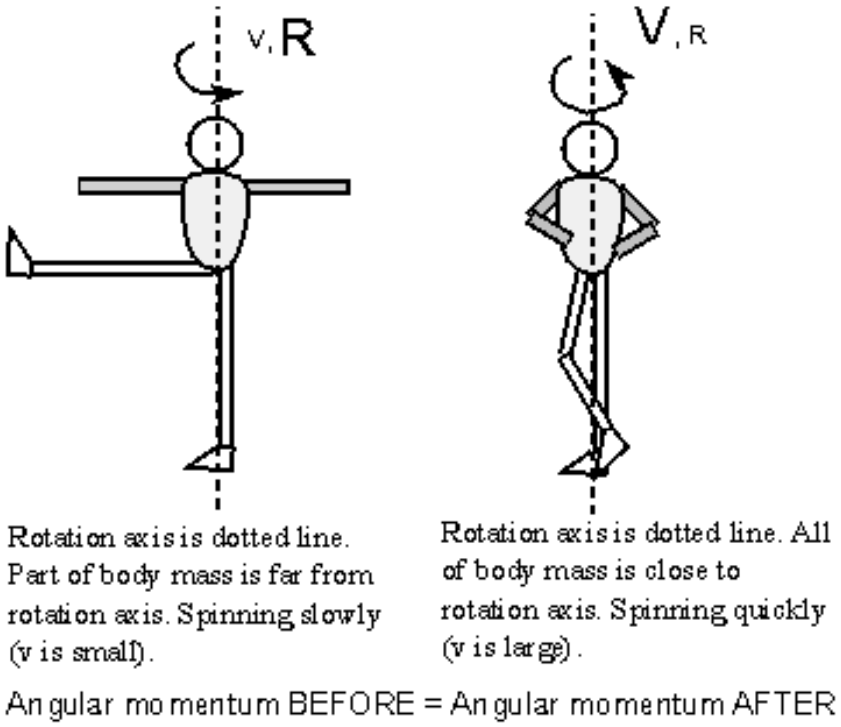


FIGURE 11.1. Spinning quickly by decreasing the moment of inertia while keeping total angular momemntum constant.

FIGURE 11.2. Clerk Maxwell as a child with his kind mother answering his questions: *What's the go of that? What's the particular go of that?*

12
The Secret Agenda

> Everyone has a hidden agenda. Except me! (Michael Crichton)
>
> Later mathematicians will regard set theory as a disease from which one has recovered. (Henri Poincaré)

.

Here is a document describing a secret agenda that will lead to resolving the mystery of the Navier-Stokes equations and thus the mystery of flight.

1. **1d problem**:
2. motion of a body B along a 1d straight line with coordinates x
3. $x(t)$ position of B at time t
4. velocity $v(t)$ as change dx of position $x(t)$ per unit time step dt: $dx = vdt$
5. accelleration $a(t)$ as change dv of velocity $v(t)$ per unit time step: $dv = adt$
6. Newton's 2nd Law $F = Ma$ or $a = \frac{F}{M}$ with F force and M mass
7. time stepping: With $v^n = v(ndt)$, $x^n = x(ndt)$, do for $n = 0, 1, 2, ...$,
8. $v^{n+1} = v^n + a^n dt = v^n + \frac{F^n}{M} dt$
9. $x^{n+1} = x^n + v^n dt$

10. B moves with zero force: $F = 0$:
11. $v(t) = v$ is constant
12. $x(t) = tv + \bar{x}$
13. B moves with constant force F: $a = F/M$ constant:
14. $v(t) = ta + \bar{v}$ straight line
15. $x(t) = t^2/2a + t\bar{v} + \bar{x}$ curved line.
16. **same in 2d or 3d**:
17. B moves in 2d Euclidean plane with coordinates $x = (x_1, x_2)$ or $x = (x_1, x_2, x_3)$
18. $x(t)$ position of B depending on time t
19. velocity $v(t)$ as change dx of position $x(t)$ per unit time step dt: $dx = vdt$
20. accelleration $a(t)$ as change dv of velocity $v(t)$ per unit time step: $dv = adt$
21. Newton's 2nd Law $F = Ma$ or $a = \frac{F}{M}$ with F force and M mass
22. Time stepping: With $v^n = v(ndt)$, $x^n = x(ndt)$, do for $n = 0, 1, 2, ...$,
23. $v^{n+1} = v^n + a^n dt = v^n + \frac{F^n}{M} dt$
24. $x^{n+1} = x^n + v^n dt$
25. B moves with $F = 0$, F constant, $F(t)$ variable
26. connect B to spring with spring force depending on position of B
27. mass-spring systems of several masses and springs in 1-3d
28. mass-spring systems with viscosity
29. **derivative with respect to space coordinate: space derivative**:
30. gradient, Laplacian
31. Euler's and Navier-Stokes' equations of fluid mechanics
32. Navier's equations of solid mechanics
33. waves in fluids and solids: wave equation
34. Fourier's equation for heat conduction/diffusion
35. Maxwell's equations for electromagnetics
36. discover secret of flight.

FIGURE 12.1. Two approaches to describing and understanding the World.

FIGURE 12.2. Describing the World.

FIGURE 12.3. Describing the World: Can you identify the equations, and the persons behind the equations?

12.1 Watch

- Being in Time according to Heidegger
- Task of Thinking
- Wittgenstein on Language Games
- The World Is All That Is The Case
- Wovon man nicht sprechen kann darüber muss mann schweigen

13
Global Warming?

Climate change should be seen as the greatest challenge to face man and treated as a much bigger priority in the United Kingdom. (Prince Charles)

The issue of climate change is one that we ignore at our own peril...What we can be scientifically certain of is that our continued use of fossil fuels is pushing us to a point of no return. And unless we free ourselves from a dependence on these fossil fuels and chart a new course on energy in this country, we are condemning future generations to global catastrophe. (Barack Obama)

I want to testify today about what I believe is a planetary emergency - a crisis that threatens the survival of our civilization and the habitability of the Earth. (Al Gore)

All across the world, in every kind of environment and region known to man, increasingly dangerous weather patterns and devastating storms are abruptly putting an end to the long-running debate over whether or not climate change is real. Not only is it real, it's here, and its effects are giving rise to a frighteningly new global phenomenon: the man-made natural disaster. (Barack Obama)

FIGURE 13.1. Common picture in the climate debate. True?

13.1 Climate Sensitivity?

Once Icarus and Daedalus have escaped from the Labyrinth of Ignorance, they wiil be ready to take on problems. What is the biggest problem facing humanity today? Is it Global Warming because of increasing CO_2 in the atmosphere from burning fossil fuels like oil and coal, and from humans breathing and cows letting out? Can we all go on breathing or will it be reserved for the rich?

The key question is *climate sensitivity*, which is how much global mean temperature will increase if the concentration of CO_2 doubles from the present 0.038%. The United Nations Intergovernmentak Panel on Climate Change IPCC tells us that the increase can be about 3 degrees Celcius C, give and take 1.5 C, that is up to 4.5 C, which if true would end human civilization as we know it. Jurrasic Park would be back. But from where does the 4.5 C or more of potentially catastrophical global warming, come? Can the climate sensitivity really be so catastropically large? And if so, then why?

To start our studies, let's try our hands on this problem. As always we ask

- What are the physical laws?
- What are the numbers?

Once we know the answers to these basic questions, we can use mathematics to produce some answer.

OK, so what do we have here? Well, the Earth with atmosphere is heated by the Sun through incoming radiation of all wavelengths, short to long,

and the Earth with atmosphere (troposphere plus stratosphere) radiates longwave infrared light to outer space. The reason the outgoing radiation is infrared is that the Earth is not nearly as hot as the Sun, and a colder object (a so called black body) tends to transform incoming shortwave to outgoing longwave. It acts like a transformer transforming high voltage to low voltage current.

We don't need to know the details of the absorption-emission process of a black body, only *Stefan-Boltzmann's Radiation Law* stating that energy from a surface of a black body (like the Sun or the Earth with atmosphere) is proportional to the fourth power of the surface temperature in degrees Kelvin K. We recall that 273 Kelvin = 0 Celcius and 373 K = 100 C, so that temperature increase measured in Kelvin K or Celcius C is the same. Stefan-Boltmann's law applies to any isolated bodies, in particular the Earth with its atmosphere if the temprerature is that of the upper atmosphere.

This means that doubling the temperature increases the radiated energy by the factor $2^4 = 16$. That is the physical law. Now to the numbers:

- The surface temperature of the Sun can be estimated to 5700 K.
- The radius of the Sun is 695.000 kilometers about 2.3 light seconds.
- The distance between the Earth and the Sun is 500 light seconds.
- The ratio is about 220, and since the area of a sphere scales like the radius squared, the intensity of the incoming Sun light is decreased by a factor 220^2.
- We assume that the incoming light is distributed evenly over the surface of the Earth (with atmosphere), which is 4 times bigger than the disc area as seen from the Sun, and thus the radition from the Sun is diluted by the factor 4×220^2.

Now we start the mathematics combining laws with numbers:

- Whatever the Earth (with atmosphere) absorbs from the Sun has to be emmitted according to Stefan-Boltzmann's law. What ratio r of the temperatures of Earth atmosphere and of the Sun, then gives $r^4 = 4 \times 220^2 \approx 194000$? Your pocket calculator gives $r \approx 21$.
- We conclude that the temperature on the top of the Earth atmosphere must be $\frac{5700}{21} \approx 273$ K or 0 C. This agrees pretty well the observed temperature at the top of the stratosphere.
- If the Earth did not have an insulating atmosphere this would also be the surface temperature of the Earth. Without insulating atmopshere the Earth could be covered by ice at 0 C. This agrees with the observation that the night temperature in Sahara (with clear sky an very dry air) is about 0 C.

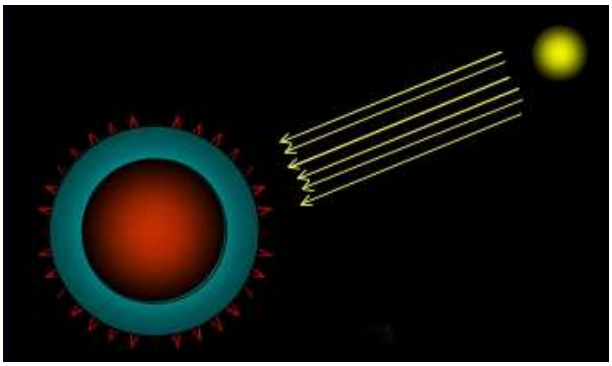


FIGURE 13.2. The Earth with insulating atmosphere receiving light from the Sun and radiating from the upper atmosphere at 0 C, while the surface temperature is 15 C.

- Luckily, the Earth has an atmosphere, which increases the mean Earth surface temperature to 15 C, by acting like an insultaing window connecting to the top of the stratosphere at 0 C.

- Climate sensitivity measures change of temperature vs change of radiative forcing. We know the total radiative forcing from the Sun to be about 270 Watts (per square meter), from the Stefan-Bolzmann Law and also from measurements. We can then say that total climate sensitivity equals $\frac{15}{280} \approx 0.06$ K (squaremeter) per Watt.

- Now IPCC informs us that the extra radiative forcing from doubling of CO_2 should be something like $2-4$ Watts per squaremeter. If we used the total climate sensivity just computed, we can estimate the corresponding global warming to $0.12-0.24$ C.

- We see that IPCC's 4.5 C is $20-40$ time bigger than our result, and thus IPCC must assume very large positive feed back, if started from the basic computation we have made.

So there we stand now: We have using physics and mathematics estimated the crucial climate sensitivity to be certainly less than 0.5 C, which is not alarming at all. Pooh! But is our calculation correct?

IPCC sends out an alarm by suggesting that the climate sensitivity can be 10 times bigger, apparently assuming very large positive feed-back. What is the truth? Your further studies will help to find the answer, from physical laws and numbers. Is global warming a real threat or only imagined? Only science and mathematics can give an answer. Politics and relgion cannot.

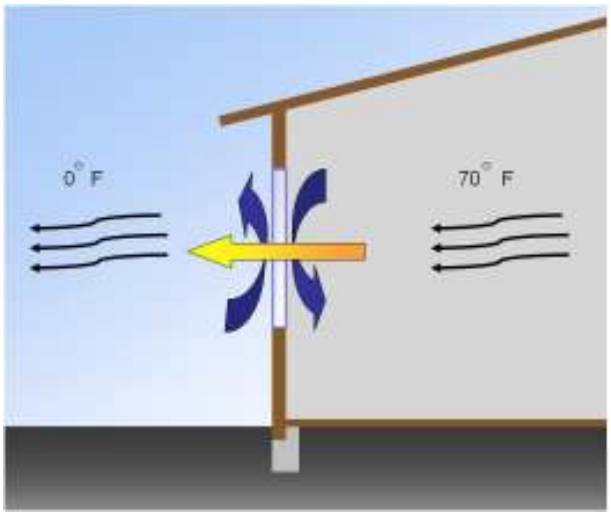


FIGURE 13.3. Does the atmosphere act like a window with a certain U-factor? Does the heat transfer through a window follow Stefan-Boltzmann's Law?

13.2 A More Familiar Example

Suppose you are a poor student living in one-room student dwelling heated by a 280 Watt lamp maintaining 15 C inside at an outdoor temperature of 0 C. A common situation for a student in Sweden.

Suppose now by chance you have acquired an additional heat source of 3 Watts. What increase of the temperature can you expect? Can you see the similarity with the above? Yes, the answer is the same $\frac{15}{280} \times 3 = 0.16$ C. From 15.0 C to less than 15.2 C say. Impossible to detect. Do you see the implication?

The U-factor of a window measures heat transfer per square meter and degree Kelvin. The U-factor of the atmosphere is thus $\frac{280}{15}$. Doubling CO_2 with radiative forcing of 2.8 W would correspond to a decrease of U by 1% and a corresponding global warming of 1% of 15 C, that is 0.15 C.

FIGURE 13.4. Selling windows with small U-factors.

14
Escaping from Ignorance

> Because the world is round, it turns me on; Because the world is round, Ah...
> Because the wind is high, it blows my mind; Because the wind is high, Ah... (The Beatles: Because)

14.1 Space and Time

We start with the intuitive ideas of *space* and *time* we all have: We perceive that everthing there is in the physical world, has a place in some form of big container with three independent directions which we call *space*. We further experience that things can *change shape and position in space*, the rate of which we measure using clocks recording *time* by periodic motion. We will record position in space by x and time by t, where x and t represent numbers. Points in space-time can then be recorded as pairs of numbers (x, t).

We live in three-dimensional or 3d space and a point x can be identified by three coordinates or numbers x_1, x_2 and x_3, which we can collect into a triple (x_1, x_2, x_3) and we can write $x = (x_1, x_2, x_3)$.

If we restrict the world to two dimensions or 2d, that is to a plane, then we need only two space coordinates x_1 and x_2, which we can collect into a pair $x = (x_1, x_2)$.

If we restrict the world further to one space dimension or 1d, that is to a line, then just one coordinate is enough and we have $x = x_1$.

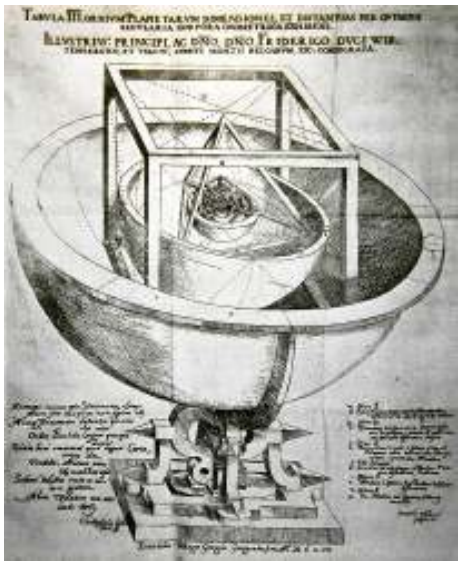


FIGURE 14.1. Keplers model of the planetary system.

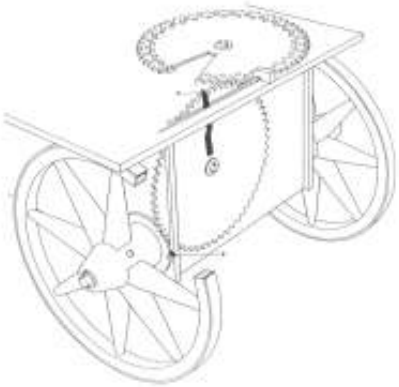


FIGURE 14.2. The hodometer was used by the Romans to measure distances, e.g. along roads. Can you figure out how it worked?.

FIGURE 14.3. The World is a'changing.

We start using *rational numbers* expressed as *decimal numbers* like

$$3.142 = \frac{3142}{1000} = 3 + 10^{-1} + 4 \times 10^{-2} + 2 \times 10^{-3}$$

using the base 10 and denoting here multiplication by $\times$. As usual $10^{-2} = \frac{1}{100}$, and more generally e.g. $10^3 = 10 \times 10 \times 10$, $10^{-3} = \frac{1}{10^3}$.

We recall that a rational number r is the quotient $r = \frac{p}{q}$ of two *integers* p and q (of the form $0, \pm 1, \pm 2, \pm 3, ...$) with $q \neq 0$, and the non-negative integers $0, 1, 2, 3, ...$ are called *natural numbers*.

We denote the set of natural numbers by $\mathbb{N}$, and the set of rational numbers by Q.

14.2 SI Standards of Length and Time

The SI Standard of unit of time is *second*, which is the duration of a certain number of oscillations of a certain caesium atom. More precisely:

- one *second* is the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom at 0 Kelvin.

The SI Standard of unit of length is a *lightsecond*, which is the distance traveled by light in one second, and *meter* as

FIGURE 14.4. 3d vision.

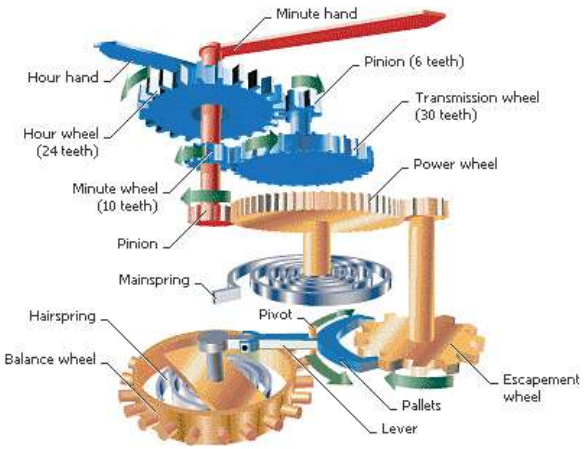


FIGURE 14.5. Mechanical clock with wheels, gears, escapement and weight.

FIGURE 14.6. Entering into the Inner World.

- the distance traveled by light during a time interval of 1/299 792 458 of a second.

14.3 Coordinate Systems in 1d, 2d and 3d

In 1d, for example a horisontal line, we mark the coordinates in meter using a laser beam and a clock measuring the time it takes for light to pass from a given point, which we called the *origin*, to different points to the right and left marking the points to the left with a minus sign.

In 2d, for example a horisontal plane, we choose two perpendicular 1d directions which we mark separately as in 1d.

In 3d we choose three perpendicular directions and mark each direction as in 1d. We can think of these directions as South-North, East-West, down-up.

14.4 The Time Step

We will denote by dt a *smallest unit of time*, which can be different in different situations. The smallest dt we can measure is the time of one oscillation

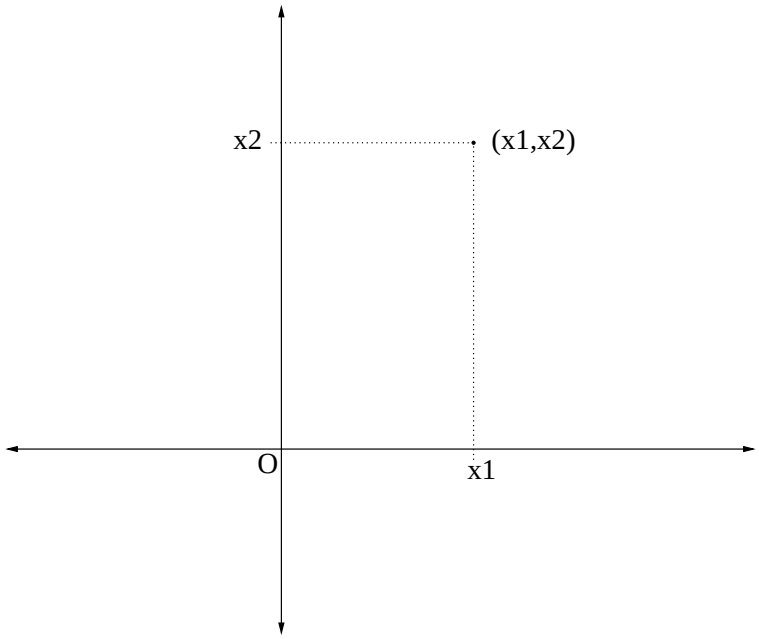


FIGURE 14.7. Coordinate system in the plane.

of a caesium atom, about $0.0000000001 = 10^{-10}$ seconds. Depending on the setting, dt may be one second, minute, hour, day, month, year,...

We will describe as *Newton's World* the physical world with a smallest dt as indicated, and by *Leibniz' World* a mathematical fictional world where dt is assumed to be smaller than any given finite value, or *vanishingly small*.

We shall find that in Leibniz fictional world with a vanishingly small time unit, many mathematical expressions and formulas become easier to manipulate and easier to understand on a conceptual level, than in Newton's real world with a finite smallest time unit.

In computational simulations we have to use a finite time step, since the computer can only perform a finite number of operations per time unit.

We shall use tools from Leibniz world when we construct computational digital simulations of Newton's real analog world, because these are efficient tools, but in the computational simulations effectively use a finite time step or time unit.

Leibniz world is like the ideal world of Plato, which is useful for thinking but untouchable in reality or simulation. IT is like the world we can imagine by using language, which is different from the real world. But we also know that the real world can be more remarkable than any world we may imagine.

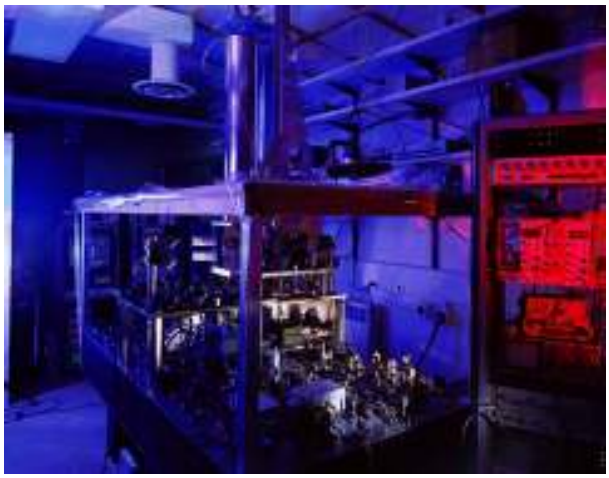


FIGURE 14.8. The Caesium reference clock at NIST Laboratory, Colorado. The Caesium atom is large and vibrates slowly. It also reacts with water.

14.5 Point, Vector and Distance = Vector Norm

Let $x = (x_1, x_2, x_3) \in \mathbb{Q}^3$, where $\mathbb{Q}^3$ is the set of triples (x_1, x_2, x_3) with $x_i \in \mathbb{Q}$, be a point in a 3d coordinate system. We can to the point x associate a *vector* also denoted by x as the directed straight line segment from the origin O to the point x, or arrow from O to x. We write $x \in \mathbb{Q}^3$ also for a vector x.

By *Pythagoras Theorem*, the distance from O to the point x, which is also the length or *norm* $|x|$ of the vector x, is given by

$$|x| = \sqrt{x_1^2 + x_2^2 + x_3^2} \quad \text{or} \quad |x|^2 = x_1^2 + x_2^2 + x_3^2. \tag{14.1}$$

which we can think of as the length of the straight line from the origin to the point x, also referred to as the vector x.

14.6 Scalar Product

If $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3)$ are two vectors in 3d space, that is $x, y \in \mathbb{Q}^3$, then we define their *scalar product* $x \cdot y$ as follows

$$x \cdot y = x_1 y_1 + x_2 y_2 + x_3 y_3 \tag{14.2}$$

We will say that if $x \cdot y = 0$, then the vectors x and y are *orthogonal* or *perpendicuar*. We note that the length of the vector x, or distance from the origin $(0, 0)$ to the point (x_1, x_2) is defined in terms of the scalar product

FIGURE 14.9. Pythagoras 580-495 BC.

as follows:

$$x \cdot x = |x|^2. \tag{14.3}$$

The distance between two points x and y is then given by $|x - y|$, where $x - y = (x_1 - y_1, x_2 - y_2, x_3 - y_3)$.

Further, the *angle* θ between two (non-zero) vectors x and y (with x an arrow with tail at $(0, 0, 0)$ and head at (x_1, x_2, x_3)) is connected to the scalar product $x \cdot y$ by the following formula (which you will derive yourself below):

$$\cos(\theta) = \frac{x \cdot y}{|x||y|}. \tag{14.4}$$

The central quantities of geometry of distance and angle, are thus computable in terms of the scalar product. Neat!

14.7 Change of Position/Time Unit = Velocity

Velocity v is defined as change dx of position x per unit time step dt, that is,

$$v = \frac{dx}{dt}. \tag{14.5}$$

FIGURE 14.10. Nude descending a staircase by Marcel Duchamp.

FIGURE 14.11. Measuring time.

14.8 Change of Velocity/Time Unit: Acceleration

Accelleration a is defined as change dv of velocity v per unit time step dt, that is,

$$a = \frac{dv}{dt}. \tag{14.6}$$

14.9 Particles and Forces

Everything which happens in physical space can be thought of as an interaction between material particles each one occupying a specific point in space at a given time, with the interaction mediated by certain *forces*.

14.10 Newton's 2nd Law: $F = Ma$

The most basic law of physics is *Newton's 2nd Law* stating that

$$F = Ma \tag{14.7}$$

where F is *force*, M is *mass* and a is accelleration. Since $a = \frac{dv}{dt}$ Newton's 2nd law can be written

$$\frac{dv}{dt} = \frac{F}{M}, \tag{14.8}$$

FIGURE 14.12. Mixing real reality and virtual reality by Rene Magritte.

or normalizing to $M = 1$,

$$\frac{dv}{dt} = F. \tag{14.9}$$

This law connects the world of particles, the world of velocities of particles, with the world of forces.

If we think of the world as consisting of particles interacting by forces, we understand that somehow the effect of forces acting on particles must be specified and Newton's 2nd is the basic law making this specification.

14.11 How to Motivate Newton's 2nd Law?

Is it possible to understand why Newton's 2nd Law holds? Or is it simply a definition of force $F \equiv Ma$? Or a definition of mass $M = \frac{F}{a}$? We will return to this question below, when we are prepared to give an answer. As of now, let us accept it and use it in our description of the World.

FIGURE 14.13. Max Planck being struck by the idea of quantum of energy.

14.12 To Think About

- How is length and time measured?
- Is the constancy of the speed of light in vacuum, a definition or physical fact?
- What is the difference between a definition and physical fact? Or is there no difference?
- What is Planck's constant?
- Things That Don't Exist

14.13 Watch

- The Planck Herschel European Space Mission
- What are photons?
- Planck's quantum of energy?

> Time is not a thing, thus nothing which is, and yet it remains constant in its passing away without being something temporal like the beings in time. (Martin Heidegger)

> When modern physics exerts itself to establish the world's formula, what occurs thereby is this: the being of entities has resolved itself into the method of the totally calculable. (Heidegger)

15
Aristotle's Physics

> It is the mark of an educated mind to be able to entertain a thought without accepting it. — All paid jobs absorb and degrade the mind. — The gods too are fond of a joke. — All human actions have one or more of these seven causes: chance, nature, compulsion, habit, reason, passion, and desire. — All men by nature desire knowledge. — For the things we have to learn before we can do them, we learn by doing them. (Aristotle)

Aristotle (384 BC - 322 BC), Greek philosopher, student of Plato, teacher of Alexander the Great, is one of the most important founding figures in Western philosophy. Aristotle's writings constitute a first at creating a comprehensive system of Western philosophy, encompassing morality and aesthetics, logic and science, politics and metaphysics. Aristotle's views on the physical sciences profoundly shaped medieval scholarship, and their influence extended well into the Renaissance, although they were ultimately replaced by Newtonian physics.

The aim of Aristotle's logical treatises was to develop a universal method of reasoning by means of which it would be possible to learn everything there is to know about reality.

To get a perspective on the basic notions of space, time, motion and change, you can amuse yourself by browsing Aristotle's Physics, from which we cite:

- Everything that is in motion must be moved by something.

- The question, what is place? presents many difficulties. An examination of all the relevant facts seems to lead to divergent conclusions. Moreover, we have inherited nothing from previous thinkers, whether in the way of a statement of difficulties or of a solution.
- Now it has three dimensions, length, breadth, depth, the dimensions by which all body also is bounded. But the place cannot be body; for if it were there would be two bodies in the same place.
- What in the world then are we to suppose place to be?
- By asking these questions, then, we must raise the whole problem about place-not only as to what it is, but even whether there is such a thing.
- Time is a measure of motion and of being moved, and it measures the motion by determining a motion which will measure exactly the whole motion, as the cubit does the length by determining an amount which will measure out the whole. Further 'to be in time' means for movement, that both it and its essence are measured by time (for simultaneously it measures both the movement and its essence, and this is what being in time means for it, that its essence should be measured).
- Since time is the measure of motion, it will be the measure of rest too-indirectly. For all rest is in time. For it does not follow that what is in time is moved, though what is in motion is necessarily moved. For time is not motion, but 'number of motion': and what is at rest, also, can be in the number of motion. Not everything that is not in motion can be said to be 'at rest'-but only that which can be moved, though it actually is not moved, as was said above.
- Time is an aspect of change.

15.1 Perspectives

- Aristotle's Physics
- Aristotle's Natural Philosophy

15.2 Watch

- Aristotle's Metaphysics

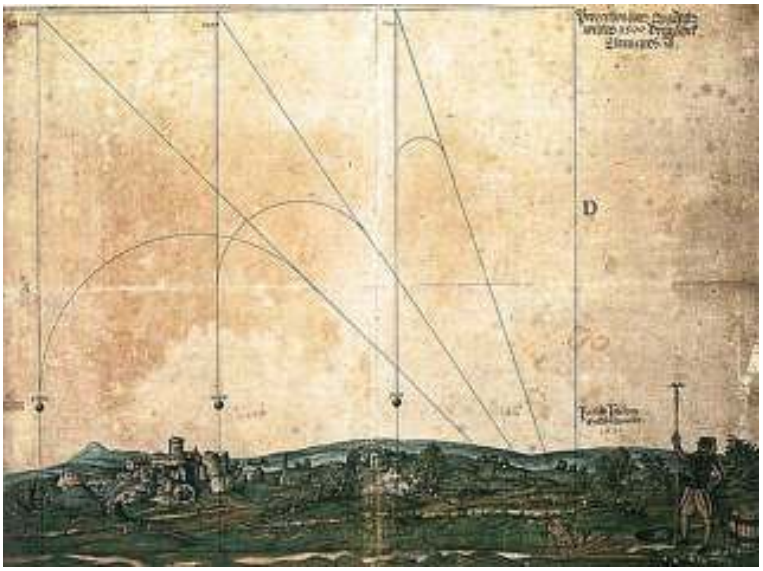


FIGURE 15.1. Trajecories of projectiles according to Aristotle. Correct?

- Aristotle Playing Soccer
- About Scientific Publication and Peer-Review
- All Men by Nature Desire to Know

> The great begins great, manintains itself only through the free recurrence of greatness within it, and if it is great ends also in greatness. So it is with the philosophy of the Greeks. It ended in greatness with Aristotle. (Martin Heidegger)

> The Greeks called the essent as a whole *physis*.We oppose the psychic, the animated, the living to the "physical". But for the Greeks all this belonged to *physis* and continued to do do even after Aristotle. (Heidegger)

> The meaning of *physis* is further restricted by contrast with *techne*, which denotes neither art nor technology, but a knowledge, the ability to play and organize freely. *Techne* is creating, building in the sense of deliberate pro-ducing. (Heidegger)

Part II

Newton's World of Mechanics

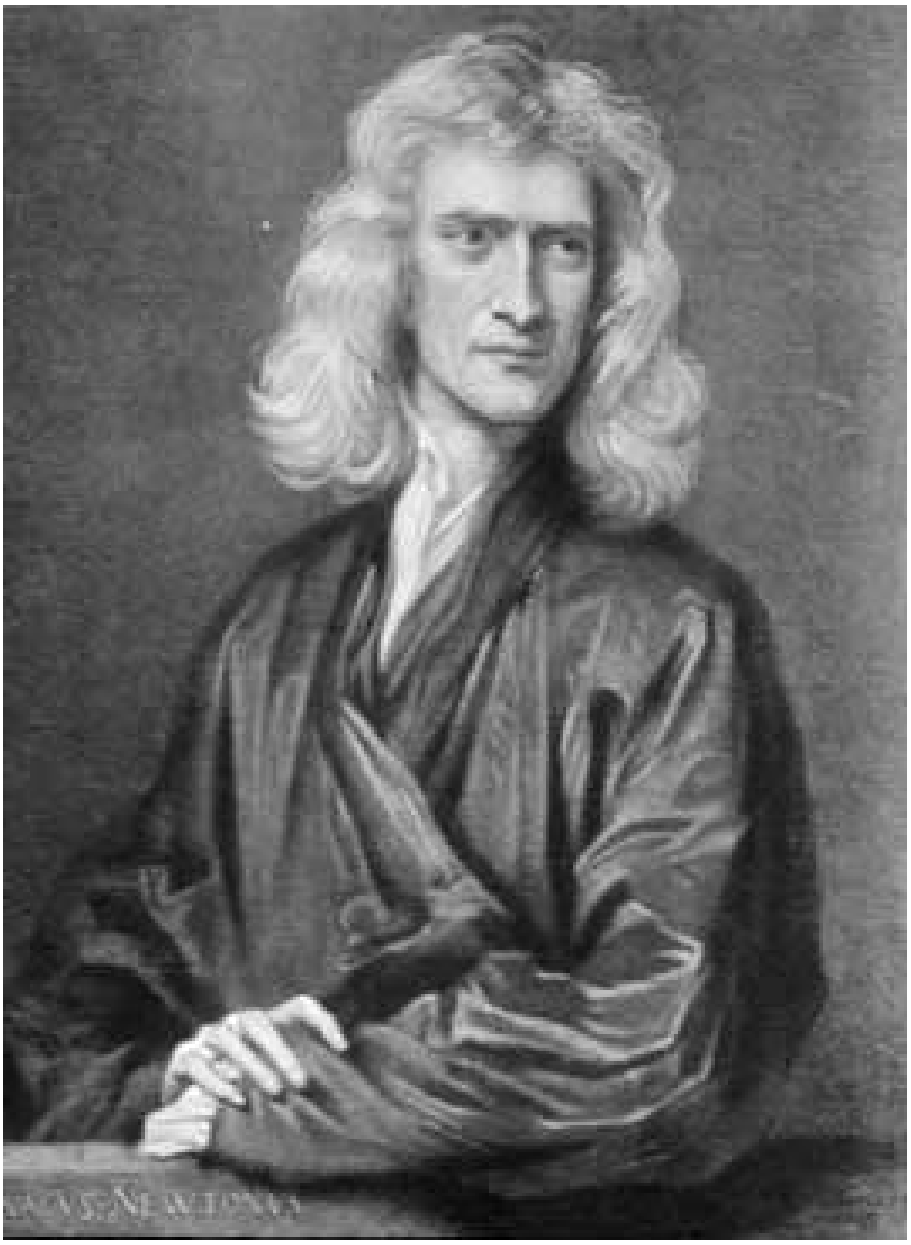


FIGURE 15.2. Newton in 1689: *I was like a boy playing on the sea-shore, and diverting myself now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me....I can calculate the motion of heavenly bodies, but not the madness of people.*

16
Particles Interacting by Forces

> I do not keep up with the details of particle physics. (Murray Gell-Mann, Nobel Prize in Physics 1969 for the standard model of elementary particles)

We start from the idea that the physical world consists of *elementary particles*, such as electrons, protons and neutrons forming atoms and molecules, which interact by four different *forces*: gravity, electromagnetic, weak and strong nucelar force. We will refer to this physical "real world" as *Newton's World of Mechanics*.

In Newton's World we assume that there is a smallest unit of time, or smallest "tick", which we denote by dt. It may *Planck time* with $dt \approx 10^{-43}$ seconds or one oscillation of a caesium atom with $dt = 10^{-10}$ seconds, or something in between of relevance in chemical reactions on atomic scales of size femtoseconds $= 10^{-15}$ seconds . We may also think of larger time units such as year on galactic astronomical scales, although galaxies are made up of atoms. Planck time unit is unimaginably small, seemingly way beyond any physically meaningful scale.

A quartz watch has a time scale of $\frac{1}{32768}$ seconds (compared to the reference of a caesium atom of $\frac{1}{9192631770}$ and rubidium $\frac{1}{6834682611}$).

16.1 Watch

- Newton's 2nd Law
- Newton in 7 minutes

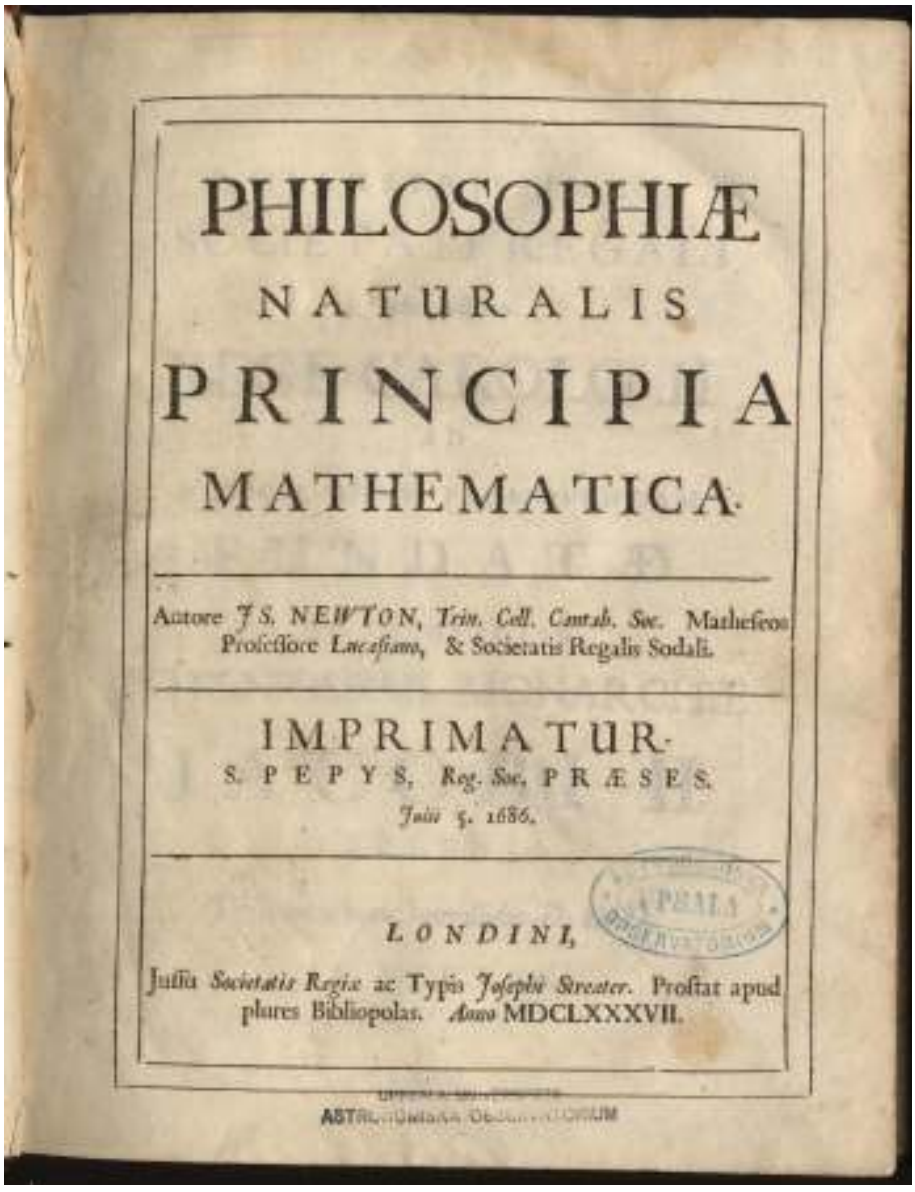
FIGURE 16.1. Particle systems in the form of flocks of birds.

FIGURE 16.2. Making Observations

- Newton: Physics Simulator
- Newton Game Dynamics

16.2 Read More

- Physics Engine
- Open Dynamics Engine
- Interactive Physics Simulation



PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA.

Autore *J S. NEWTON*, *Trin. Coll. Cantab. Soc.* Matheseos Professore *Lucasiano*, & Societatis Regalis Sodali.

IMPRIMATUR.
S. PEPYS, *Reg. Soc.* PRÆSES.
Julii 5. 1686.

LONDINI,
Justu *Societatis Regiæ* ac Typis *Josephi Streater*. Prostat apud plures Bibliopolas. *Anno* MDCLXXXVII.

FIGURE 16.3. Newton's Principia Mathematica.

17
Newton's Laws of Motion

- 1st Law: In the absence of a net force, a body either is at rest or moves in a straight line with constant speed.
- 2nd Law: A body experiencing a force F experiences an acceleration a related to F by F = ma, where m is the mass of the body. Alternatively, force is equal to the time derivative of momentum.
- 3rd Law: Whenever a first body exerts a force F on a second body, the second body exerts a force -F on the first body. F and -F are equal in magnitude and opposite in direction.

17.1 Time-Stepping Newton's Equations of Motion

Newton's World is based on the following *incremental equations of motion* with smallest unit of time dt:

$$dx = vdt, \quad dv = adt, \tag{17.1}$$

as another way of writing

$$\frac{dx}{dt} = v, \quad \frac{dv}{dt} = a, \tag{17.2}$$

which combined with Newton's 2nd Law $F = a$ assuming $M = 1$, take the form:

$$dx = vdt, \quad dv = Fdt, \tag{17.3}$$

or

$$\frac{dx}{dt} = v, \quad \frac{dv}{dt} = F. \tag{17.4}$$

These equations are solved by time-stepping with time step dt:

$$dx^n = v^n dt \quad dv^n = a^n dt, \tag{17.5}$$

where

$$dx^n = x^{n+1} - x^n, \quad dv^n = v^{n+1} - v^n, \tag{17.6}$$

and $x^n = x(ndt)$ and $v^n = v(ndt)$ are the position and velocity at time ndt after n successive time steps with time step dt.

With each tick of time, velocity and position are thus updated according to

$$v^{n+1} = v^n + F^n dt, \quad x^{n+1} = x^n + v^n dt, \quad \text{for } n = 0, 1, ..., \tag{17.7}$$

from given initial values $v(0)$ and $x(0)$ at initial time $t = 0$, where $F^n = F(ndt)$ is the force acting on the body at time ndt. We refer to this update formula as *Euler's method* also called *Forward Euler*.

An alternative update formula is obtained by updating first velocity to v^{n+1} and using this value when updating to x^{n+1}:

$$v^{n+1} = v^n + F^n dt, \quad x^{n+1} = x^n + v^{n+1} dt, \tag{17.8}$$

which we will refer to as *Smart-Euler's method*. You will soon discover the difference between Euler and Smart-Euler.

A variant of Smart-Euler is

$$v^{n+1} = v^n + F^n dt, \quad x^{n+1} = x^n + \frac{1}{2}(v^n + v^{n+1})dt, \tag{17.9}$$

where the mean velocity $\frac{1}{2}(v^n + v^{n+1})$ is used instead of either v^n or v^{n+1}.

Below we shall meet variants with F^n depending on x^{n+1}. The basic method of this form is the *Trapezoidal Method*:

$$v^{n+1} = v^n + \frac{1}{2}(F^n + F^{n+1})dt, \quad x^{n+1} = x^n + \frac{1}{2}(v^n + v^{n+1})dt, \tag{17.10}$$

where $F^n = F(ndt, x^n)$ and $F^{n+1} = F((n+1)dt, x^{n+1})$, which requires iteration because F^{n+1} depends on x^{n+1}, which depends on v^{n+1}.

Below we shall recover Midpoint Euler in the form of the *continuous Galerkin cG(1)*, and *Backward Euler* with v^n and F^n in (17.7) replaced by v^{n+1} and F^{n+1}, as *discontinuous Galerkin dG(0)*.

We also refer to the Trapezoidal Method as *Midpoint Euler*, with Forward and Backward Euler as "Endpoint Euler".

We distinguish between *explicit methods* like Forward Euler with direct update, and *implicit* methods requiring *iteration*, like Midpoint Euler or Backward Euler, where the update formula for v^{n+1} and F^{n+1} is repeated with latest values inserted in the righthand side. With a (small) fixed number of iterations, implicit methods can be viewed as explicit direct update methods.

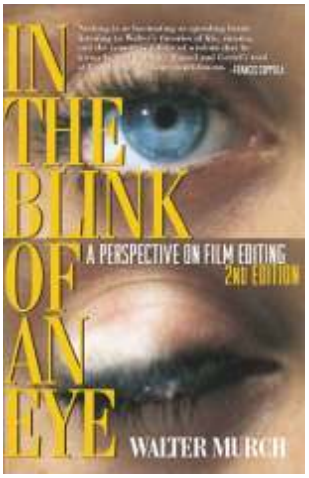


FIGURE 17.1. Eyeblink as time step.

17.2 Basic Solutions of the Equations of Motion

Newton's equations of motion, with given initial position x^0 and velocity v^0, take the form

$$\frac{dv}{dt} = \frac{F}{m}, \quad \frac{dx}{dt} = v \quad \text{for } t > 0, \quad x(0) = x^0,\ v(0) = v^0. \tag{17.11}$$

If $F = 0$, then the solution is given by

$$v(t) = v^0, \quad x = v^0 t + x^0 \quad \text{for } t \geq 0, \tag{17.12}$$

because if $v(t) = v^0$ then $dv = 0$, and if $x(t) = v^0 t$ then $dx = v^0 dt$.

If $F = 2$, $m = 1$ and $v^0 = 0$, then the solution is given by

$$v(t) = 2t, \quad x = t^2 + x^0 \quad \text{for } t \geq 0, \tag{17.13}$$

because if $v(t) = 2t$ then $dv = 2dt$, and if $x(t) = t^2$ then $dx = (t+dt)^2 - t^2 = (t + dt + t)\,dt \approx 2t\,dt$ using the formula

$$a^2 - b^2 = (a + b)(a - b). \tag{17.14}$$

By combination, we thus obtain the following solution formula for the basic case with F constant:

$$v(t) = \frac{F}{m}t + v^0, \quad x(t) = \frac{F}{m}\frac{t^2}{2} + v^0 t + x^0. \tag{17.15}$$

It is important that you understand the derivation of this formula. The key is to understand that $\frac{dv}{dt} = 1$ if $v = t$ and that $\frac{dx}{dt} = 2t$ if $x = t^2$.

17.3 The Fight: Newton vs Leibniz

Leibniz and Newton developed the basics of Calculus independently, in the second half of the 17th century. Newton accused Leibniz for plagiarism backed by the Royal Society of London, which made Leibniz very unhappy in his later years, see Newton vs. Leibniz. Of course, since Leibniz was such an honest scientist, he did not steal anything from Newton. In fact, it is Leibniz' Calculus which is now taught, which is a machine for symbolic and numerical computation with derivatives and integrals, and not Newton's theory of fluxions based on geometric arguments which is very difficult to understand and use.

17.4 Crash Test

- Crash test simulation
- Crash experiment

17.5 Watch

- Newton's Laws
- Newton's 2nd Law
- Conservation of Momentum.

17.6 Conservation of Momentum and Kinetic Energy

The *momentum* m of a body of mass M traveling with velocity v is defined by $m = Mv$. If the body is not acted upon by any force ($F = 0$), then *momentum is conserved*:

$$\frac{dm}{dt} = \frac{d}{dt}(Mv) = M\frac{dv}{dt} = Ma = F = 0 \tag{17.16}$$

If the body is acted upon by a force F, then momentum m changes according to

$$\frac{dm}{dt} = F \tag{17.17}$$

If we multiply this equation by v and interprete $Fv = W$ as *rate of work* W, then we we can write

$$\frac{dk}{dt} = \frac{d}{dt}\frac{Mv^2}{2} = Mv\frac{dv}{dt} = M\frac{dv}{dt}v = Fv = W, \tag{17.18}$$

where $k = \frac{Mv^2}{2}$ is the *kinetic energy*. We here used the fact that $\frac{d}{dt}v^2 = 2v\frac{dv}{dt}$, which we will prove shortly. We conclude that the kinetic energy changes according to

$$\frac{dk}{dt} = W = Fv. \tag{17.19}$$

In particular, if $F = 0$ then the kinetic energy is conserved.

For a system of particles interacting by elastic collisions, total momentum and kinetic energy are conserved if exterior forces vanish, because interior forces and work cancel.

If the velocity is a a vector $v = (v_1, v_2, v_3)$, so is momentum Mv, while kinetic energy K is a number (scalar)

$$k = \frac{M|v|^2}{2} = \frac{M(v_1^2 + v_2^2 + v_3^2)}{2}. \tag{17.20}$$

If we agree to generalize *conservation of momentum* to $\frac{dm}{dt} = F$ and *conservation of kinetic energy* to $\frac{dk}{dt} = W(= Fv)$, then we understand that

- Conservation of momentum is the same as Newton's 2nd Law.
- Conservation of kinetic energy is obtained by multiplying Newton's 2nd Law by velocity.

You will find these insights very helpful below.

17.7 Does Time-Stepping Respect Conservation of Kinetic Energy?

When you start to compute with Forward Euler, Smart Euler and Midpoint Euler, you will find that Forward Euler gains kinetic energy, Smart Euler loses kinetic energy, while Midpoint Euler as a compromise essentially conserves kinetic energy, in problems where kinetic energy should be conserved. You will also discover that the loss and gain decrease with decreasing time step.

We shall meet conservation of energy in a more general context in the next chapter, as conservation of *total energy* as the sum of kinetic energy and *potential/elastic energy*

17.8 To Think About

- How did Newton discover the 2nd Law?
- Who won the War of Calculus, Newton or Leibniz?

FIGURE 17.2. Heraclitus in Raphael's School of Athens: *Pantha rei...Everything flows...There is nothing permanent except change... .Much learning does not teach understanding...No man ever steps in the same river twice, for it's not the same river and he's not the same man...The eyes are more exact witnesses than the ears...Justice will overtake fabricators of lies and false witnesses...Big results require big ambitions...The way up and the way down are one and the same... Man is most nearly himself when he achieves the seriousness of a child at play...Men who wish to know about the world must learn about it in its particular details.* .

- Given the velocities of two elastic spheres about to impact, seek the velocities after impact. Conservation of (total) momentum? Conservation of (total) kinetic energy?

17.9 To Think About: Airbus 340-600

Consider the following data for an Airbus 340-600:

- take-off weight W: 368 tons
- wing area S: 439 square meter
- wing load $\frac{W}{S}$: 8383 Newton/square meter
- sea-level thrust T: 4×25.4 tons
- $\frac{W}{T} = 3.62$

- seats 380.

What is the take-off time and distance? What is the power required at crusing? Why is the engine power given as thrust in tons rather than horse-powers?

17.10 To Think About: Fokker 50

Consider the following data for a Fokker 50:

- take-off weight W: 20 tons
- wing area S: 70 square meter
- wing load $\frac{W}{S}$: 3000 Newton/square meter
- engine power P: 2×2050 kWatts
- $\frac{P}{W}$: 100 Watts/kilo
- crusing speed: 526 km/hour
- seats 50.

How many kW are required at crusing if $F = 10$ (which means that the thrust is 2 tons)? Are the engines oversized (for crusing?)?

17.11 To Watch: Airbus 340-600

- Crash 340-600 April 16 2009.
- Take off
- Emergence landing on Hudson River
- Construction in 116 seconds.

17.12 To Watch: Spitfire

- The story
- Spitfire vs MX2
- Start off

FIGURE 17.3. Airbus 340-600 and Supermarine Spitfires on mission.

- Under bridge

The Supermarine Spitfire is a British single-seat fighter aircraft used by the Royal Air Force and many other Allied countries through the Second World War. Specifications (Spitfire Mk Vb): max weight 3000 kg, engine Rolls-Royce Merlin 45 supercharged V12 engine, 1,470 hp at 9,250 ft (1,096 kW at 2,820 m), max speed 605 km/hour.

17.13 To Think About: Take-Off

To accellerate an airplane of weight W kp from rest to 60 meter/second in 60 seconds, requires an accelleration of 1 meter/second squared, that is a force of W Newton. For a jumbojet of 400 tons a thrust of 40 tons is required (because 1 kp is about 10 Newton), and the length of the starting lane is $\frac{1}{2}60^2 = 1.800$ meters. Doubling the thrust to 80 tons, reduces the time to 30 seconds and the starting lane to 900 meters, which is more realistic. To cruise at a finesse $F = 20$ requires a thrust of $\frac{400}{20} = 20$ tons, about a quarter of the thrust needed for take-off.

Can you figure out how much the length of the starting lane increases if you take into account that the drag increases as velocity squared, and thus the engine power available for acceleration decreases with speed (until the maximum speed is attained and no further accelleration is possible).

17.14 To Think About: Galileo's Experiment

Suppose you drop at the same time a tennis ball and a much heavier similar size pétanque (boule) ball from the Tower of Pisa, like Galileo did? How much quicker will the pétanque ball reach the ground? Compare the Reference Frame. Can you scale the balls so that they fall equally fast?

18
Particle-Spring System

> Fear always springs from ignorance (Ralph Waldo Emerson, American Poet, Lecturer and Essayist, 1803-1882)

Let $x(t)$ be the position at time t of a unit point mass or *particle* moving without friction along a line subject to a linear spring force $F(x) = -x$. See Intro to Springs.

Newton's equations of motion take the form:

$$dx = vdt, \quad dv = -xdt. \tag{18.1}$$

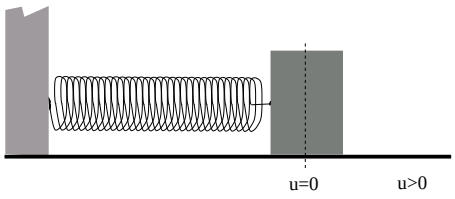


FIGURE 18.1. Particle-spring system: One particle/mass gliding without friction along a line attached to one of a spring attached to a fixed wall: Here $u(t) = x(t)$ is the position at time t measured from some the reference point with zero spring force

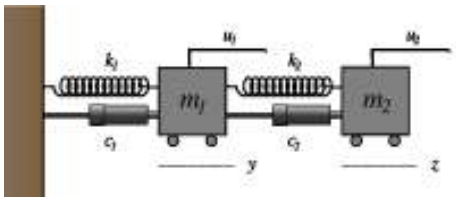


FIGURE 18.2. A 2particle-2spring system with dampers

18.1 Watch

- Particle-Spring
- Particle-Spring Frequency Response
- Particle-Spring Cows on Ice

18.2 To Think About

- How does a spring function?
- How to motivate that spring force is proportional to elongation?

18.3 Conservation of Total Energy

The total energy E a particle connected to a linear spring modeled by $\dot{x} = v$ and $\dot{v} = -x$ is defined by

$$E = \frac{1}{2}(x^2 + v^2). \tag{18.2}$$

Let us now prove that Midpoint Euler conserves the total energy. This follows by multiplying the time-stepping equations

$$x^{n+1} - x^n = \frac{1}{2}(v^{n+1} + v^n)dt, \quad v^{n+1} - v^n = -\frac{1}{2}(x^{n+1} + x^n)dt$$

by $\frac{1}{2}(x^{n+1} + x^n)$ and $\frac{1}{2}(v^{n+1} + v^n)$, respectively, to get by summation and reordering (using that $(a+b)(a-b) = a^2 - b^2$),

$$E^{n+1} \equiv \frac{1}{2}((x^{n+1})^2 + (v^{n+1})^2) = \frac{1}{2}((x^n)^2 + (v^n)^2) \equiv E^n \tag{18.3}$$

which expresses conservation of the total energy as $E^{n+1} = E^n$.

We understand that as the particle moves back and forth, kinetic energy is transformed into elastic energy stored as the spring stretches or compresses, which is transformed back into kinetic energy as the stretching and compression is eased.

19
Planetary System

> I demonstrate by means of philosophy that the earth is round, and is inhabited on all sides; that it is insignificantly small, and is borne through the stars. (Johannes Kepler)

The equations of motion for a planet (viewed as a pointlike particle) of unit mass orbiting a fixed Sun of unit mass centered at the origin, take the form

$$dx = vdt, \quad v = Fdt, \tag{19.1}$$

where

$$F(x) = -\frac{x}{|x|^3} \tag{19.2}$$

is the *gravitational force*. This is a force acting at distance, because the origin is the Sun at the origin, and it acts at x with distance $|x|$ from the origin.

Note that (19.2) is Newton's famous inverse square law of gravitation stating that the magnitude of the gravitational force F between two bodies with mass M_1 and $M-2$ at distance r is given by

$$F = G\frac{M_1 M_2}{r^2}, \tag{19.3}$$

where G is the gravitational constant.

We shall prove below that (19.2) this is a consequence of the fact that the gravitational potential satisfies a certain differential equation named Laplace's equation, and we shall uncover the assumptions leading to Laplace's equation. We can this way motivate that the exponent in Newton's Law is 2 and nothing else.

FIGURE 19.1. Jupiter.

FIGURE 19.2. The Crab nebulosa: A macroscopic particle system.

FIGURE 19.3. Galileo presenting mathematical arguments to disbelieving Catholic priests.

FIGURE 19.4. Galileo's telescope.

19.1 Watch

- Poincaré and the chaos of the three-body problem.

19.2 To Think About

- What are Kepler's Laws?
- What is the simplest solution of a 2-body problem?

19.3 To Read

- BS How to prove Kepler's laws yourself.
- BS Solar System

19.4 Watch

- Kepler's Laws I
- Kepler's Laws II.

20
Local Interaction

> The apple cannot be stuck back on the Tree of Knowledge; once we begin to see, we are doomed and challenged to seek the strength to see more, not less. (Arthur Miller)

When two bodies come in direct contact they may exchange contact forces, without the need of any medium transmitting the force. Contact forces between two bodies in contact are symmetric in the sense that the forces on ecah body have the same magnitude and opposite directions. Contact forces are easy to grasp by intuition as a form of direct contact pressure.

20.1 To Think About

- What is Newton's 3rd Law?
- What is pressure?

20.2 Watch

- Reaction forces
- Saturn V launch
- Apollo 8 launch

FIGURE 20.1. Local interaction.

FIGURE 20.2. Contact.

- Apollo 11 launch
- Challenger Explosion
- ABC Challenger Explosion

21
Action at Distance

> The real lover is the man who can thrill you by kissing your forehead or smiling into your eyes or just staring into space...I believe that everything happens for a reason. People change so that you can learn to let go, things go wrong so that you appreciate them when their right, you believe lies so you eventually learn to trust noone but yourself, and sometimes good things fall apart so better things can fall together. (Marilyn Monroe)

If local contact forces in a sense are easy to envision, *action at distance* by definition is mysterious. If we can see a chain connecting the source/cause to a distant effect, the action at distance can be explained as a form of chain reaction based on repeated local interaction like a row of domino bricks falling one after the other, with each domino brick knocking down the next.

The gravitational and electromagnetic forces are key examples of action at distance because the medium carrying the action seems to be a vaccum or nothingness.

Physicists like to believe that forces between elementary particles are transmitted through certain other particles carrying forces over distance. The gravitational force is conjectured to be transmitted by a hypothetical particle named *graviton*, but nobody has been able to detect a particle like that.

The standard view of gravitation acting at distance is that the presence of a mass, like the Sun, creates a *gravitational field* or *gravitational potential*,

FIGURE 21.1. Action at distance.

the variation of which gives rise to a gravitational force in the same way as a variation of pressure in the air can give rise to a pressure force.

We shall below present an alternative view only based on local interaction without any gravitons, where it is instead the gravitational field which creates the mass. This view is like a hen as gravitational field laying an egg as mass, while the standard view is an egg as mass generating a hen as gravitational potential.

We believe it is more difficult to explain how an egg can create a hen, than how a hen can lay can egg.

FIGURE 21.2. Computing a derivative is like digging where you stand..

21.1 Perspectives

- The Hen and the Egg of Gravitation
- Does the Earth Rotate?

21.2 Local vs Global in Digital Simulation

We shall meet the aspect of local interaction vs action at distance, or local vs global, in both Calculus of derivatives and integrals and in digital computation.

Computation of a derivative of a function is a local operation involving comparison of function values at nearby points in space/time, while computation of an integral of a function is a global operation involving summation over many function values points in space/time which are not close.

Differentiation is like digging a whole where you stand, while integration is like a rumour spreading over distance by mouth-to-mouth communication. In general ntegration requires more computational work than differentiation because information needs to spread. Differention is a local process, while intergration is global.

Derivatives of analytical functions can be computed analytically/symbolically, while integrals in general cannot.

In digital computation the aspect of processing of local vs global information relates to how information is stored in a computers memory, and how fast it can be accessed. The memory storage pattern can reflect physics so that nearby points in space are stored nearby in the memory, but in digital computation action at distance is possible, by addressing any point in the memory. This is like sending information by email instead of by person-to-person mouth-to-mouth.

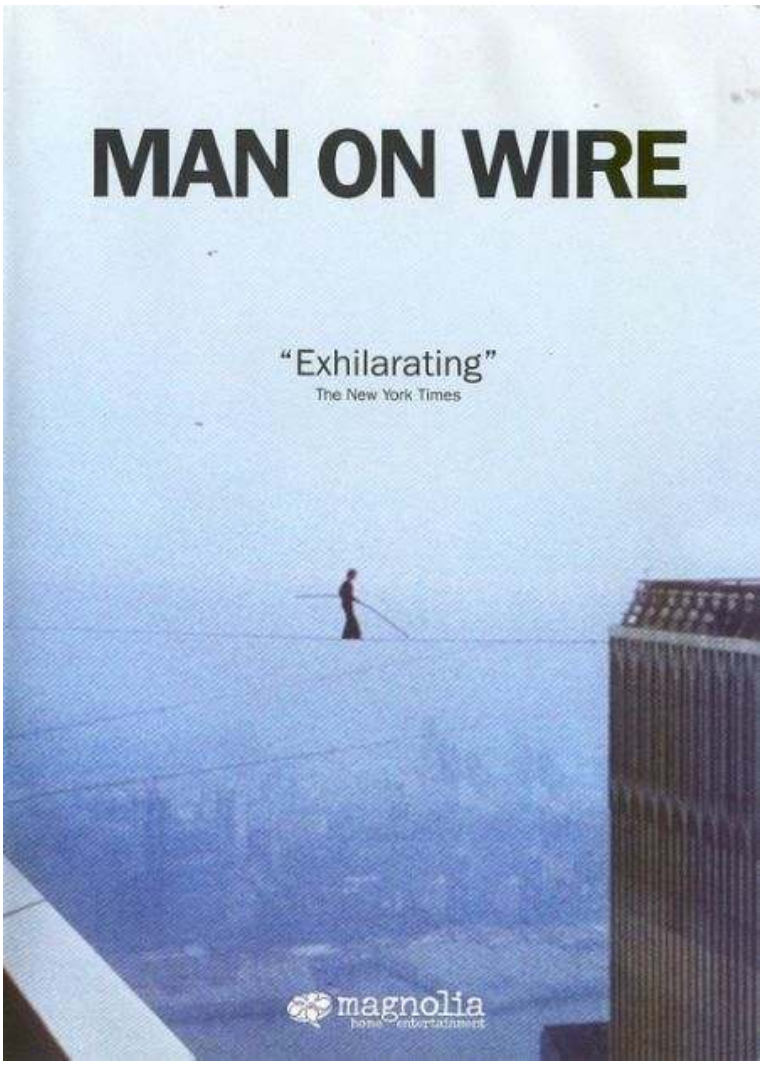


FIGURE 21.3. Computing an integral is like walking from one point to another, step by step.

FIGURE 21.4. Person to person communication reflecting meshpoint-to-meshpoint communication in computational mesh.

FIGURE 21.5. Hermes carrying a message over distance.

In computational digital solution of differential equations, information is processed on a computational mesh reflecting a physical structure. If physical flow of information between nearby material particles in space/time is reflected computatinally by communication only between nearby meshpoints, then the digital flow of information mimics the physical flow and thus can be termed as "physical". But in digital solution also communication between distant points is possible, which as we will see can speed up the computational process, like email communication can speed up communicationby surface snail-mail.

We will meet computational processes with direct meshpoint-to-meshpoint communication in the form of *explicit methods* (of time-stepping), while *implict methods* will involve more or less global communication. Explicit methods are "physical" and "simple" but sometimes slow, while implicit methods are "artificial" and more "complex" but possibly much faster.

FIGURE 21.6. Floating on magnetic forces.

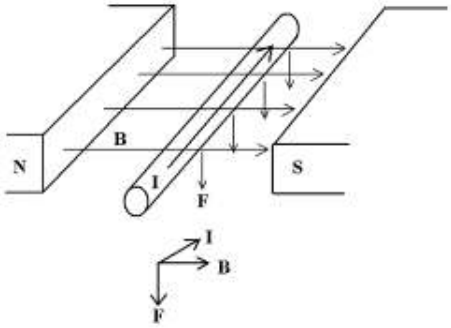


FIGURE 21.7. The mysterious Lorentz force acting perpendicular to electrical current and magnetic field.

21.3 To Think About

- How is gravitational force transmitted?
- How is light transmitted?
- How is an email transmitted?
- How is sound transmitted?

22

Newton: Flight is Impossible!

> If I have seen further than others, it is by standing upon the shoulders of giants... A man may imagine things that are false, but he can only understand things that are true, for if the things be false, the apprehension of them is not understanding... Errors are not in the art but in the artificers...I can calculate the motion of heavenly bodies, but not the madness of people. (Isaac Newton)

Newton computed the lift L of a thin flat rectangular wing of (one-sided) surface area S traveling with velocity V through still air at an angle of attack α (in radians), to be

$$L \approx \alpha^2 V^2 S \tag{22.1}$$

by considering the lower part of the wing to be deviating incoming air downwards in the direction of the wing. In other words, Newton argued that the lift coefficient $c_L = \alpha^2$, so that for $\alpha = 0.1$ as a common angle of attack, $c_L = 0.01$. For a human being of weight 1000 Newton (100 kp) traveling on a wing of surface area $S = 10$ a velocity $V = 100$ meter/second (360 km/hour) would be required. Or with $V = 10$, a wing area of $S = 1000$ square meters would be needed. Impossible! The flight of birds must have been totally inexplicable to Newton. And of course: No hope for Icarus!

22.1 To Think About

- How did Newton argue to come up with (22.1)?

To understand what is correct, one has to also understand what is not correct.

22.2 Kutta and Zhukovsky: Flight is Posssible!

Newton's computation ruled aerodynamics for more than 200 years until the two brothers Wilbur and Orwille Wright in 1903 showed that powered human flight was possible. Newton's lift coefficient was then quicky increased to

$$c_L = 2\pi\alpha \tag{22.2}$$

by the two mathematicians Kutta and Zhukovsky, based on a different mathematical argument. For $\alpha = \frac{10}{180}$ this gave $c_L = 0.3$ and theory and observation was no longer in glaring contradiction.

In reality c_L can be bigger for well designed wings $c_L \approx 18\alpha$ that is with α in degrees $c_L \approx 0.1\alpha$ so that $c_L \approx 1$ for $\alpha = 10$ degrees (100 times bigger than Newton's!!). See The Secret of Flight.

22.3 Flight in a Nutshell

The basics of flight can be summarized in the followin two formulas

$$W = L = c_L V^2 S, \quad P = DV = \frac{L}{F} V, \tag{22.3}$$

where W is weight (Newton), L is lift (Newton), V velocity (meter/sec), S wing area (square meter), P power (Watt), D is drag, and $c_L \approx 0.1\alpha$ with α angle of attack (degrees) is lift coefficient and $F = \frac{L}{D} = 10 - 20$ is finesse coefficient.

With $c_L = 1.0$ we obtain the *wing loading* $\frac{W}{S} = V^2$ ranging from 10 Newton/square meter for a Gossamer Condor, 25 for a common tern, 100 for a wandering albatross, and up to 10 000 for an Airbus 340 at take-off. The velocity ranges from 5 meter per second for a Gossamer Condor, 10 for a starling, over 30 for a Canada goose up to 250 for an Airbus 380.

The quantity $\frac{P}{WV} = \frac{D}{L} = \frac{1}{F}$ measures energy consumption per meter traveled distance and ranges from 0.15 for pigeons, 0.05 for albatrosses and Boeing 747, 0.035 for Lance Armstrong and French TGV, 0.025 for sailplanes.

The power P ranges from 1 Watt for a starling at 10 meter/sec, 450 kWatts for a Spitfire and 200 000 kWatts for a Boeing 747. A human being is capable of 200 Watt.

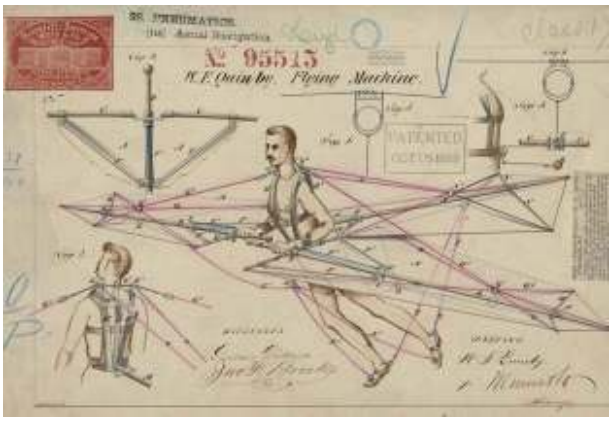


FIGURE 22.1. Patent of Flying Machine 1869.

FIGURE 22.2. Albatross in flight.

23

Computational vs Analytical Mechanics

> The reader will find no figures in this work. The methods which I set forth do not require either constructions or geometrical or mechanical reasonings: but only algebraic operations, subject to a regular and uniform rule of procedure. (Lagrange)
>
> For since the fabric of the universe is most perfect and the work of a most wise Creator, nothing at all takes place in the universe in which some rule of maximum or minimum does not appear. (Euler)
>
> Madam, I have come from a country where people are hanged if they talk. (Euler in Berlin to the Queen Mother of Prussia on his return from Russia)

23.1 Classical Analytical Mechanics

Newton's mechanics initiated the scientific and industrial revolution in the late 17th century. Newton's masterpiece was to solve the 2-*body problem* for a small planet orbiting around a fixed big Sun, showing the orbit to be an ellipse with the Sun in one of the focal points of the ellipse. Newton appeared like a Master of the Universe in charge of all motion according to his immutable marvellous analytical mathematics! Now man was in charge of his fate with the possibility to control the World according to his wishes, if only the analytical mathematics would work out.

But it did not, really: Newton could not even solve the 3-*body problem* with two planets (and nobody else either), not to speak of the N-body problem with $N \geq 3$ bodies, and to keep the illusion of a World governed by analytical mathematics, intense efforts were made by mathematicians including in particular the top stars Euler and Lagrange to reformulate Newton's Laws mathematically as Euler-Lagrange equations characterizing system states as minimizing *Lagrangians* being combinations of potential and kinetic energies. The objective of the reformulation was to describe system states with as few degrees of freedom (variables) as posssible, so that analytical solution could become possible, like for the 2-body problem but with more complicated formulas.

This developed into the discipline of analytical mechanics which has been taught as a core subject in engineering and science education with a long tradition, essentially unchanged during the last 100 (or 200) years.

Analytical mechanics is focussed on rigid-body mechanics, because the motion of a rigid (non-deformable) body can be described with few degrees of freedom, like the position of its moment of inertia and rotation around some axis, if the total mass and moment of interia is known.

High points of analytical mechanics are the 2-body and a spinning top. But problems quickly become very difficult to solve, and various tricks have been developed, which over the centuries have caused head-ache to engineering students.

Analytical mechanics is difficult: special formulation is tricky and the solution work is done by symbolic computation by pen and paper.

A standard classical course in (analytical) mechanics includes:

- Special simple cases of rigid-body mechanics.
- Reformulation of Newton's Laws as Euler-Lagrange equations.
- Tricky combinations of position and angular variables.
- Clever choices of coordinate systems.
- Highly inventive teaching required: Performance by Prof.Levin.
- A typical rigid-body problem.

Contact between rigid-body mechanics is expressed as constraints on motion, which can be tricky to express mathematically. Contact forces between rigid bodies are determined implicitely by global force balance, and thus can be tricky to compute.

23.2 Computational Mechanics

The computer now opens entirely new possibilities to use Newtonian mechanics to model and simulate the World, e.g as a large N-body problem

FIGURE 23.1. Newton playing Master of the Universe.

as in the Millennium Run with $N = 10^{10}$, simply by solving Newton's equations by a computer instead of analytical mathematics. This brings fundamental changes to the teaching, science and engineering practice of mechanics, by changing both the scope and the tools: The analytically impossible N-body problem becomes a simple computational problem, and so it goes:

Basically any thinkable problem of mechanics becomes possible to model and simulate computationally, the only limit being computer power, which is already impressive and continues to increase according to Moore's law with doubling every 18 months.

Computational mechanics is useful: general formulation is not tricky and the solution work is done by computer. BodyandSoul includes a lot of computational mechanics:

- General particle-spring N-body mechanics with N large.
- General deformable-body mechanics or continuum mechanics.
- General continuum fluid-solid mechanics.
- Standard choice of variables in standard coordinate systems.

Contact between deformable bodies can be expressed through local elastic spring forces easily implemented in computational models.

We see that there is little overlap between a classical analytical mechanics course (special rigid-body) and modern computational mechanics (general deformable-body fluid-solid). Of course the general essentially covers also the special: Classical building is collapsing...

23.3 Perspectives

Take a look at:

- Virtual cat walk
- Crash test
- Dummy crash test
- Volvo S80 crash test
- Airbag simulation and experiment
- Real vs virtual testing?
- Earth quake simulation
- Shake-out earth-quake simulation
- Big simulation
- What is water?
- Black hole terror.

And recall the Circus Cow:

- Flying Circus Cow: Exterior
- Flying Circus Cow: Interior

showing the essence of the argument.

23.4 Looking Forward

- Session: Analytical Mechanics

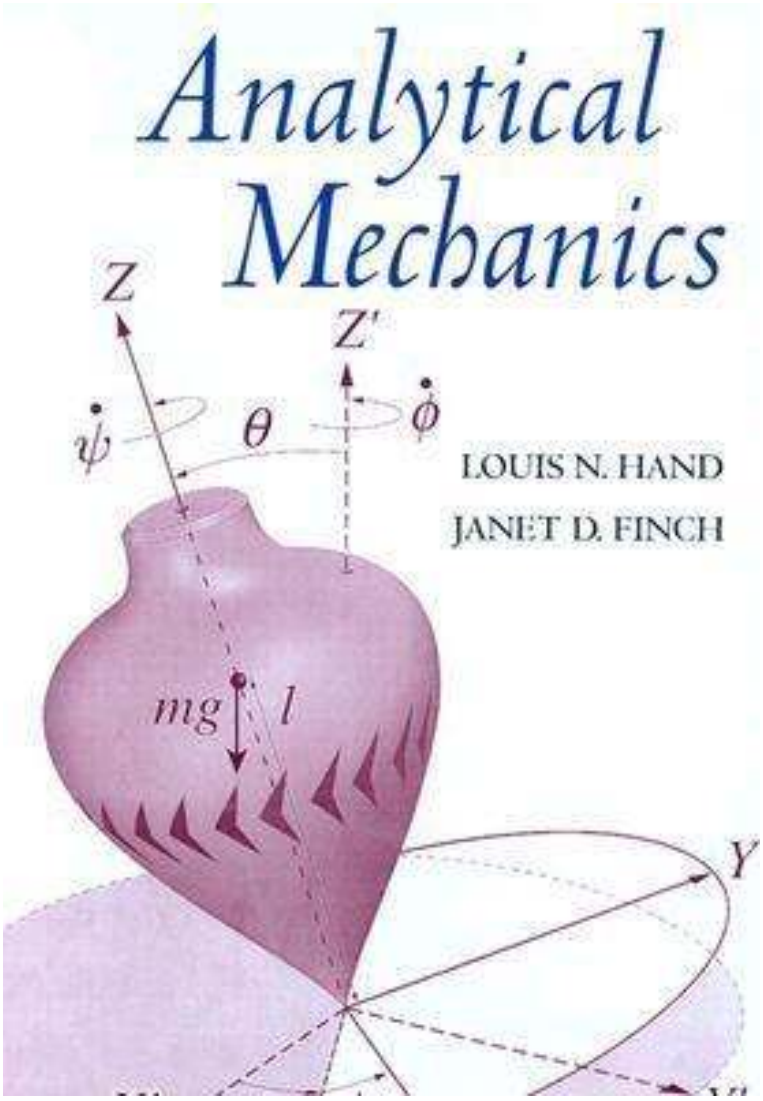


FIGURE 23.2. High point of classical analytical mechanics: Spinning top!

Part III

World of Games

24
Creating Virtual Worlds

> In virtual reality or "la realite virtuelle", characters, objects, and images take on the phantasmagoric force of alchemy's visionary internal dramas. (Antonin Artaud, French playwright, poet and actor)

24.1 Interactive Virtual Worlds as Games

There are different types of games:

- between people based on real physics: soccer, tennis...
- between people based on virtual physics: computer games: World of Warcraft...
- between people based on non-physics: chess, cards...

All games are based on *interaction* between players and the elements of the game (balls, cards, guns...).

Computer games are now booming as they offer interesting challenging environments at small costs as compared to playing real games with real guns and real sweat.

Computational mathematics is an efficient tool for constructing interactive virtual worlds, and therefore serves as the engine of computer games.

FIGURE 24.1. Cave Painting Virtual World in the Lascaux Cave.

An area of science can be viewed as virtual physics created by scientists and used by scientists to discover truths about real physics. Scientist interact with their virtual physics models by giving input and studying the output.

A scientist interacting with a virtual physics model can be viewed as playing a form of game with the scientist giving input to the model in response to output from the model, with the purpose of obtaining maximal information, a game in which also other scientists participate. Often a tough merciless game...

Science and technology are like computer games booming as computational mathematics allows the construction of affordable virtual physical worlds as an alternative to expensive real experimental labs.

As a student of Simulation Technology you are certainly interested in mathematics/science/technology as virtual physics, but your interest in playing games may vary.

In any case it is useful to view virtual physics as games, because the input-output aspect of the underlying mathematical model then has to be made clear.

You will now meet this approach in your studies of the real world, in a sequence of progressively more complex games based on virtual physics, starting with the most simple and arriving at surprisingly complex physics.

To construct virtual physics game you need to specify the physics involved (e.g. Newton's 2nd Law) and the rules for interaction. This is like constructing a machine with certain control knobs.

The game is played using a computer to run the machine and by using the knobs to control the machine to specific goals.

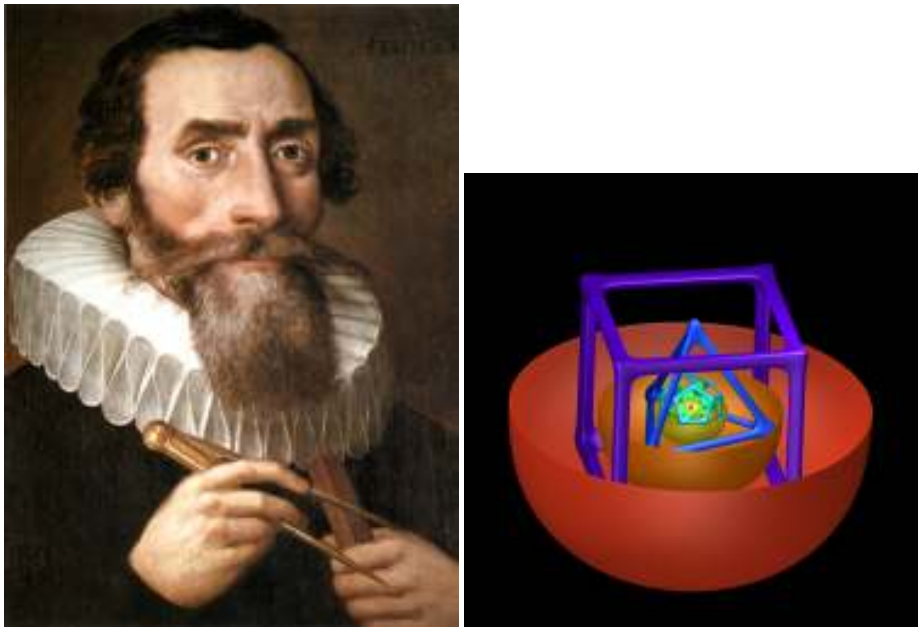


FIGURE 24.2. A scientist with a copy of his virtual physics model.

To construct computer games based on virtual physics you need knowledge/skills in

- mathematics: calculus, linear algebra and geometry,
- physics: Newton's laws, spring forces,
- programming.

In this chapter you will be directly confronted with these aspects in a progression from the most simple to the more complex, without noticing how "difficult" and understanding how "advanced", it is in fact. In sections entitled *Demo + Lab* you find material allowing you to

- play a game
- inspect the computer code of the game
- modify the code
- write your own code for your own game variants.

24.2 Python

The demo codes are written in Python. You find an introduction to Python here.

FIGURE 24.3. The (Ultimate) Game of Life against Death.

24.3 Simulation

You will find that a few basic principles allows you to simulate increasingly realistic simulations such as:

- Circus Cow.
- Flying Bird
- Volvo Car
- Flapping Flag

You will learn below how this is done (mathematics + programming) and how to improve realism and add interactivity. With this knowledge you be a scientist, engineer, teacher, computer game inventor and more...

24.4 Geometry Preparation

To get some know-how about 2d and 3d space, browse

- Analytic Geometry in 2d.
- Analytic Geometry in 3d.

24.5 Watch

- Best computer games 2009

FIGURE 24.4. From World of WarCraft.

FIGURE 24.5. Heavy Rain trailer.

- Computer game design
- Make games online
- Interactive Python programming in 1 min

24.6 Another Story: Heavy Rain Gameplay

Who will write the first Icarus Gameplay?

25
Homo Ludens: Playing Man

25.1 Homo Ludens and Homo Faber

Homo Sapiens (Knowing Man) has many other names indicating different characters or qualities including:

- Homo Faber: Making Man
- Homo Ludens: Playing Man
- Homo Amans: Loving Man
- Homo Politicus: Political Man
- Homo Ridens: Laughing Man
- Homo Technologicus: Technological Man
- Homo Discens: Learning Man
- Homo Investigans: Investigating Man
- Homo Musicus: Man Playing Music
- ...

As a student of simulation technology you get a chance to develop several of these qualities of your personality: constructing games as Homo Faber and Technologicus and playing games as Homo Ludens...and more...

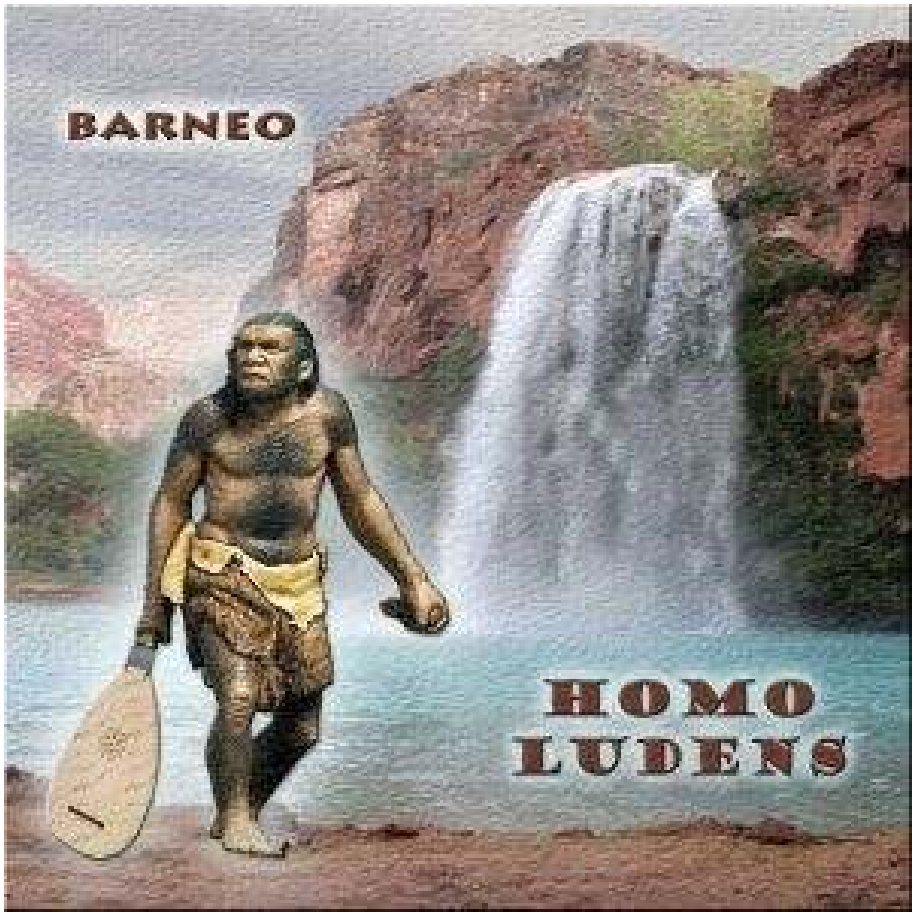


FIGURE 25.1. Early Playing Man.

25.2 Huizinga

Johan Huizinga (1872-1945) suggested in his book Homo Ludens that play is primary to and a necessary (though not sufficient) condition of the generation of culture. From the Foreword:

- *A happier age than ours once made bold to call our species by the name of Homo Sapiens. In the course of time we have come to realize that we are not so reasonable after all as the Eighteenth Century, with its worship of reason and its nave optimism, thought us; hence modern fashion inclines to designate our species as Homo faber: Man the Maker. But though faber may not be quite so dubious as sapiens it is, as a name specific of the human being, even less appropriate, seeing that many animals too are makers. There is a third function, however, applicable to both human and animal life, and just as important as reasoning and making namely, playing. It seems to me that next to Homo Faber, and perhaps on the same level as Homo Sapiens, Homo Ludens, Man the Player, deserves a place in our nomenclature.*

A central idea of Huizinga is thus that all culture is rooted in play, but he carries the analysis further by noting that as culture evolves or is refined the its playful element is suppressed. This can be seen in the transition from folk and jazz music allowing musicians to play following their own inspiration, to classical music dictated by composers and conductors with limited freedom of interpretation.

Huizinga identifies the following characteristics of play:

1. *Play is free, is in fact freedom.*
2. *Play is not "ordinary" or "real" life.*
3. *Play is distinct from "ordinary" life both as to locality and duration.*
4. *Play creates order, is order. Play demands order absolute and supreme.*
5. *Play is connected with no material interest, and no profit can be gained from it.*

You may amuse yourself by playing with the following ideas by Huizinga:

- *Play is a uniquely adaptive act, not subordinate to some other adaptive act, but with a special function of its own in human experience.*
- *It is the goal of the American university to be the brains of the republic.*
- *If we are to preserve culture we must continue to create it.*
- *History is the interpretation of the significance that the past has for us.*

FIGURE 25.2. Kermesse (1567-8) by Pieter Brueghel

- *History can predict nothing except that great changes in human relationships will never come about in the form in which they have been anticipated*

- *In Europe art has to a large degree taken the place of religion. In America it seems rather to be science.*

- *A superstition which pretends to be scientific creates a much greater confusion of thought than one which contents itself with simple popular practices.*

- *Culture means control over nature.*

- *Culture must have its ultimate aim in the metaphysical or it will cease to be culture.*

- *Systematic philosophical and practical anti-intellectualism such as we are witnessing appears to be something truly novel in the history of human culture.*

- *You only live a short time... and you are dead a long time.*

- *The second fundamental feature of culture is that all culture has an element of striving.*

25.3 Roger Caillois

Roger Caillois (1913-78) builds on the theories of Johan Huizinga and disputes many of them, adding a more comprehensive review of play forms:

- The Game Design Reader
- Man, Play and Games
- Man, Play and Games on Google Books

25.4 A Flavor of Mathematical Game Theory

- Mathematical Game Theory: John Nash
- Theorie des Jeux par Pierre-Louis Lions

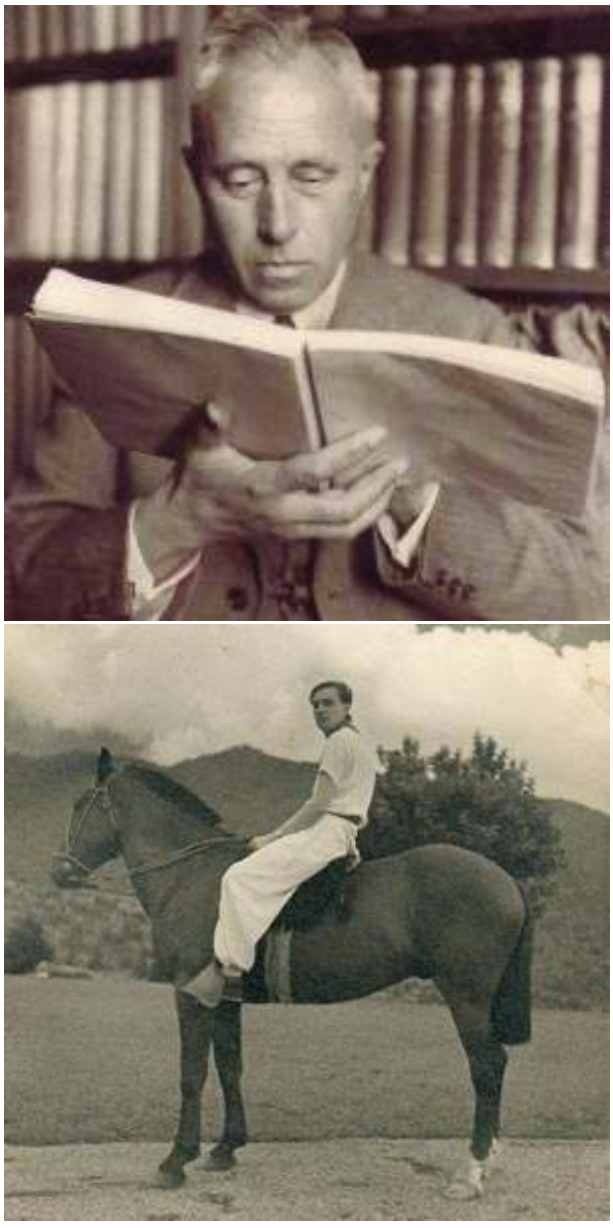


FIGURE 25.3. Huizinga and Caillois.

26
Pong 1d

Pong (marketed as PONG) is one of the earliest arcade video games, and is a tennis sports game featuring simple two-dimensional graphics. The aim is to defeat an opponent either computer-controlled or a second player by earning a higher score. The game was originally manufactured by Atari Incorporated (Atari), who released it in 1972. Pong was created by Allan Alcorn as a training exercise assigned to him by Atari founder Nolan Bushnell. Bushnell based the idea on an electronic ping-pong game included in the Magnavox Odyssey, which later resulted in a lawsuit against Atari. Surprised by the quality of Alcorn's work, Atari decided to manufacture the game.

26.1 Game

Two players interact with a ball free to move in the interval $[0, 1]$ of a 1d coordinate axis. One of the players can reverse the velocity by clicking the mouse when the ball hits $x = 0$ and the other at $x = 1$. If the mouse is clicked to early or too late the game ball is lost.

26.2 Mathematics

Assume the ball moves with constant speed v. The equation of motion is dx=vdt as long as the ball is inside the interval $[0, 1]$. If $x((n+1)dt) < 0 < x(ndt)$ or $x(ndt) < 1 < x((n+1)dt$, then a player can reverse the direction of motion, by a mouse click.

26.3 Realization

- Update position by $x^{n+1} = x^n + vdt$.
- Reverse the sign of v by mouse click when $x((n+1)dt) < 0 < x(ndt)$ or $x(ndt) < 1 < x((n+1)dt$.

26.4 Demo + Lab

- Test, Modify and Create Yourself (pong1d)

26.5 Generalization

1. Given variable speed $v(t)$.
2. Add given force $F(t)$ and update $dv = \frac{F}{M}dt$, $dx = vdt$.

26.6 Perspective

- Short Course of Calculus.

27
Pong 2d and 3d

27.1 Game

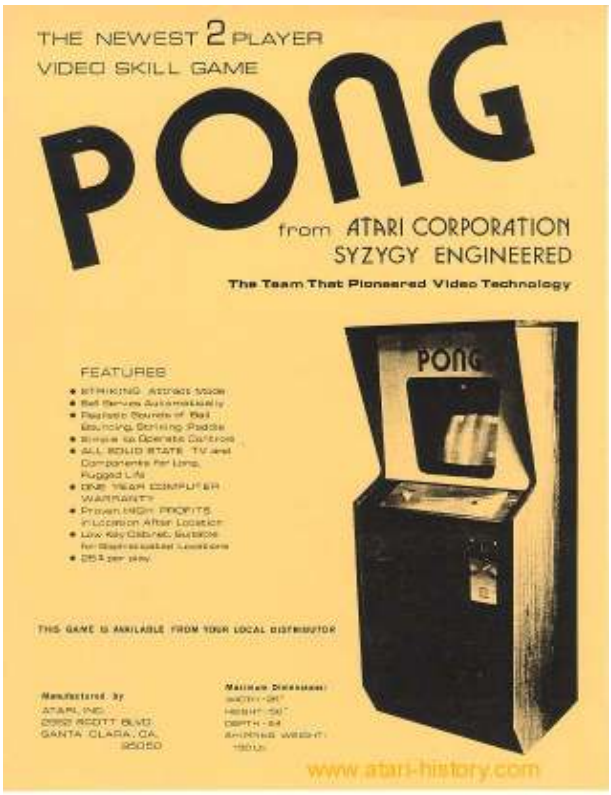
Two players interact with a ball of unit mass free to move inside the unit square $Q = [0,1] \times [0,1]$ of a 2d plane by reflection. At reflection the appropriate velocity component switches sign. One of the players can reverse the velocity by clicking the mouse when the ball hits $x_1 = 0$ and the other at $x_1 = 1$. If the mouse is clicked to early or too late the game ball is lost.

27.2 Mathematics

Assume the ball moves with constant speed $v = (v_1, v_2)$. The equation of motion is dx=vdt as long as the ball is inside Q. If $x_1((n+1)dt) < 0 < x_1(ndt)$ or $x_1(ndt) < 1 < x_(n+1)dt)$, then a player can return the ball by reversing v_1 by a mouse click by also moving a cursor moving in the x_2-directionto the point of reflection.

27.3 Realization

- Update position by $x^{n+1} = x^n + vdt$.
- Reverse the sign of v by mouse click when $x_1((n+1)dt) < 0 < x_1(ndt)$ or $x_1(ndt) < 1 < x - 1((n+1)d)t$, with an x_2 cursor at the point of reflection.



27.4 Demo + Lab

- Test, Modify and Create Yourself 2d (pong2d)
- Test, Modify and Create Yourself 3d (pong3d)

27.5 Generalization

1. Same in 3d.
2. Variable speed $v(t)$.
3. Add force $F(t)$ and update $dv = \frac{F}{M}dt$, $dx = vdt$.

28
Viscous Pong

28.1 Mathematics

Newton's 2nd Law for a particle subject to a viscous (friction) force $F = -\mu v$ in the opposite direction to motion, takes the form

$$M\dot{v} = -\mu v \quad \text{or } \dot{v} + \mu v = 0, \tag{28.1}$$

where μ is a non-negative viscosity coefficients and v velocity. The corresponding velocity update formula is (with mass $M = 1$) given by

$$v^{n+1} = v^n - \mu v^n dt = (1 - \mu dt)v^n. \tag{28.2}$$

Without additional force the viscous force will act as damping and will eventually reduce the velocity to zero. With small viscosity μ, this will take many time steps.

28.2 Demo + Lab

- Test, Modify and Create Yourself (viscouspong)

FIGURE 28.1. Lucifer being expelled from Paradise. Why?

29
Pendulum

> Any person, brought into the presence of this fact, stops for a few moments and remains pensive and silent; and then generally leaves, carrying with him forever a sharper, keener sense of our incessant motion through space... The phenomenon develops calmly, but it is invisible, unstoppable. One feels, one sees it born and grow steadily; and it is not in one's power to either hasten or slow it down. (Leon Foucault 1819-1868)

29.1 Game

Two players interact with a pendulum by controling the direction, magnitude and duration of a force. The players have the same finite total amount of force times duration, with e.g. the objective of getting the pendulum at a given final time to rotate in a specific direction: e.g clockwise for player 1 and counter-clockwise for player 2.

We think of a pendulum as a mass connected to one end of a stiff weightless arm with the other end fixed at a frictionless hinge. The force is applied to the mass.

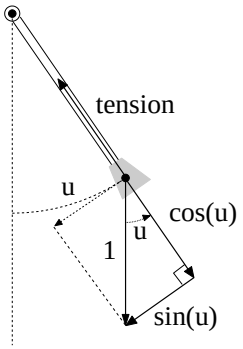


FIGURE 29.1. Pendulum

29.2 Mathematics

Consider a pendulum length L and mass M free to rotate around the origin. Let the position at time t be given by the angle $u(t)$ measured from the vertical with $u = 0$ in the rest position of the pendulum (with the mass under the hinge).The equation of motion is given by Newton's 2nd Law as follows, without player force,

$$L\dot{u} = v, M\dot{v} = -MG\sin(u(t))$$

since the velocity v is given by $L\dot{u}$ and the force perpendicular to the pendulum axis, in the direction of circular motion, is given by $-Mg\sin(u(t))$, where G is the gravitational constant. Assuming $L = G = 1$, the equations of motion thus are

$$\ddot{u}(t) + \sin(u(t)) = F(t) \quad \text{for } t > 0, \tag{29.1}$$

with given initial conditions $u(0)$ and $\dot{u}(0) = 0$, and where $F(t)$ is the force from the players.

29.3 Realization

$$du = vdt, \quad dv = -\sin(u)dt + Fdt \tag{29.2}$$

29.4 The Inverted Pendulum

The equations for the inverted pendulum are the same. What is the difference in mechanical action?

FIGURE 29.2. Foucault's Pendulum.

29.5 Demo + Lab

- Test, Modify and Create Yourself (pendulum)

30
Double Pendulum

Already a two-body problem in the form of a double pendulum, has a very complex dynamics:

- double pendulum
- double pendulum experiments
- Rondo in C
- Two-legged double pendulum

The standard way to model a double pendulum is to consider the two arms to be rigid (with fixed lengths) with a friction-free joint and to describe the position in terms of two angles as indicated in Fig. 30. But it is not so easy to write down the equations of motion in these coordinates, which you will discover if you try.

A more direct way is to assume that the two arms act like stiff springs and to model the system like a two-particle two-springs system. The equations of motion can then formulated directly and simulations can star directly.

Classical analytical mechanics focussed on rigid bodies described by few cleverly chosen (angular) coordinates or independent variables, and required tricky mathematics.

In computational mechanics you can afford to go beyond rigid body mechanics and use particle-spring models or 3d elasticity models for which the equations of motion are given once and for all. Computational mechanics is thus both more useful and easy to understand, than classical analytical rigid-body mechanics.

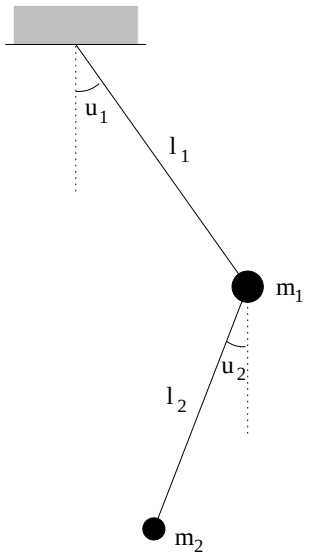


FIGURE 30.1. Coordinates for double pendulum as rigid body system.

30.1 Demo + Lab

- Test, Modify and Create Yourself (doublependulum)

30.2 Game

Construct games based on a double-pendulum. Extend to multiple-pendulum.

30.3 Read More in BS

- Lagrange and the Principle of Least Action
- Double Pendulum as Rigid-Body System

31
Tour de France

> When I got off the bike at the end of the Tour, it was finished...Pain is temporary. Quitting lasts forever. (Lance Armstrong after Tour de France)

31.1 To Read

- Newton's Law of Motion

31.2 Game

Consider a race between two bikers travling on a varying level horisontal straight road from A to B subject to a variable wind speed. Suppose each biker has the same total energy to spend on a race over the distance AB.

31.3 Mathematics

The laws of motion for each biker are (with an approximate horisontal momentum balance):

$$\dot{x} = v, \quad \dot{v} = F - c_D v^2 - w(t) - g(x) \tag{31.1}$$

FIGURE 31.1. Lance Armstrong fighting wind and rain.

where $v(t)$ is the bike velocity, $w(t)$ is a wind force, c_D is a drag coefficient, $g(x)$ is the component of the gravitational force parallel to road elevation curve, and $F(t)$ the force supplied by the biker. The side condition is that

$$\int_0^T F(t)v(t)\,dt = E, \tag{31.2}$$

where E is the total energy. We assume that $g(t)$ is given to the biker, but that the wind drag $w(t)$ is subject to stochastic fluctuations. Each biker seeks to get from A to B in minimal time with the resource E.

32
The Wright 1903 Flyer

32.1 To Read

- Human Powered Aircraft
- Newton's Laws of Motion
- Flight Is Possible

32.2 Game

Construct a flight game based on given lift and drag curves for a wing.

32.3 Mathematics

Suppose the total lift L and drag D of an airplane, like the Wright 1903 Flyer, are given by

$$L = c_L \alpha v^2 S, \quad D = \frac{L}{F}, \tag{32.1}$$

where $c_L \approx 10$ is a given lift coefficient, $F = \frac{L}{D} \approx 10$ is a given finesse coefficient, $v \approx 10$ is the airplane velocity in meter/sec, $S \approx 50$ is wing area in square meters, $\alpha \approx 0.1$ is the angle of attack of the wing. The

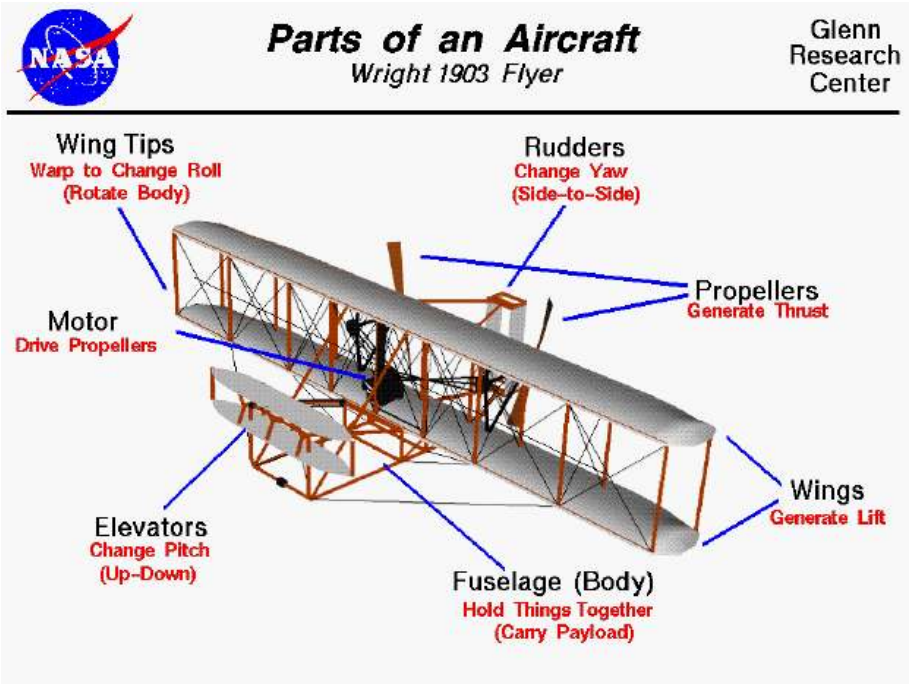


FIGURE 32.1. The 1903 Flyer.

equations of motion in a $(x-1, x_2)$-coordinate system with x_2 vertical and x_1 horisontal, are the following:

$$\frac{dx}{dt} = v, \quad M\frac{dv}{dt} = (F - D + f, L - W) \tag{32.2}$$

where F is the engine force (assumed to be horisontal for simplicity) in Netwon, $W \approx 3000$ Newton is the total weight of the airplane and $M \approx \frac{W}{10}$ the mass in kg, and f is a given variable head wind force in Newton.

The total energy spent on a trip equals $\int Fv\,dt$, and the objective may be to cover a certain distance in shortest time with a given amount of energy to spend with a given engine, possibly with requirements of reaching certain altitudes on the way.

33
Americas Cup 1851

33.1 To Read

- Flight Is Possible
- Newton's Laws of Motion

33.2 Game

Construct a sailing game based on given lift and drag curves for sail and keel for different angles of attack. Start with motion against the wind with close-hauled sails parallel to the boat centerline.

33.3 Mathematics

Define

- $\bar{\alpha}$ true angle of incoming wind vs boat centerline,
- α apparent wind angle = angle of attack of (close-hauled) sails,
- β angle between boat centerline and direction of motion = angle of attack of keel,
- v boat speed vs water,

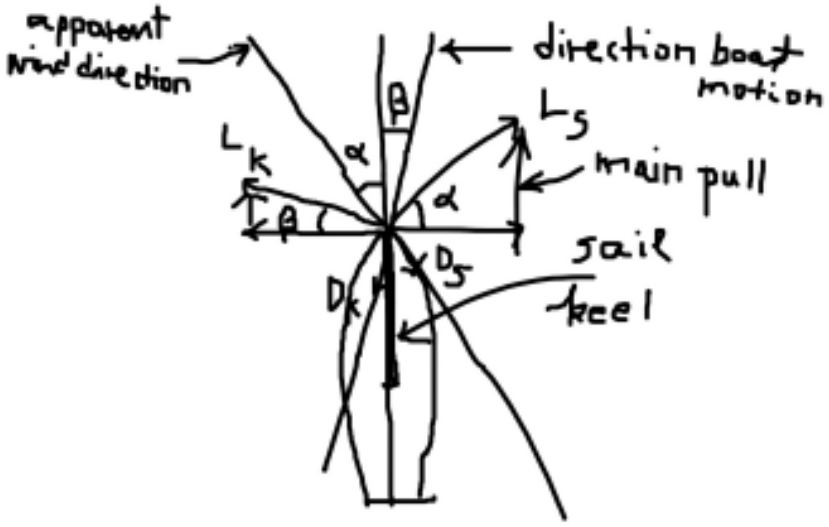


FIGURE 33.1. Lift Forces L_S and L_K, and drag forces D_S and D_K, from sail and keel.

- w apparent wind speed.

Assume the boat is currently pointing in the x_2-direction of a (x_1, x_2) coordinate system. Force balance of lift L_S and drag D_S from sail and lift L_K and drag from keel D_K in the coordinate directions give (under stationary conditions without accelleration):

$$\begin{aligned} L_s \sin(\alpha) + L_K \sin(\beta) + D_2 &= D_S \cos(\alpha) + D_K \cos(\beta) \\ L_S \cos(\alpha) + D_1 \sin(\alpha) + D_1 &= L_K \cos(\beta) + D_K \sin(\beta), \end{aligned} \tag{33.1}$$

where (D_1, D_2) is the drag force from boat hull in water (in the (x_1, x_2)-system), assuming there is no drag from hull and rigg in air. Assume

$$\begin{aligned} L_S &= c_S \alpha (w + v \sin(\alpha + \beta))^2, \quad D_S = \frac{L_S}{F_S}, \\ L_K &= c_K \beta v^2, \quad D_K = \frac{L_K}{F_K}, \\ D_2 &= c_2 (v \cos(\alpha))^2, \quad D_1 = c_1 (v \sin(\alpha))^2, \end{aligned} \tag{33.2}$$

with c_S and c_K lift coefficients (including areas and fluid densities), and F_S and F_K lift/drag are finesse factors for sail and keel. Of course, the apparent wind velocity can be computed from the true wind and boat velocity.

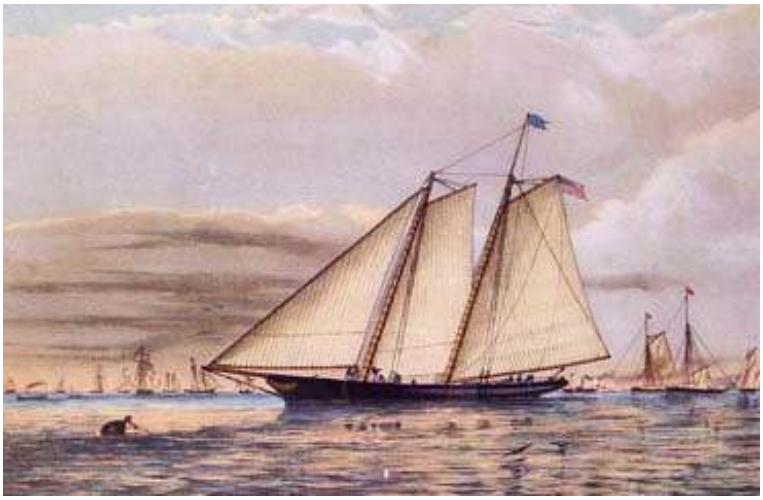


FIGURE 33.2. *America* at the first Americas Cup 1851.

The sailor seeks to control the direction of the boat vs the true wind so that the boat advances as quickly as possible in a given direction, with close-hauled sails typically in the direction opposite to the wind.

During a game the true wind changes speed and direction and the helmsman is supposed to chaange the direction of the boat accordingly.

Note that the main pull forward comes from the component $L_S \sin(\alpha)$ of sail lift L_S, which is to balance drag from sail, keel and hull, with the heeling force $L_S \cos(\alpha)$ roughly balanced by the keel force component $L_K \cos(\beta)$. Typical values are $\alpha = 20$ and $\beta = 10$ (degrees), $F_S = 10$ and $F_K = 20$.

34
Planetary Slalom

34.1 Game

Move a space rocket of unit mass from given start position and velocity to a given end position under the influence of gravitational forces from a given set of masses at fixed positions, using as short time as possible with a given total amount of energy W for rocket engines.

34.2 Mathematics

$$dx = vdt, \quad v = (F + f)dt, \tag{34.1}$$

where F is the combined gravitational force from the masses and f is the force from the rocket engines, under the side condition

$$\int_f dx \leq W \tag{34.2}$$

expressing that the total work $\int_f dx$ does not exceed W.

34.3 Demo + Lab

- Test, Modify and Create Yourself (planetaryslalom)

35
Arrow

35.1 Game

Throw an arrow of unit mass as far as possible. Control initial speed and launching angle. Assume different friction forces depending on position and velocity.

35.2 Mathematics

Equations of motion:

$$dx = vdt, \quad dv = Fdt, \tag{35.1}$$

with $x(t) = (x_1(t), x_2(t)$ position of arrow at time t in a (x_1, x_2) coordinate system with x_1 horisontal and x_2 vertical axis. Assume the force $F = ((0, -G) + (f_1, f_2)$ with $G = 1$ gravitational constant and $f = (f_1, f_2)$ frictional force depending on the speed $|v(t)|$.

35.3 Analytical Mathematics

The solution of

$$\dot{x}_1 = 1, \quad \dot{x}_2 = v_2, \quad \dot{v}_2 = 2, \tag{35.2}$$

is given by

$$x_1(t) = t, \quad x_2(t) = t^2 \tag{35.3}$$

FIGURE 35.1. The Art of Bow and Arrow.

which represents a parabola.

$$x_2 = x_1^2 \tag{35.4}$$

35.4 Experiments

- Experimental setup
- Experiments

35.5 Demo + Lab

- Test, Modify and Create Yourself (arrow)

36
Achilles and the Tortoise

Achilles had overtaken the Tortoise, and had seated himself comfortably on its back. "So you've got to the end of our race-course?" said the Tortoise. "Even though it does consist of an infinite series of distances? I thought some wiseacre or other had proved that the thing couldn't be done?"

"It can be done," said Achilles. "It has been done! Solvitur ambulando. You see the distances were constantly diminishing; and so –"

"But if they had been constantly increasing?" the Tortoise interrupted "How then?"

"Then I shouldn't be here," Achilles modestly replied; "and you would have got several times round the world, by this time!"

"You flatter me – flatten, I mean" said the Tortoise; "for you are a heavy weight, and no mistake! Well now, would you like to hear of a race-course, that most people fancy they can get to the end of in two or three steps, while it really consists of an infinite number of distances, each one longer than the previous one?"

"Very much indeed!" said the Grecian warrior, as he drew from his helmet (few Grecian warriors possessed pockets in those days) an enormous note-book and a pencil. "Proceed! And speak slowly, please! Shorthand isn't invented yet!" (Lewis Carroll 1895)

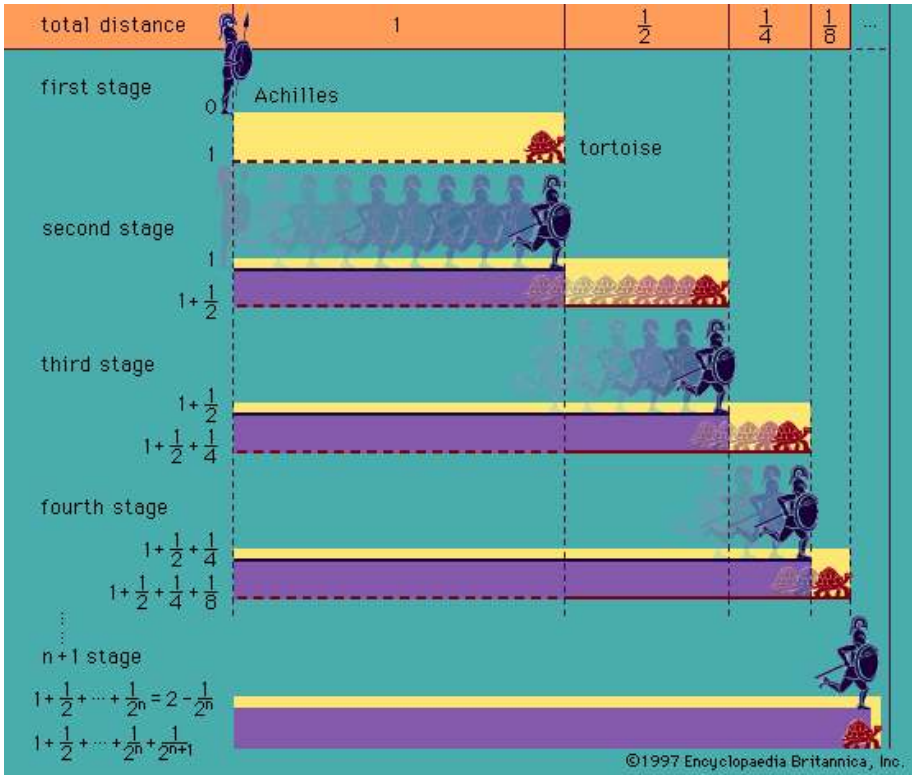


FIGURE 36.1. Achilles and the Tortoise.

36.1 Geometric Series

We have all heard the mathematical argument that the fast Achillles cannot overtake the slow Tortoise, or not even move a given distance of length 1 say, since first he has to move the half distance $\frac{1}{2}$ with $\frac{1}{2}$ still to go, and then half of $\frac{1}{2}$ that is $\frac{1}{4}$, and then $\frac{1}{8}$, and so on with always half of the last step still to go. The traveled distance is given by

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\frac{1}{32}... < 1 \tag{36.1}$$

It seems that Achilles would be unable to run a distance of 1. Can you resolve the paradox? Let's seek an answer in a game.

36.2 Experiments

- Experiment 1.

- Experiment 2.
- Zeno's Paradox.
- Resolution of Zeno's Paradox.

36.3 Game

Player 1 is Achilles and Player 2 the Tortoise and the game is a 100 meter dash. The Tortoise starts first and then Achilles. Will he overtake within 100 meters? What is the dependence on the respective velocities?

36.4 Mathematics: The Sum of a Geometric Series

Let $0 < a < 1$ and consider the sum

$$S(n) = 1 + a + a^2 + a^3 + ... + a^n \tag{36.2}$$

Multiplying by a we get

$$aS(n) = a + a^2 + a^3 + ... + a^{n+1} \tag{36.3}$$

and thus by subtraction term by term $S(n) - aS(n) = 1 - a^{n+1}$, which gives

$$1 + a + a^2 + a^3 + ... + a^n = \frac{1 - a^{n+1}}{1 - a}. \tag{36.4}$$

If we now let n increase without bound, so that n gets bigger than any given natural number, then a^{n+1} becomes smaller than any given positive number. We thus write with the dots indicating that n increases without bound:

$$1 + a + a^2 + a^3 + .. = \frac{1}{1 - a}, \tag{36.5}$$

which we write symbolically (with $a^0 = 1$)

$$\sum_{m=0}^{\infty} a^m = \frac{1}{1 - a}. \tag{36.6}$$

37
Nobel Peace Prize: Climate Sensitivity

During my service in the United States Congress, I took the initiative in creating the Internet. (Al Gore)

The day I made that statement, about the inventing the internet, I was tired because I'd been up all night inventing the Camcorder. (Al Gore)

Our world faces a true planetary emergency. I know the phrase sounds shrill, and I know it's a challenge to the moral imagination. (Al Gore)

37.1 Al Gore and Global Warming

Al Gore recieved the Nobel Peace Prize in 2007 together with Pachauri (about mathematics at time 15.25) chairman of the Intergovernmental Panel of Climate Change IPCC, for their alarm reports on Anthropgenic Global Warming (AGW) claiming a catastrophical climate sensitivity of up to 6 degrees Celsius global warming upon doubling of the CO2 concentration in the atmosphere.

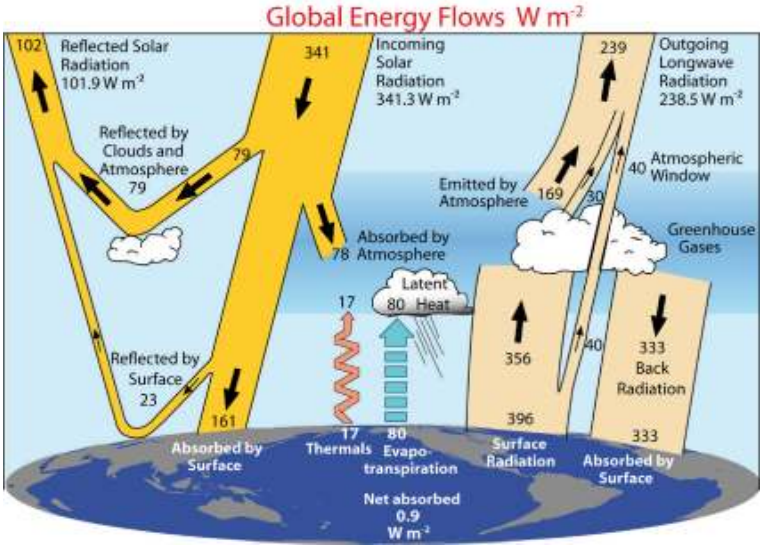


FIGURE 37.1. Common picture of "global radiative budget". Correct?

37.2 Mathematics

A (very) simple model for the thermodynamics of the Earth takes the form

$$\dot{T} = Q + aT - bT = Q - (b-a)T = Q - cT, \tag{37.1}$$

where T can be the sea surface temperature, Q is direct radiative forcing from the Sun, bT is outgoing radiation from the troposphere, and aT is feedback radiative forcing from water vapour, clouds et cet. The net model is thus

$$\dot{T} + cT = Q, \tag{37.2}$$

where c is a positive coefficient.

37.3 Climate Sensitivity

Climate sensitivity S is temperature change vs change of radiative forcing in stationary state with $\dot{T} = 0$, that is $S = \frac{1}{c}$. Can we determine the climate sensitivity S by observing the change of temperature for given perturbations of the radiative forcing, over the seasons of the year, or from day to night, and from such observations deduce the vale of c? For input, see Climate Sensitivity and Feedback 1-4 and Climate Sensitivity.

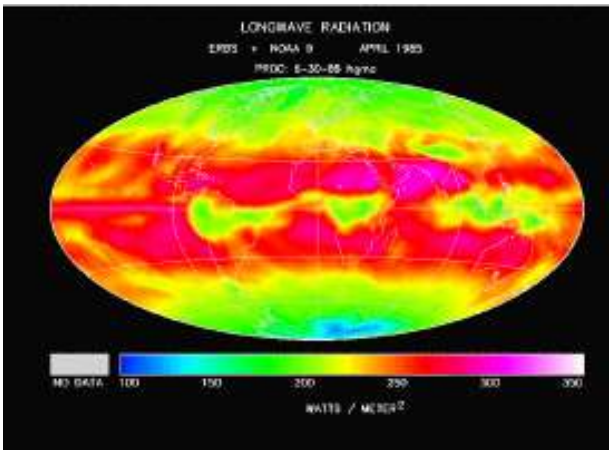


FIGURE 37.2. ERBE data.

We observe that if $Q = \sin(t)$, then the solution of (37.2) with $T(0) = -\frac{1}{1+c^2}$ is given by

$$T(t) = \frac{1}{1+c^2}(-cos(t) + c\sin(t)). \tag{37.3}$$

We observe that for c small, $T(t) \approx -\cos(t) = \sin(t - \frac{\pi}{2})$ equal to the forcing with a $\frac{\pi}{2}$ phase shift. For c large, $T \approx \frac{1}{c}\sin(t)$.

37.4 Game

Player 1 subjects the heat forcing $Q = \sin(t)$ to a secret perturbation and computes the correspond temperature variation $T(t)$ from (37.2) with some secret coefficient c. The objective of the Player 2 is to recover c from given variation of $T(t)$, e.g. by fitting of (37.3).

Player 1 and 2 switch roles, and the winner gets a Nobel Prize.

37.5 Data

Real data are given by Earth Radiation Budget Experiment (ERBE) and On the Determination of Climate Feedbacks from ERBE Data by Lindzen and Choi.

FIGURE 37.3. The winner gets a Nobel Prize.

37.6 Extended Model

The above basic model can be extended in many ways, for example by making a distinction between a mean temperature T_1 at the Equator and T_2 at the North Pole:

$$\dot{T}_1 = Q_1 - c_1T - c_3(T_1 - T_2), \quad \dot{T}_2 = Q_2 + c_3(T_1 - T_2) - c_2T_2 \qquad (37.4)$$

where Q_1 and Q_2 represent incoming radiative forcing, and c_1 and c_2 are coefficients of net outgoing radiation, at the Equator and North Pole, and c_3 is a heat exchange coefficient.

The objective of the game would be to identify the coefficients c_i from observations of T_1 and T_2 with Q_1 and Q_2 known seasonal forcing subject to unknown perturbation, with the exchange coeficient being of special interest as it determines the temperature difference $T_1 - T_2$ of crucial importance for glaciation.

37.7 Glaciation vs Eccentricity and Tilt of Earth's Orbit

The period of glaciations is roughly 100.000 years, which coincides with the period of varying eccentricity of the Earth's orbit around the Sun (Milankovitch cycles) giving a varying radiative forcing over the year. The eccentricity varies about 5 percent and higher eccentricity increases the seasonal contrast over the Northern and Southern Hemisphere depending also on the tilt which is about 23 degrees.

Figure 37.6 shows that we are at the end the peak of an interglacial period and are facing a rapdily approaching new ice age to last another 100.000 years...Real estate prices in Sweden are already turning down...only very long term investors have a chance to survive...

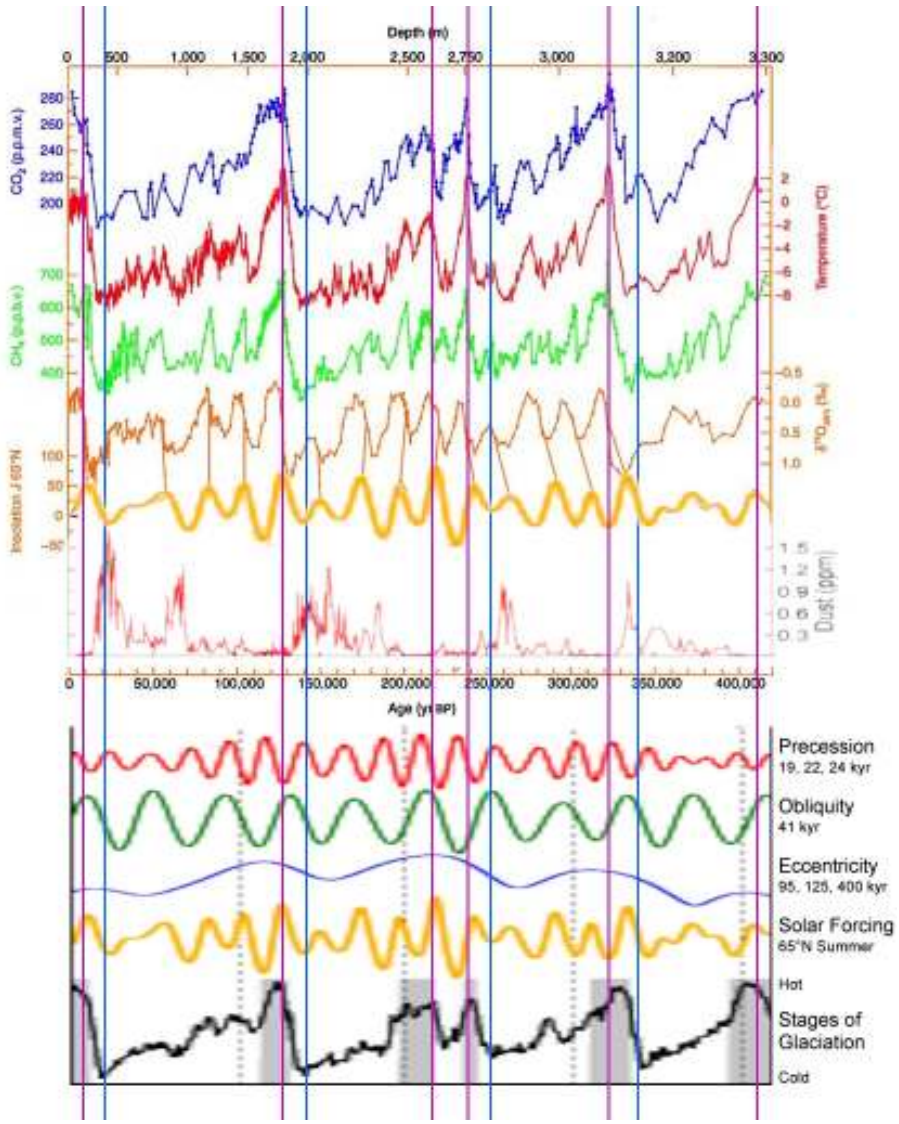


FIGURE 37.4. Climate data over 400.000 years. Notice in particular the connection between eccentricity and glaciation, and of course that the CO2 shadows temperature with a lag of 800 years (not the other way around).

It seems reasonal to expect that it is the seasonal contrasts that drive glaciation and not the global mean temperature which has changed just 2 degrees Celcius since Sweden was covered by ice 10.000 years ago. Climate change alarmists today claim that the small variation of global mean temperatures over periods of glaciation indicates a strong climate sensitivity to global mean temperatures. The political goal of a 2 degree limit is rooted in such an argument. But this may not be scienticfially correct, if in fact the global mean temperature is not strongly coupled to climate change. The above data indicates that we may be quickly approaching a new ice age, while the global temperature stays almost constant. The Winter 2010 is an example: Lots of ice an snow in the Northern Hemisphere, while the global temperature is about 0.6 degrees Celscius above 20th century mean value.

With larger eccentricity the NH Summer short and hot, while during the long NH Winter heat is flowing from the Equator, with the net effect that ice is not building up over Sweden...

37.8 Yearly Dynamics of Climate vs Glaciation

Consider the follwing simple model for coupled Equator-North Pole system:

$$\dot{T}_1 = Q_1 - e(T_1 - T_2) - r_1 T_1, \quad \dot{T}_2 = Q_1 + e(T_1 - T_2) - r_2 T_2, \qquad (37.5)$$

where T_1 and T_2 are temperatures, r_1 and r_2 outgoing radiation coefficients, and Q_1 and Q_2 are incoming insolations at the Equator and the North Pole, respectively, and e is a heat exchange coefficient.

Denote yearly average by a bar, and assume that Q_1 is year-periodic with $\bar{Q}_2 = 0$ and that $Q_1 > 0$ is constant. Taking the averages of the equations, we obtain assuming T_1 and T_2 to be year perodic so that the averages of $\dot{T}_1$ and $\dot{T}_2$ are zero:

$$\bar{Q}_1 = e(\bar{T}_1 - \bar{T}_2) + r_1 \bar{T}_1, \quad r_2 \bar{T}_2 = e(\bar{T}_1 - \bar{T}_2), \qquad (37.6)$$

from which follows assuming $\bar{Q}_1 = 1$:

$$\bar{T}_1 = \frac{e + r_2}{e(r_1 + r_2) + r_1 r_2}, \quad \bar{T}_2 = \frac{e}{e(r_1 + r_2) + r_1 r_2}. \qquad (37.7)$$

In the extreme case of $e = 0$, we get $\bar{T}_1 = \frac{1}{r_1}$ and $\bar{T}_2 = 0$, and in the other extreme case with e very large, we get $\bar{T}_1 = \bar{T}_2 = \frac{1}{r_1 + r_2}$, as expected.

The total poleward transferred heat/year equals

$$e(\bar{T}_1 - \bar{T}_2) = \frac{r_2}{e(r_1 + r_2) + r_1 r_2}$$

which does not depend on the yearly dynamics of the NP radiative forcing Q_2, only upon outgoing radiation and exchange.

Let us now make the heat exchange nonlinear, and then consider the extreme case with a short hot Summer so that Q_2 is large positive for a short time and slightly negative for a long time, with still $\bar{Q}_2 = 0$. This would correspond to an extreme eccentricity. The NP temperature T_2 then will have a somewhat delayed large peak for a short time with possibly $T_2 > T_1$. Suppose we eliminate the corresponding heat flow from the NP towards the Equator, by changing the linear term to $e(T_1 - T_2)^+$ with $v^+ = v$ is $v > 0$ and $v = 0$ else.

The average of the corresponding exchange term can be approximated by taking the average only over the long period when $T_1 > T_2$, which can be approximated by

$$e\bar{T}_1 - \gamma e\bar{T}_2 \tag{37.8}$$

with the coefficient $0 < \gamma < 1$ depending on the size of the oscillation of T_2 around its mean value. We obtain as above

$$\bar{T}_1 = \frac{\gamma e + r_2}{e(\gamma r_1 + r_2) + r_1 r_2}, \quad \bar{T}_2 = \frac{e}{e(\gamma r_1 + r_2) + r_1 r_2}. \tag{37.9}$$

We see as expected that $\bar{T}_2$ increases as γ decreases. The model thus illustrates that the year dynamics of the radiative NP heating can influence the poleward heat flow and thus increase the average NP temperature (and prevent glaciation).

38
Ping-Pong

Ping-pong is similar to Pong in 3d, with additional restriction of net, return velocity, return spin et cet.

38.1 Demo + Lab

- Test, Modify and Create Yourself 1d (pingpong1d)
- Test, Modify and Create Yourself 2d (pingpong2d)
- Test, Modify and Create Yourself 3d (pingpong3d)

FIGURE 38.1. The importance of focussing.

39
Particle-Spring Systems

I do not keep up with the details of particle physics. (Murray Gell-Mann)

It's indeed surprising that replacing the elementary particle with a string leads to such a big change in things. I'm tempted to say that it has to do with the fuzziness it introduces. (Edward Witten)

There are no Quantum Jumps, nor are there Particles! (H. D. Zeh)

39.1 Equations of Motion

We now generalize from pointlike hard particles to flexible systems consisting of hard particles connected by elastic springs, referred to as *particle-spring systems*. We start with a system consisting of two particles of mass M_1 and M_2, the positions of which we record by the coordinates $x^1(t)$ and $x^2(t)$. We connect the particles by an elastic spring with rest length L_{12} and spring constant E, which establishes a force between the particles, with the force acting on M_1 given by

$$F_{12} = E(r_{12} - L_{12})e_{12}, \tag{39.1}$$

where $r_{12} = |x^1 - x^2|$. This is an attractive force if $|r_{12}| < L_{12}$ and repulsive if $|r_{12}| > L_{12}$, and let

$$e_{12} = \frac{x^2 - x^1}{r_{12}} \tag{39.2}$$

be the vector of unit length pointing from x^1 to x^2. Of course (why?) the force F_{21} acting on particle M_2 is the reverse of F_{12} so that $F_{21} = -F_{12}$ (Newton's 3rd Law).

We say that this is a *linear spring* since the spring force is directly proportional to the elongation $|r_{12}| - L_{12}$ from the rest length.

The equations of motion are

$$\begin{aligned} \dot{x}^1(t) &= v^1(t), \quad \dot{x}^2(t) = v^2(t), \\ \dot{v}^1(t) &= \frac{F_{12}}{M_1}, \quad \dot{v}^2(t) = \frac{F_{21}}{M_2}, \end{aligned} \tag{39.3}$$

or in incremental form using Smart-Euler:

$$\begin{aligned} v^{1,n+1} &= v^{1,n} + \frac{F_{12}^n}{M_1}dt, \quad v^{2,n+1} = v^{2,n} + \frac{F_{21}^n}{M_2}, \\ x^{1,n+1} &= x^{1,n} + v^{1,n+1}dt, \quad x^{2,n+1} = x^{2,n} + v^{1,n+1}dt, \end{aligned} \tag{39.4}$$

where $x^{i,n} = x^i(ndt)$, $v^{i,n} = v^i(ndt)$ for $i = 1, 2$, and $F_{12}^n = E(r_{12}^n - L_{12})e_{12}^n$ with $e_{12}^n = \frac{x^{2,n} - x^{1,n}}{r_{12}^n}$ and $r_{12}^n = |x^{2,n} - x^{1,n}|$.

39.2 Experiments

- Flying Circus Cow
- Inside Flying Circus Cow
- Particle-spring elastic system

39.3 Generalization

We can directly generalize to any number of particles connected by any set of linear springs, including crossing springs. We can generlize to non-linear springs with a non-linear relation between spring force and spring elongation. See e.g. N-Body Systems.

39.4 Demo + Lab

- Test, Modify and Create Yourself (particlespring)

40
Elastic Pong 1d

40.1 Demo + Lab

- Test, Modify and Create Yourself 1d (elasticpong1d)

40.2 Computational vs Analytical Elastic Collision

There are two possibilties of describing elastic collision, between two bodies coming into contact or one body colliding with a solid wall:

1. Formulate a analytical law of collision, like reflection.
2. Insert an elastic spring with short range of action and resolve the collision process by time-stepping.

Option 2. can be used in general and replaces possibly tricky analytical mathematics required for 1. by simple computation; imagine e.g. a multiple-body collision.

41
Elastic Pong 2d and 3d

Use short-range elastic springs to model elastic collisions and reflections at boundaries.

41.1 Demo + Lab

- Test, Modify and Create Yourself 2d (elasticpong2d)
- Test, Modify and Create Yourself 3d (elasticpong3d)

42
Elastic Ping-Pong

42.1 Demo + Lab

- Test, Modify and Create Yourself (elasticpingpong)

43
Billiards

43.1 Analytical Mathematics

One can find the analytical laws of elastic collision and reflection by using that momentum and kinetic energy is conserved and that the direction after impact of the ball at rest before impact, is given by the normal to the plane of contact.

43.2 Computational Mathematics

In computation short-range elastic springs can be used instead of analytical collision/reflection laws. Simpler and more general. Contact with friction can easily be added.

43.3 Spinning Cue-Balls

Spinning the cue-ball (which you do with the cue) is useful to position the cue ball appropriately after collision. Spin can be modeled by a suitable additional force on the cue ball. The impact between the cue and the ball must come along with a friction force, unless the direction of the cue points at the center of the ball, in order for the ball to move in the direction of the cue. It is this friction force which causes the spin. Friction can be modeled

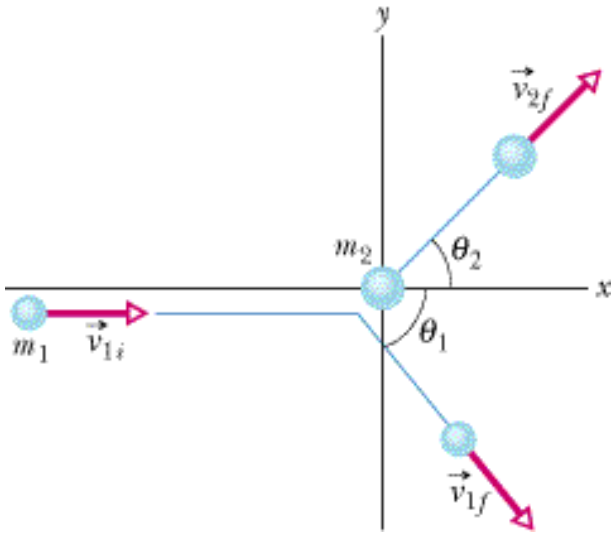


FIGURE 43.1. Billiard angles.

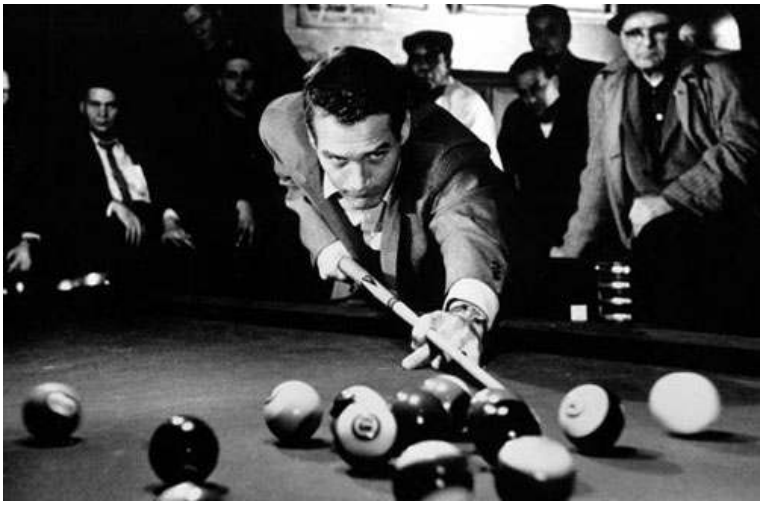


FIGURE 43.2. Paul Newman in The Hustler

by adding a force proportional to the spring force in the plane tangent to the point of contact.

43.4 Watch

- Colliding Elastic Spheres
- Billiards Simulator

43.5 Game

Construct a billiards simulator.

44
Curling

Curling describes that the stone can be given a rotation at start resulting (how?) in a sideway force making the trajectory (path) of the stone slightly curved, more so at the end of the trajectory when the speed is small.

44.1 Watch

- Curling Rocks
- Curling Basics

44.2 Game

Construct a curling simulator and game.

FIGURE 44.1. Olympic Champion Anette Norberg.

45
Elastic String

We consider a string occupying the interval $[0,1]$ consisting of equal particles initially at the points $x^i = ih$, $i = 0, 1, ..., J+1$, where h is the distance between two consecutive particles. Assume the particles are connected by linear springs with spring constant $\frac{E}{h}$ and zero rest length. Notice that the spring constant is normalized with the length h of each spring. Assume the particles at the end points x^0 and x^{J+1} are held fixed. Let $u^i(t)$ be the *displacement* of particle i from its initial position at ih. In particular, we have $u^0(t) = 0$, $u^{J+1}(t) = 0$.

The spring force between particle i and $i+1$ acting on particle i is given by

$$F_{i,i+1} = \frac{E}{h}(u^{i+1} - u^i) \tag{45.1}$$

and the spring force between $i-1$ and i acting on particle i, is given by

$$F_{i,i-1} = -F_{i-1,i} = -\frac{E}{h}(u^i - u^{i-1}). \tag{45.2}$$

The net force acting on particle i thus is given by

$$F_i = F_{i,i+1} + F_{i,i-1} = \frac{E}{h}(u^{i+1} - 2u^i + u^{i-1}). \tag{45.3}$$

The equation of motion for ball i, assuming that the mass of each particle is Mh (total mass M), is given by

$$Mh\dot{u}^i(t) = v^i(t), \quad \dot{v}^i(t) = \frac{E}{h}(u^{i+1} - 2u^i + u^{i-1}) \tag{45.4}$$

that is, assuming $\frac{E}{M} = 1$ for simplicity,

$$\dot{u}^i(t) = v^i(t), \quad \dot{v}^i(t) = \frac{u^{i+1} - 2u^i + u^{i-1}}{h^2}, \quad i = 1, ..., J. \tag{45.5}$$

Using the notation $\ddot{u} = \frac{d}{dt}\dot{u} = \dot{v}$, we can write the equations of motion

$$\ddot{u}^i(t) = \frac{u^{i+1} - 2u^i + u^{i-1}}{h^2} \quad i = 1, ..., J. \tag{45.6}$$

45.1 Demo + Lab

- Test, Modify and Create Yourself (elasticstring)

45.2 Space Derivative

The above derivation of the equations of motion for an elastic string of particles leads to the definition of space derivative of the displacement u as change of displacement per unit space step as $\frac{du}{dx}$ where $dx = h$ and $du^i = u^{i+1} - u^i$. We then have with $E = 1$

$$F_{i,i+1} = \frac{du^i}{dx} = \frac{u^{i+1} - u^i}{h}. \tag{45.7}$$

This leads to define the derivative $u'(x)$ of a function $u(x)$ as

$$u'(x) = \frac{du}{dx} = \frac{u(x+h) - u(x)}{h} \tag{45.8}$$

for small h.

We then define $u''(x)$ as the derivative of $u'(x)$, that is

$$u''(x) = \frac{u'(x+h) - u'(x)}{h} = \frac{\frac{u(x+h)-u(x)}{h} - \frac{u(x)-u(x-h)}{h}}{h} \tag{45.9}$$

that is

$$u''(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} \tag{45.10}$$

for small h.

We also write

$$u'' = \frac{d}{dx}\frac{du}{dx} = \frac{d^2u}{dx^2}. \tag{45.11}$$

Letting the number of particles increase and their mutual distance h decrease, we are led to the following equations for an *elastic string*: Find

the displacement $u(x,t)$ defined for $x \in [0,1]$ and $t \geq 0$, such that:

$$\begin{aligned} \ddot{u}(x,t) &= u''(x,t) \quad \text{for } x \in (0,1), t > 0, \\ u(0,t) &= u(1,t) = 0 \quad \text{for } t > 0, \\ u(x,0) &= u^0(x), \quad \dot{u}(x,0) = \dot{u}^0(x) \quad \text{for } x \in (0,1), \end{aligned} \tag{45.12}$$

where $u^0(x)$ and $\dot{u}^0(x)$ are given functions.

Note that here we consider small displacements $u(x,t)$ *along the string* assuming the string is kept straight. Below we will consider *transversal* displacements perpendicular to the string.

The equation

$$\ddot{u} = u'' \tag{45.13}$$

is called the *1d wave equation*. This equation also describes the propagation of sound waves in a straight tube with the air molecules oscillating back and forth as if connected by linear springs.

We can write the wave equation as a first order system as follows

$$\begin{aligned} \dot{u} &= v \\ \dot{v} &= F' \\ F &= u' \end{aligned} \tag{45.14}$$

where u is displacement, v velocity and F spring tension and F' spring force acting on particles.

45.3 To Think About

- String theory

46
Visco-Elastic String

An elastic string with a viscous (damping) force proportional to the velocity v changes Newtons 2nd Law to

$$\dot{v} + \nu v = F \tag{46.1}$$

where ν is a non-negative viscosity. The wave equation for an elastic string with viscous damping thus takes the following form with $\nu = 1$:

$$\dot{u} = v, \quad \dot{v} + v = u'' \tag{46.2}$$

which is solved by time stepping and discretization of u'' as above.

46.1 Demo + Lab

- Test, Modify and Create Yourself (viscoelasticstring)

47
Elastic Net

Direct generalization to in-plane displacement of an elastic net occupying $[0,1] \times [0,1]$ consisting of particles at the points $x^{i,j} = (ih, jh)$, $i, j = 0, 1, ..., M$, with in-plane displacements $u^{i,j} = u(ih, jh)$, satisfying

$$\begin{aligned} \dot{u}^{i,j}(t) &= v^{i,j}(t), \\ \dot{h}^2 v^{i,j}(t) &= \frac{E}{h} h(u^{i+1,j} - 2u^{i,j} + u^{i-1,j}) + E(u^{i,j+1} - 2u^{i,j} + u^{i,j-1}) \\ &= E(u^{i+1,j} + u^{i-1,j} + u^{i,j+1} + u^{i,j-1} - 4u^{i,j}), \end{aligned} \tag{47.1}$$

with mass of each particle $= h^2$ and the spring constant $\frac{E}{h}$ multiplied by h to get correct scaling.

With a vanishingly small mesh size h these equations take the following form if $E = \frac{1}{h^2}$:

$$\ddot{u} = \Delta u, \tag{47.2}$$

with

$$\Delta u = \frac{\partial^2 u}{\partial dx_1^2} + \frac{\partial^2 u}{\partial dx_2^2}. \tag{47.3}$$

We will see below that u can also represent the transversal displacement of an elastic membrane.

With damping we get a model of the form

$$\ddot{u} + \dot{u} = \Delta u. \tag{47.4}$$

47.1 Demo + Lab

- Test, Modify and Create Yourself (elasticnet)

48
Elastic Body

Generalizing to 3d we are led to the following wave equation for the dynamic vibration of an elastic body occupying the volume Ω and being fixed at the boundary Γ (compare an elastic cube without fixation of the boundary):

$$\begin{aligned} \ddot{u}(x,t) &= \Delta u(x,t) \quad \text{for } x \in \Omega, t > 0, \\ u(x,t) &= 0 \quad \text{for } x \in \Gamma, t > 0, \end{aligned} \tag{48.1}$$

with $u(x,0)$ and $\dot{u}(x,0)$ given initial values in Q, and

$$\Delta u = \frac{d^2u}{dx_1^2} + \frac{d^2u}{dx_2^2} + \frac{d^2u}{dx_3^2}. \tag{48.2}$$

48.1 Watch

- Another elastic cube.

Viscosity is introduced by adding a term $\nu\dot{u}$ as above.

48.2 Demo + Lab

- Test, Modify and Create Yourself (elasticbody)

49
Elast String: Tranversal Motion

The particle-spring string model (45.6)

$$\ddot{u}^i(t) = \frac{u^{i+1} - 2u^i + u^{i-1}}{h^2} \quad i = 1, ..., M. \tag{49.1}$$

can be given a different interpretation with u^i being the vertical (transversal) displacement of particle i of a horisontal string of beads, instead of the horisontal displacement as above. To arrive at this interpretation, we assume that the tension in the string is uniform so that all springs have the same spring force of unit strength. The inclination of the spring from particle $i-1$ to particle i is given by $\frac{u^i - u^{i-1}}{h}$ and that from particle i to particle $i+1$ equals $\frac{u^{i+1} - u^i}{h}$. The vertical component of the spring force on particle i is thus given by

$$\frac{u^{i+1} - 2u^i + u^{i-1}}{h} \tag{49.2}$$

which gives (49.1) from Newton's 2nd Law with accelleration force $h\ddot{u}^i$ attributing the mass h to each particle.

We conclude that the 1d wave equation (45.12) also models transversal vibration of an elastic string.

49.1 Game

Player 1 controls the left endpoint of the particle-spring string, and Player 2 the right. The objective is to kick a wabe back and forth by suitable vertical motion of the endpoints.

49.2 Realization

50
Music = Vibrating Elastic Strings

Music is the pleasure the human mind experiences from counting without being aware that it is counting. (Leibniz)

It is this way that in mathematics speculative theorems and practical canons are reduced by analysis to definitions, axioms and postulates. (Leibniz)

There is geometry in the humming of the strings, there is music in the spacing of the spheres. (Pythagoras)

50.1 Harmonics of a Vibrating Strings

Show that functions $u_n(x,t)$ of the form

$$u_n(x,t) = (a_n \cos(\pi nt) + b_n \sin(\pi nt)) \sin(\pi nx), \quad n = 1, 2, 3, ..., \quad (50.1)$$

where a_n and b_n are real constants, solve the wave equation (45.12) for a tranversally vibrating elastic string. The functions $\sin(\pi nx)$ are called the *harmonics* of the string and correspond to different musical tones with different frequencies n, with the *base tone* given by $n = 1$ and the *overtones* by $n > 1$.

If the string is plucked, initiated with $u^0 \neq 0$ and/or $\dot{u}^0 \neq 0$, it will generate a tone which is a sum

$$u(x,t) = \sum_{n=1}^{N} u_n(x,t), \quad (50.2)$$

of the harmonics $\sin(nx)$ with amplitudes $(a_n \cos(\pi nt) + b_n \sin(\pi nt))$ varying with time. If you have several strings, of different lengths/tension as on a guitar, then you can generate *chords* consisting of two or more tones, including overtones.

50.2 The Pythagorean Scale

The Pythagorean scale or tuning is generated from the harmonics based on the ratio $3:2$ of a fifth according the the circle of fifths as follows:

- base note: C: 1 (normalized frequency)
- octave: high C: 2
- fifth: G: $\frac{3}{2}$
- second: D: $\frac{9}{8}$
- sixth: A: $\frac{27}{16}$
- third: E: $\frac{81}{64}$
- seventh: H: $\frac{243}{128}$
- fourth: F: $\frac{4}{3}$,

with the last fifth, which would come out as $F\#$ or F-sharp of frequency $\frac{729}{512} = 1.423828125$, is replaced by F with the simplest possible ratio $\frac{4}{3} = 1.3333....$

50.3 The Equally-Tempered Scale

In the *equally-tempered* scale with an octave of 12 half-notes, the relative frequency between successive half-notes is equal to $x = 2^{\frac{1}{12}}$ as the positive solution of the equation $x^{12} = 2$.

Compare the Pythgorean scale to the equally-tempered major scale consisting of the following sequence of intervals: whole-whole-half-whole-whole-whole-half, with a whole interval equal to two half-notes.

Compare:

- Pythagoras and Music 1
- Pythagoras and Music 2
- Pythagorean scale.
- Non-Pythagorean scales.

FIGURE 50.1. Pythagoras discovering the mathematics of music, and the Universe

50.4 Musical Game

Arrange a Song Contest based on e.g the Pythagorean scale.

50.5 Demo + Lab

- Test, Modify and Create Yourself (transversalelasticstring)

51
Elastic Membrane

The bead-spring string model with transversal partcile motion generalizes to a vibrating membrane, described by

$$\begin{aligned} \ddot{u} - \Delta u &= f \quad \text{for } x \in \Omega,\ t > 0, \\ u(x,t) &= 0 \quad \text{for } x \in \Gamma,\ t > 0, \\ u(x,0) = u^0(x), \dot{u}(x,0) &= \dot{u}^0, \quad \text{for } x \in \Omega, \end{aligned} \tag{51.1}$$

where $u(x,t)$ is the vertical displacement at position x at time t of a horisontal menbrane covering the domain Ω with fixed zero displacement at its boundary Γ, acted upon by the vertical force dsitribution $f(x,t)$.

51.1 Demo + Lab

- Test, Modify and Create Yourself (transversalelasticmembrane)

51.2 Game

Players can hit the membrane by supplying the force $f(x,t)$ and/or changing the initial conditions.

51.3 Realization

52
Bungy Jump

Simulate a bungy jump:

52.1 Demo + Lab

- Test, Modify and Create Yourself (bungyjump)

FIGURE 52.1. Free-falling body attached to elastic chord.

53
Spin-Ping-Pong

A spinning ball flying in the air will follow a curved path (without gravitation) because the spin causes non-symmetric separation with generates a force perpendicular to the axis of spin and the direction of flight. Add this effect to your Pong games.

53.1 Perspective

- Why a Topspin Tennis Ball Curves Down.

53.2 Demo + Lab

- Test, Modify and Create Yourself (spinpingpong)

54
Elastic Spin-Ping-Pong

Extend to elastic spinning ball.

54.1 Demo + Lab

- Test, Modify and Create Yourself (elasticspinpingpong)

55
Golf

Construct your own game using experience from spin-ping-pong. Watch

- 3d Minigolf Challenge (Iphone)
- Classic game
- Game with real clubs
- More realistic game

FIGURE 55.1. Tiger Woods swing.

56
Tennis

Construct your own game using experience from spin-ping-pong. Watch

- Top Spin 3 Borg-Becker
- Haptic interaction

FIGURE 56.1. Björn Borg top spin.

57
Squash

Construct your own game using experience from spin-ping-pong. Watch

- Rally 2009
- Champions 2009
- Computer game.

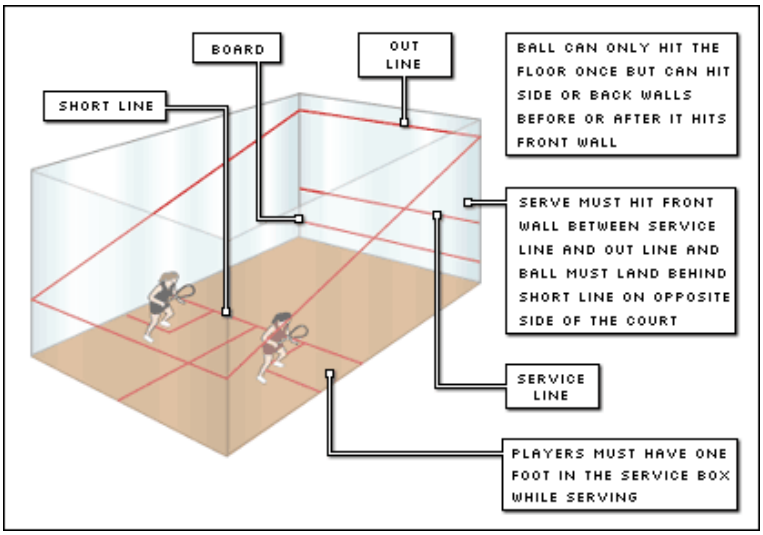


FIGURE 57.1. Squash court

58
Badminton

Construct your own game using experience from viscous-ping-pong. Watch

- Computer game
- Long rally
- How to play.

FIGURE 58.1. Badminton court.

59
Electrostatic Barrier

Build a barrier by suitably lining up electrical charges with the objective of preventing an equally charged approaching particle to get through. Use Coulomb's Law for the respulsive force between electrical charges of equal sign, which analogous to Newton's law of gravitation with a change of sign (repulsion instead of attraction).

59.1 Game

Design a game including electrostatic walls and moving charged particles.

Part IV

Leibniz' World of Calculus

FIGURE 59.1. Gottfried Wilhelm Leibniz: *There are also two kinds of truths: truth of reasoning and truths of fact. Truths of reasoning are necessary and their opposite is impossible; those of fact are contingent and their opposite is possible... When a truth is necessary, the reason for it can be found by analysis, that is, by resolving it into simpler ideas and truths until the primary ones are reached.*
.

60
Differential Equations of Motion

> 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)
>
> Making the simple complicated is commonplace; making the complicated simple, awesomely simple, that's creativity. (Charles Mingus)

60.1 Initial Value Problem IVP

We recall that velocity v is defined as the change dx of position x per unit time step dt:

$$v = \frac{dx}{dt}\,. \tag{60.1}$$

We also refer to $\frac{dx}{dt}$ as the *derivative of* x with respect to t. Similarly, we have

$$\frac{dv}{dt} = a\,, \tag{60.2}$$

where a is the accelleration. In other words:

- Derivative of position with respect to time = velocity,
- Derivative of velocity with respect to time = accelleration.

We will denote derivative with respect to time, alternatively with a dot:

$$\dot{x} = \frac{dx}{dt}, \quad \dot{v} = \frac{dv}{dt}, \tag{60.3}$$

where we in the spirit of Leibniz assume the time step to be vanishingly small. Newton used $\dot{x}$ and Leibniz used $\frac{dx}{dt}$ to denote the derivative of x with respect to time t. Newton's age is sometimes referred to as the *dot-age* (while our age may be referred to as the *dot-com-age*, right?).

The process of computing the derivative $\dot{x}(t)$ of a function $x(t)$ is called *differentiation*, or more precisely, *differentiation with respect to* t.

Using Newton's notation we can thus write the equations of motion, assuming $M = 1$ for simplicity, as *differential equations*:

$$\dot{x} = v, \quad \dot{v} = F, \tag{60.4}$$

which is to be interpreted as

$$\begin{aligned} &\dot{x}(t) = v(t), \quad \dot{v}(t) = F(t) \quad \text{for } t > 0, \\ &x(0) = x^0, \quad v(0) = v^0, \end{aligned} \tag{60.5}$$

where $x(t)$, $v(t)$ and $F(t)$ are viewed as functions of t, and x^0 and v^0 are given *initial values* of position and velocity at an *initial time* $t = 0$.

We can generalize to $F = F(x, v, t)$ depending also on x and v, in which case the equations of motion read

$$\begin{aligned} &\dot{x}(t) = v(t), \quad \dot{v}(t) = F(x(t), v(t), t) \quad \text{for } t > 0, \\ &x(0) = x^0, \quad v(0) = v^0. \end{aligned} \tag{60.6}$$

or with $\ddot{x} = \dot{v}$

$$\begin{aligned} &\ddot{x}(t) = F(x(t), \dot{x}(t), t) \quad \text{for } t > 0, \\ &x(0) = x^0, \quad \dot{x}(0) = v^0. \end{aligned} \tag{60.7}$$

We refer to (60.5), (60.6) and (60.7) as *Initial Value Problems* or *IVPs*.

60.2 Measures of Change: Continuity, Derivative

Calculus is the mathematics of change with the *derivative* being a measure of change, and thus can be viewed as the mathematics of IVPs.

The time derivate $\dot{x}(t)$ of position $x(t)$ as function of time t, measures the change of position per unit time step. A function $x(t)$ with derivative $\dot{x}(t)$ is said to be *differentiable*.

Another basic concept of Calculus related to change is *continuity*, which is a form of poor cousin of derivative, also measuring change but in a less precise way.

You will below meet the precise definitions of *derivative* and *continuity*, as more or less precise *measures of change*. We here prepare these basic definitions with a short introductory discussion.

60.3 A Basic Example

If the velocity $\dot{x}(t)$ of position $x(t)$ is constant $\dot{x}(t) = v$ with v a constant velocity, then the position $x(t)$ changes linearly with time: $x(t) = x^0 + vt$ for $t > 0$ with x^0 the position for $t = 0$. Thus $x(t) = x^0 + vt$ is a linear function of t, since it has the form $c_0 + c_1 t$ with c_0 and c_1 constants.

If $v(t)$ is not constant, then $x(t)$ will not be linear in t, but if $v(t)$ is almost constant locally, for small changes of t, then $x(t)$ will be almost linear locally. We here meet both the concept of *continuity* and the concept of *differentiability*:

- A function $v(t)$ is continuous if $v(t)$ is locally close to a constant.
- A function $x(t)$ is differentiable (with derivative $\dot{x}(t)$) if $x(t)$ is locally close to a linear function in t.

We shall below meet the concept of continuity as *Lipschitz continuity* including a quantitative measure of the local deviation from a constant.

It is natural to generalize, an essential aspect of mathematics, to:

- A function $x(t)$ is *two times differentiable* if $x(t)$ is locally close to a quadratic function in t up to a third order term.

This connects to a particle with position $x(t)$ subject to constant acceleration a, in which case the velocity $v(t) = \dot{x}(t) = at + v^0$ and the position

$$x(t) = \frac{a}{2}t^2 + v^0 t + x^0 \tag{60.8}$$

is exactly equal to a quadratic function in t. Differentiating $\dot{x} = at$ with respect to t, we find that the *second derivative* $\ddot{x}(t) = \frac{d}{dt}\dot{x}(t) = a$, and thus (60.8) can be written

$$x(t) = x(0) + \dot{x}(0)t + \frac{\ddot{x}(0)}{2}t^2 \tag{60.9}$$

while for a general twice differentiable function $x(t)$

$$x(t) \approx x(0) + \dot{x}(0)t + \frac{\ddot{x}(0)}{2}t^2 \quad \text{for } |t| \text{ small}, \tag{60.10}$$

up to a term of order $|t|^3$ (allowing t to also be negative).

We shall below recover this expression as an example of *Taylor's formula* expressing a general function locally as a polynomial with coefficients given by the values of the function and its derivatives at a specific point.

60.4 Perspectives

- Return of Descartes

- Soul as Simulation of Body
- Zeno's Paradox of Particle Motion
- Slinky as Resolution of Zeno's Paradox

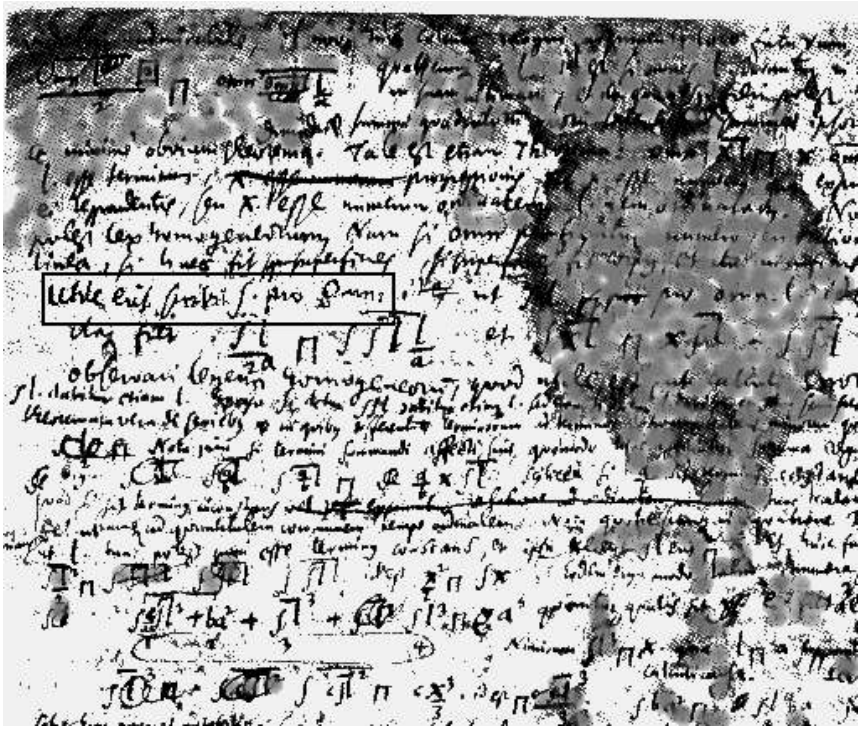


FIGURE 60.1. Leibniz manuscript from October 29, 1675, introducing the integral sign $\int$ (in the box): *It is useful to denote summation (omnia) by* $\int$. Yes, it has shown to be very useful.

60.5 To Think About

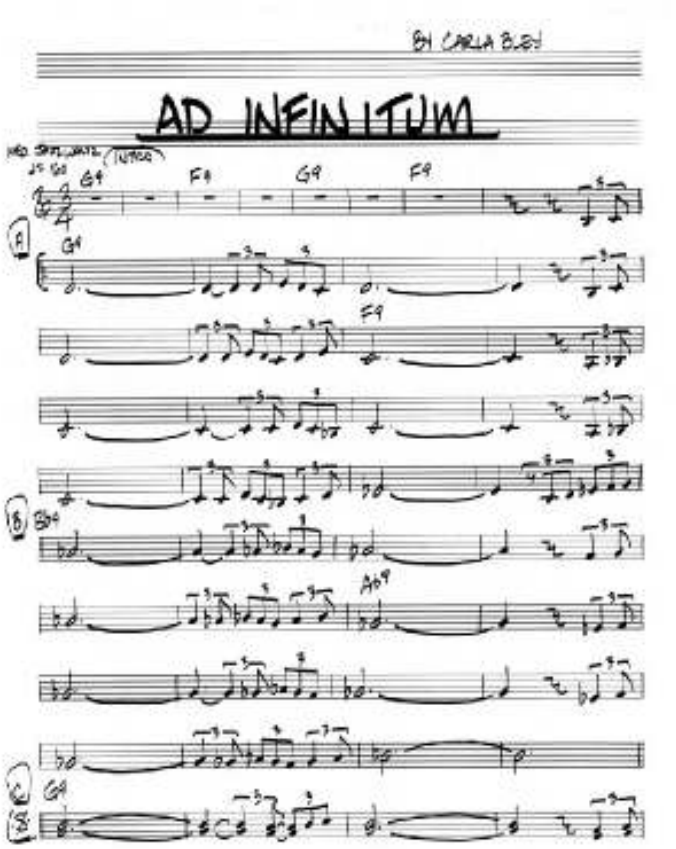
- What did Galileo say about motion?
- How is the derivative of a function defined? Computed?
- How does a speedometer on a bike work?
- How does a distance meter on a bike work?

60.6 Watch

- Newton, Leibniz or Gore?, Leibniz Calculus rap
- Importance of Calculus, Leibniz Monadology
- Quinton about Leibniz
- Feynman on the Relation of Mathematics and Physics

> Seeing then that truth consisteth in the right ordering of names in our affirmations, a man that seeketh precise truth had need to remember what every name he uses stands for, and to place it accordingly; or else he will find himself entangled in words, as a bird in lime twigs; the more he struggles, the more belimed. And therefore in geometry (which is the only science that it hath pleased God hitherto to bestow on mankind), men begin at settling the significations of their words; which settling of significations, they call definitions, and place them in the beginning of their reckoning. (Leviathan, Thomas Hobbes)

> When man reasoneth, he does nothing else but conceive a sum total, from addition of parcels; or conceive a remainder, from subtraction of one sum from another: which, if it be done by words, is conceiving of the consequence of the names of all the parts, to the name of the whole; or from the names of the whole and one part, to the name of the other part. And though in some things, as in numbers, besides adding and subtracting, men name other operations, as multiplying and dividing; yet they are the same: for multiplication is but adding together of things equal; and division, but subtracting of one thing, as often as we can. These operations are not incident to numbers only, but to all manner of things that can be added together, and taken one out of another. For as arithmeticians teach to add and subtract in numbers, so the geometricians teach the same in lines, figures (solid and superficial), angles, proportions, times, degrees of swiftness, force, power, and the like; the logicians teach the same in consequences of words, adding together two names to make an affirmation, and two affirmations to make a syllogism, and many syllogisms to make a demonstration; and from the sum, or conclusion of a syllogism, they subtract one proposition to find the other. (Leviathan, Thomas Hobbes)


BY CARLA BLEY
AD INFINITUM
G9
F9
G9
F9
A
G9
F9
B
B♭9
A♭9
C
G9

61
Functions $f : \mathbb{Q}^m \to \mathbb{Q}^n$

> All things are subject to interpretation whichever interpretation prevails at a given time is a function of power and not truth. (Friedrich Nietzsche)
>
> The function of muscle is to pull and not to push, except in the case of the genitals and the tongue. (Leonardo da Vinci)
>
> The supreme function of reason is to show man that some things are beyond reason. (Blaise Pascal)
>
> The function of education is to teach one to think intensively and to think critically... Intelligence plus character - that is the goal of true education. (Martin Luther King, Jr.)

We denote by Q the set of rational numbers, that is, the numbers with finite or periodic decimal expansion.

We write $f : D(f) \to \mathbb{Q}$ if for each given value of x in some set, called the *domain* $D(f)$ of f, a rational number $f(x)$ is assigned, that is, for each $x \in D(f)$ the function value $f(x) \in \mathbb{Q}$ is assigned.

The set of values $f(x)$ for $x \in D(f)$ forms the *range* $R(f)$ of f. We can thus write $f : D(f) \to R(f)$ stating that for each $x \in D(f)$ there is assigned a value $f(x) \in R(f)$, and for each value $y \in R(f)$ there is at least one value $x \in D(f)$ such that $y = f(x)$.

In particular, writing $f : \mathbb{Q} \to \mathbb{Q}$, means that for each $x \in \mathbb{Q}$, a function value $f(x) \in \mathbb{Q}$ is assigned.

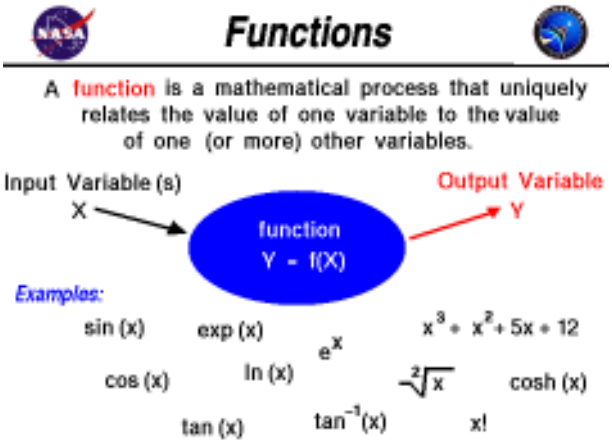


FIGURE 61.1. NASA definition of function $f(x)$

Writing $f : \mathbb{Q}^m \to \mathbb{Q}^n$ means that for each m-vector $x \in \mathbb{Q}^m$, an n-vector $f(x)$ with rational coefficients is assigned.

We often write e.g. $f : \mathbb{Q}^m \to \mathbb{Q}^n$ without explicitly specifying the domain $D(f) \subset \mathbb{Q}^m$, or the range $R(f) \subset \mathbb{Q}^n$.

It is common to denote by $\mathbb{R}$ the set of all rational numbers together with the numbers with an infinite non-periodic decimal expansion, referred to as the set of *real numbers*.

We shall see a function $f : \mathbb{Q}^m \to \mathbb{Q}^n$ can be extended to a function $f : \mathbb{R}^m \to \mathbb{R}^n$, if the function $f(x)$ is continuous in a way to be specified precisely below.

We denote by $\mathbb{R}_+$ the set of positive real numbers.

61.1 Read More

- What is a Function?

61.2 To Think About

- How did the notion of *function* develop?
- Who introduced the concept and terminology?

62
$x(t) = \int_0^t v(s)ds$ solves $\dot{x}(t) = v(t)$

> Without mathematics we cannot penetrate deeply into philosophy. Without philosophy we cannot penetrate deeply into mathematics. Without both we cannot penetrate deeply into anything. (Leibniz)

62.1 The Most Basic IVP

The solution of the IVP of finding $x : [0, T] \to \mathbb{R}$ such that

$$\dot{x}(t) = v(t) \quad \text{for } 0 < t \leq T, \quad x(0) = 0, \tag{62.1}$$

where $v : [0, T] \to \mathbb{R}$ is a given function and $[0, T]$ a given time-interval, is denoted by

$$x(t) = \int_0^t v(s)\, ds, \quad t \in [0, T], \tag{62.2}$$

and is referred to as the *integral* or *primitive function* of $v(t)$. So far the integral $\int_0^t v(s)\, ds$ is just a sign or name of the solution $x(t)$, and it remains to give it a concrete meaning. We shall see that the S-like integral sign $\int$ can be viewed as indicating a certain form of Summation, which we shall make precise. The integral sign $\int$ was the strike of genius of Leibniz, long before logotypes became the carriers of the inner meaning of companies and organizations.

The Forward Euler method for the IVP (62.1), is given by

$$x((n+1)dt) = x(ndt)+v(ndt)dt \quad \text{for } n = 0, 1, 2, ..., N, \quad \text{with } (N+1)dt = T, \tag{62.3}$$

or equally well

$$x((n+1)ds) = x(nds)+v(nds)ds \quad \text{for } n = 0, 1, 2, ..., N, \quad \text{with } (N+1)ds = T, \tag{62.4}$$

with $dt = ds$ the time step. If we replace $x(nds)$ by $x((n-1)ds) + v((n-1)ds)ds$, and so on, we see that $x((n+1)ds)$ can be expressed as as a sum

$$x((n+1)ds) = \sum_{m=0}^{n} v(mds)ds = v(0)ds+v(ds)ds+v(2ds)ds+...+v(nds)ds. \tag{62.5}$$

We are thus led to view

$$\int_0^t v(s)ds \quad \text{and} \quad \sum_{m=0}^{n} v(mds)\, ds, \tag{62.6}$$

to be similar, which we shall make precise below. We refer to the sum representation of the integral as a *Riemann sum*. We sum up so far:

Observation 1: The integral $x(t) = \int_0^t v(s)ds$ satisfies by definition

$$\dot{x}(t) = \frac{d}{dt}\int_0^t v(s)ds = v(t) \quad \text{for } 0 < t \leq T. \tag{62.7}$$

The integral $x(t) = \int_0^t v(s)ds$ represents a Riemann sum $\sum_{m=0}^{n} v(mds)\, ds$ with $(n+1)dt = t$.

Observation 2: The solution of the IVP, with possibly non-zero initial value x^0, of finding $x : [0, T] \to \mathbb{R}$ such that

$$\dot{x}(t) = v(t) \quad \text{for } 0 < t \leq T, \quad x(0) = x^0, \tag{62.8}$$

is given by

$$x(t) = x^0 + \int_0^t v(s)ds. \tag{62.9}$$

This is because the derivative of a constant function (the function $w(t) = x^0$), is zero ($\dot{w} = 0$).

Observation 3: Since $\dot{x}(t) = v(t)$ and $\dot{v}(t) = a(t)$ with $x(t)$ distance, $v(t)$ velocity and $a(t)$ accelleration, we can say that

- *distance is the integral of velocity*,
- *velocity is the integral of accelleration*.

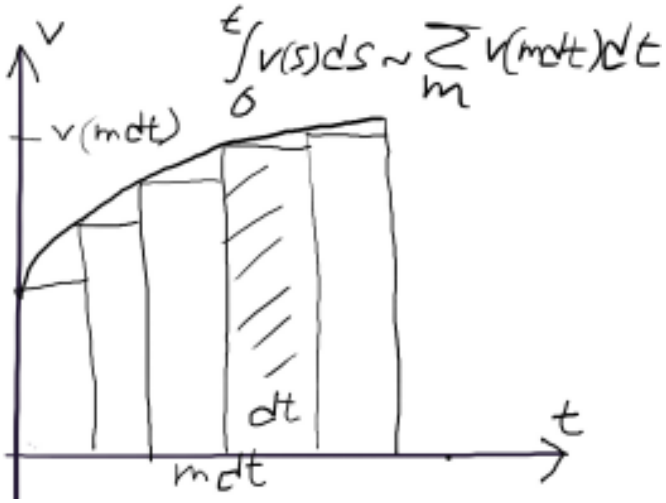


FIGURE 62.1. Integral as Riemann sum as area under graph.

62.2 Interpreting the Integral as an Area

The area $A(v,t)$ bounded by the graph of the function $v : [0,t] \to \mathbb{R}$ and the s-axis of a (v,s)-coordinate system, can viewed as a sum of rectangular strips of height $v(mdt)$ and width dt (asssuming for definiteness that $v(mdt) \geq 0$), and thus

$$A(v,t) = \sum_{m=0}^{n} v(mds)ds, \quad t = (n+1)ds. \tag{62.10}$$

We are thus led to interprete the integral as an area:

$$\int_0^t v(s)ds = A(v,t) = \text{ area under the graph of } v(s) \text{ on the interval } [0,t] \tag{62.11}$$

as illustrated in Fig. 60.1.

62.3 The Trapezoidal Rule

Replacing the shaded rectangle area in Fig. 60.1 with the area of a trapezoid right vertical of length $v((m+1)dt$ as illustrated in Fig. 60.2, we obtain the alternative Riemman sum approximation

$$\int_0^t v(s)ds \approx \sum_{m=0}^{n-1} \frac{v(mds)+v((m+1)ds)}{2}ds = \frac{v(0)}{2}ds + \sum_{m=1}^{n-1} v(mds)ds + \frac{v(t)}{2}ds. \tag{62.12}$$

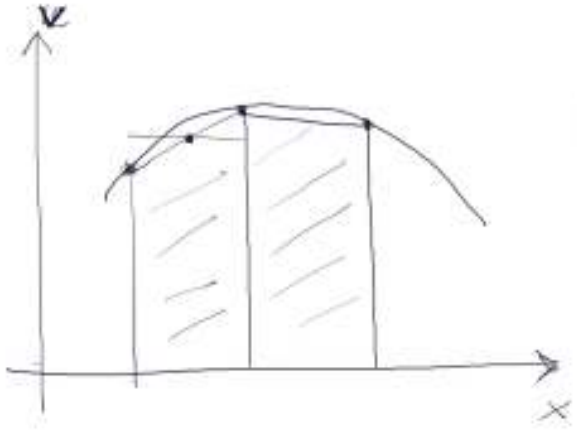


FIGURE 62.2. Piecewise linear approximation of the Trapezoidal Rule vs piecewise constant approximation of Euler Midpoint. Piecewise linear approximation is a basic element of computational mathematics including the finite elemeent method, as you will discover below...Simple and profound...

which is the *Trapezoidal Rule*. We compare with a *Midpoint Euler* method defining the height of the rectangle to be the function value at the midpoint of the interval. This value is close to the mean-value of the endpoint values used in the Trapezoidal Method, which thus is close to Midpoint Euler.

62.4 Not All Integrals are Areas

Note that distance is the integral of velocity but it is not very natural to say that distance is the area under the velocity graph.

Summing up: The integral is defined as the solution to an IVP. Some integrals can be interpreted as areas, but all integrals are not areas. Some cars (integrals) are Volvos (areas) but all cars (integrals) are not Volvos (areas). There are also Saabs...

Nevertheless, many Calculus books introduce the integral as the area under a graph, based on the pedagocial idea to define a new concept (the integral) in terms of something supposedly more familiar (area), but this is questionable from mathematical point of view and also confusing, when students discover that all integrals are not areas. To say that an integral is solution to an IVP, is not questionable, because this is what an integral *is*.

FIGURE 62.3. IVP of Usain Bolt

62.5 Watch

- Jesse Owens 1936 IVP: 100 on 10.3 sec
- Usain Bolt 2009 IVP: 100 m on 9.58 sec

63
The Fundamental Theorem of Calculus

63.1 Integration as Inverse of Differentation

The formula (62.7) is referred to as the *Fundamental Theorem of Calculus*: Integration of the function $v(t)$ followed by differentiation, gives back the function $v(t)$:

$$\frac{d}{dt}\int_0^t v(s)ds = v(t) \quad \text{for } t > 0. \tag{63.1}$$

Alternatively, The Fundamental Theorem of Calculus can be expressed as

$$\int_0^t \dot{u}(s)ds = u(t) \quad \text{for } t > 0, \tag{63.2}$$

stating: Integration of the derivative $\dot{u}(t)$ of the function $u(t)$, gives back the function $u(t)$. This follows from the fact that the derivative with respect to t of both sides of (63.2) equals $\dot{u}(t)$, combined with the fact that two functions with the same derivative taking the same value for $t = 0$, must coincide. Two cars traveling with the same velocity starting at the same time from the same location will arrive at the same time to the destination. Right?

We shall see that (63.2) can be viewed to express the following identity:

$$\text{The sum } (\int_0^t \text{ or } \sum_{m=0}^{n}) \quad \text{of the parts } (du = \dot{u}ds) \quad = \quad \text{the whole } (u(t)). \tag{63.3}$$

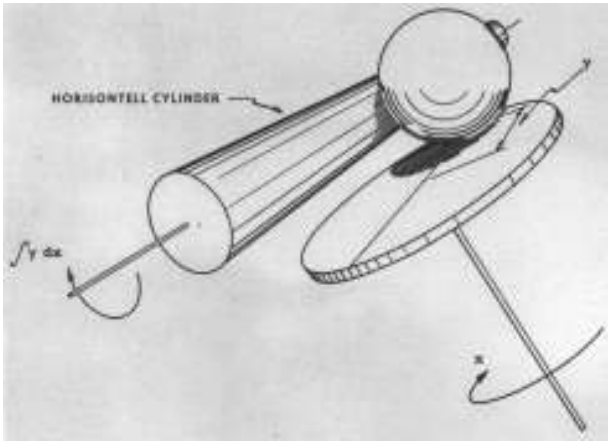


FIGURE 63.1. Analog mechanical integrator computing the integral $\int y(x)dx$ of a funtion $y(x)$. Can you explain how it works?

Integration means summing little pieces to make up the whole. In Leibniz notation this is expressed as

$$\int_0^t \frac{du}{ds}ds = \int_0^t du = u(t) - u(0). \tag{63.4}$$

Elementary and profound.

Below we shall study the dependence of the integral $\int_0^t v(s)ds$ of a given function $v(s)$ on the time step ds, and see that it is a uniquely determined number for vanishing time step, which is approximated using a finite time step, with accuracy depending on the variation of the function $f(t)$ with t. We will thus give a mathematical analysis of the meaning of the Fundamental Theoreom of Calculus, which we will refer to as a *mathematical proof* of the Fundamental Theorem.

This experience will illustrate the role and meaning of a mathematical proof as a process of dissecting the structure and meaning of a certain mathematical statement.

63.2 Read More

- Short Course in Calculus
- The Fundamental Theorem of Calculus

63.3 To Think About

- What could it mean to *prove* the Fundamental Theorem?
- Other interpretations of the whole = sum of parts?
- Suppose $u(T) = u(0) = 0$. What then about $\int_0^T u'(t)\,dt$?

63.4 Watch

- Babbages Difference Engine No. 2
- Leibniz binary ball computer
- $\sqrt{2}$ pepper grinder
- Kraftwerk Pocket Calculator
- Kraftwerk Numbers
- Computer World

> And as in arithmetic unpractised men must, and professors themselves may often, err, and cast up false; so also in any other subject of reasoning, the ablest, most attentive, and most practised men may deceive themselves, and infer false conclusions; not but that reason itself is always right reason, as well as arithmetic is a certain and infallible art: but no one man's reason, nor the reason of any one number of men, makes the certainty; no more than an account is therefore well cast up because a great many men have unanimously approved it. And therefore, as when there is a controversy in an account, the parties must by their own accord set up for right reason the reason of some arbitrator, or judge, to whose sentence they will both stand, or their controversy must either come to blows, or be undecided, for want of a right reason constituted by Nature; so is it also in all debates of what kind soever: and when men that think themselves wiser than all others clamour and demand right reason for judge, yet seek no more but that things should be determined by no other men's reason but their own, it is as intolerable in the society of men, as it is in play after trump is turned to use for trump on every occasion that suit whereof they have most in their hand. For they do nothing else, that will have every of their passions, as it comes to bear sway in them, to be taken for right reason, and that in their own controversies: bewraying their want of right reason by the claim they lay to it. (Leviathan, Thomas Hobbes)

64
The Fundamental Theorem Game

> The two questions, the first that of finding the description of the curve from its elements, the second that of finding the figure from the given differences, both reduce to the same thing. From this it can be taken that the whole of the theory of the inverse method of the tangents is reducible to quadratures. (Leibniz 1673)

64.1 Game

One player gives a function $f : [0, T] \to \mathbb{R}$ and the other player is supposed to compute its integral $u(t) = \int_0^t f(s)ds$ for $t \in [0, T]$, or simply the value $u(T)$, as quickly as possible. Return by giving a new $f(t)$ to integrate.

64.2 Mathematics

Solve the IVP $\dot{u} = f$ by time stepping.

64.3 Demo + Lab

- Test, Modify and Create Yourself (fundamental)

65
Integrals of Polynomial Functions t^p

> Does anyone believe that the difference between the Lebesgue and Riemann integrals can have physical significance, and that whether say, an airplane would or would not fly could depend on this difference? If such were claimed, I should not care to fly in that plane. (Richard Hamming)
>
> Nature laughs at the difficulties of integration. (Laplace)

65.1 Derivatives and Integrals of Polynomials

If $v(s) = s^p$, then with $x(0) = 0$ we have

$$x(t) = \frac{t^{p+1}}{p+1} = \int_0^t s^p ds, \quad p = 0, 1, ... \tag{65.1}$$

To prove this, we note that for $p = 0$ we have

$$t = \int_0^t 1 ds, \tag{65.2}$$

which is the same as

$$\frac{d}{dt}t = \frac{dt}{dt} = 1. \tag{65.3}$$

To see this we note that if $x(t) = t$, then $dx(t) = x(t + dt) - x(t) = t + dt - t = dt$ and thus $\frac{dx}{dt} = 1$.

For $p = 1$ we have

$$\frac{t^2}{2} = \int_0^t s ds, \tag{65.4}$$

which is the same as

$$\frac{d}{dt}t^2 = 2t. \tag{65.5}$$

To see this we note that if $x(t) = t^2$, then

$$dx(t) = x(t+dt)-x(t) = (t+dt)(t+dt)-t^2 = t^2+2tdt+dtdt-t^2 = 2tdt-dtdt.$$

If now dt is small then we can argue that $dtdt$ is so small that it can be neglected, and thus $\frac{d}{dt}t^2 = 2t$.

For $p = 2$ we have

$$\frac{t^3}{3} = \int_0^t s^2 ds, \tag{65.6}$$

which is the same as

$$\frac{d}{dt}t^3 = 3t^2. \tag{65.7}$$

To see this we note that if $x(t) = t^3$, then

$$\begin{aligned} dx(t) &= x(t+dt) - x(t) = (t+dt)(t+dt)(t+dt) - t^2 \\ &= t^3 + 3t^2dt + 3tdtdt + dtdtdt - t^3 = 3t^2dt + 3tdtdt + dtdtdt. \end{aligned}$$

If now dt is small then we can argue that $3tdtdt$ and $dtdtdt$ are so small that they can be neglected, and thus $\frac{d}{dt}t^3 = 3t^2$. Similarly, one can show that

$$\frac{d}{dt}t^p = pt^{p-1} \quad \text{for } p = 1, 2, 3... \tag{65.8}$$

In BS we show that his formula also holds for negative exponents

$$\frac{d}{dt}t^p = pt^{p-1} \quad \text{for } t > 0, \quad p = -1, -2, -3, ..., \tag{65.9}$$

and also for $t < 0$. In particular, we have for $t \neq 0$

$$\frac{1}{t+dt} - \frac{1}{t} = \frac{t-(t+dt)}{(t+dt)t} \approx -\frac{1}{t^2}dt, \tag{65.10}$$

which proves (65.9) for $p = 1$.

We shall also discover that the case $p = 0$ is special, and gives rise to the logarithm $\log(t)$ as the solution of

$$\dot{x}(t) = t^{-1} = \frac{1}{t}, \quad \text{for } t > 0, \quad u(1) = 0, \tag{65.11}$$

that is

$$\log(t) = \int_1^t \frac{1}{s}\, ds, \quad \text{for } t > 0. \tag{65.12}$$

The logarithm function was first constructed by the mathematician, physicist, astronomer/astrologist John Napier (1550-1617) in 1614.

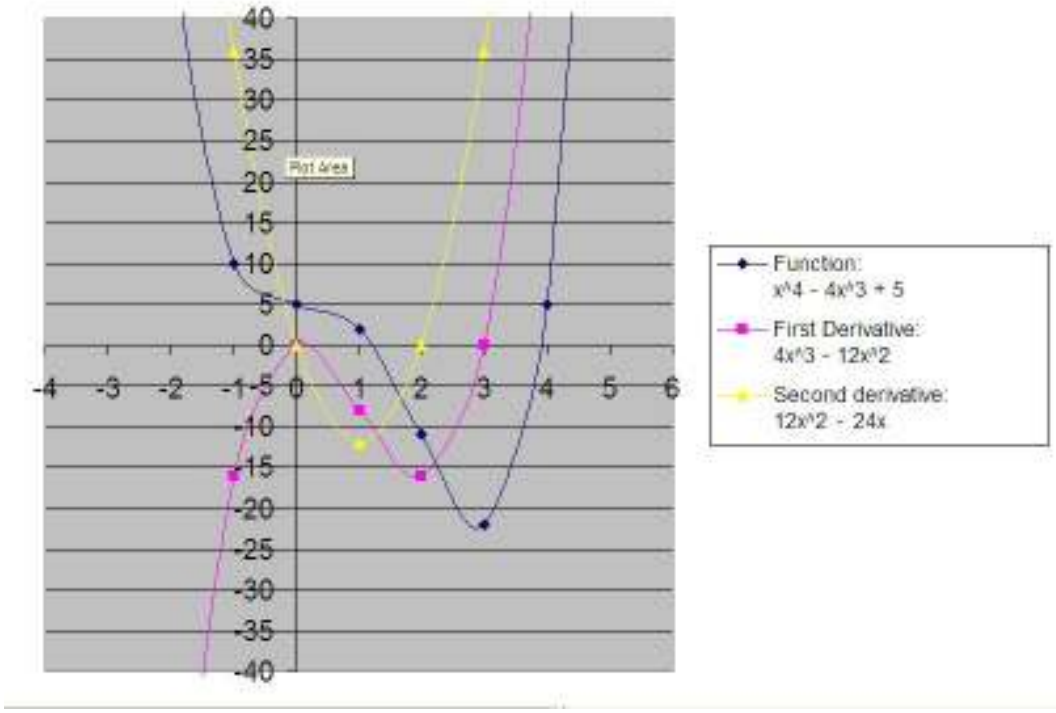


FIGURE 65.1. Derivatives of a polynomial.

65.2 To Think About

- How to prove (65.8)?

65.3 Generalization

The derivation formulas (65.8) and (65.9) generalize to

$$\frac{d}{dt}t^p = pt^{p-1} \quad \text{for } p \neq 0, \quad t \neq 0 \quad \text{for } p < 1, \tag{65.13}$$

where p is a rational (or real) number.

66
The Exponential Function $\exp(t)$

Exponential growth looks like nothing is happening, and then suddenly you get this explosion at the end. (Ray Kurzweil)

When, several years ago, I saw for the first time an instrument which, when carried, automatically records the number of steps taken by a pedestrian, it occurred to me at once that the entire arithmetic could be subjected to a similar kind of machinery so that not only addition and subtraction, but also multiplication and division could be accomplished by a suitably arranged machine easily, promptly and with sure results.... For it is unworthy of excellent men to lose hours like slaves in the labour of calculations, which could safely be left to anyone else if the machine was used.... And now that we may give final praise to the machine, we may say that it will be desirable to all who are engaged in computations which, as is well known, are the mangers of financial affairs, the administrators of others estates, merchants, surveyors, navigators, astronomers, and those connected with any of the crafts that use mathematics. (Leibniz)

66.1 Defining Differential Equation

The solution to $\frac{dx}{dt} = v$ with $v = x$ and $x(0) = 1$, that is the solution $x(t)$ to the IVP

$$\dot{x}(t) = x(t) \quad \text{for } t > 0, \quad x(0) = 1, \tag{66.1}$$

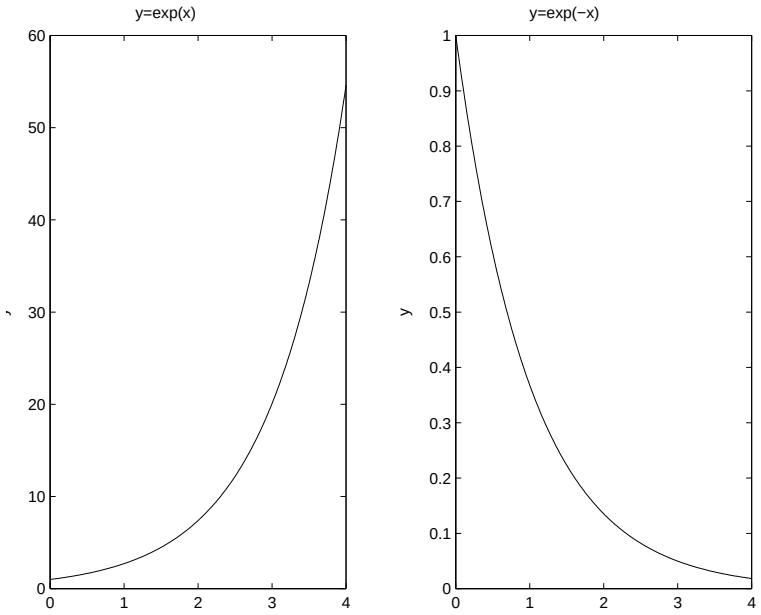


FIGURE 66.1. The exponential functions $\exp(x)$ and $\exp(-x)$ for $x \geq 0$.

is the *exponential function*

$$\exp(t) = \int_0^t \exp(s)ds, \quad \exp(0) = 1. \tag{66.2}$$

We shall see below that $\exp(t)$ extends to $t < 0$ by the same differential equation $\dot{x}(t) = x(t)$.

Listen to the amazing properties of exponential growth. from Ray Kurzweil himself.

66.2 Computing exp(t)

Updating $dx = xdt$ gives

$$x^{n+1} = x^n + x^n dt^n = (1 + dt)x^n, \quad \text{for} n = 0, 1, 2, ..., \tag{66.3}$$

and after summation, assuming $x^0 = 1$,

$$x^n = (1 + dt)^n. \tag{66.4}$$

With $t = ndt$, we thus have with Forward Euler:

$$\exp(t) \approx (1 + \frac{t}{n})^n \tag{66.5}$$

In other words, $\exp(t)$ is the compund capital after n years of interest at a yearly rate of $\frac{t}{n} = dt$ with a starting capital of 1.

66.3 Varying the Time Step

We shall see below that as n increases the approximation $\exp(t) \approx (1+\frac{t}{n})^n$ improves and can be made as small as we like. Decreasing the time step $\frac{t}{n}$ in the formula

$$\exp(t) \approx (1 + \frac{t}{n})^n \tag{66.6}$$

by increasding n, we thus obtain for $t = 1$:

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

Increasing n we can this way compute any number of decimals of the number $e = \exp(1)$, or more generally of $\exp(t)$ for any $t > 0$.

66.4 Properties of $\exp(t)$

Properties of the exponential function $\exp(t)$ can be derived from the defining differential equation $\dot{u}(t) = u(t)$. For example, the basic property of the exponential function

$$\exp(t + s) = \exp(t)\exp(s) \tag{66.7}$$

follows by noting that $u(t) = \exp(t+s)$ viewed as a a function of t, satisfies

$$\dot{u}(t) = u(t) \quad \text{for } t > 0, u(0) = \exp(s),$$

which means that $u(t) = \exp(s)\exp(t)$, which proves (66.7).

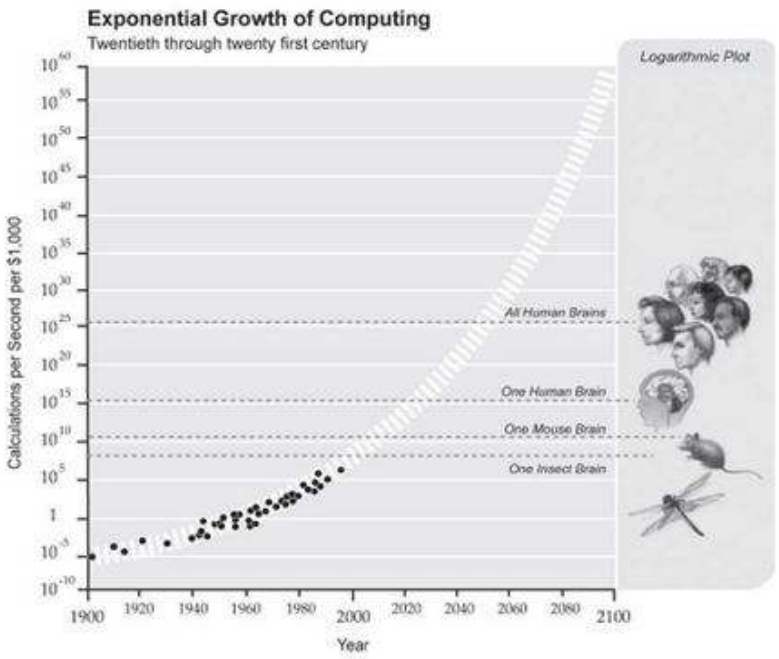


FIGURE 66.2. Exponential growth of computing power.

66.5 The Exponential exp(t) for $t < 0$

The exponential $\exp(-t)$ is the solution $x(t)$ of the IVP

$$\dot{x}(t) = -x(t) \quad \text{for } t > 0, \quad x(0) = 1, \tag{66.8}$$

If we here change variable writing $s = -t$, we have $\frac{dx}{dt} = -\frac{dx}{ds}$ and thus the IVP can also be written

$$\dot{x}(s) = \frac{dx}{ds} = x(s) \quad \text{for } s < 0, \quad x(0) = 1, \tag{66.9}$$

which extends the original definition (66.1) to $t < 0$, as we announced.

We conclude that if $f(t) = \exp(t)$, the $D(f) = \mathbb{R}$ and $R(f) = \mathbb{R}+$. (Why is $\exp(t) > 0$ for all t?)

66.6 Read More

- The Exponential.

In particular you find here proofs of the basic properties of $\exp(x)$:

$$\exp(a+b) = \exp(a)\exp(b), \quad \exp(a)^r = \exp(ra), \tag{66.10}$$

or

$$e^{a+b} = e^a e^b, \quad (e^a)^r = e^{ra}. \tag{66.11}$$

66.7 To Think About

- Why is usually the growth of an economy/BNP measured in per cent?
- Is there exponential growth in Nature?

66.8 Watch

- The Most Important Video You Will Ever See
- Exponential Growth
- Bacteria Growth

> Even if I knew nothing of the atoms, I would venture to assert on the evidence of the celestial phenomena themselves, supported by many other arguments, that the universe was certainly not created for us by divine power: it is so full of imperfectio. On the Nature of Things, Lucretius.
>
> Nothing can be created out of nothing. (Lucretius)
>
> ...explain by what forces nature steers the courses of the Sun and the journeyings of the Moon, so that we shall not suppose that they run their yearly races between heaven and earth of their own free will [i.e., are gods themselves] or that they are rolled round in furtherance of some divine plan.... (Lucretius' Physics)
>
> Let us now take as our theme the cause of stellar movements. First let us suppose that the great globe of the sky itself rotates.... There remains the alternative possibility that the sky as a whole is stationary while the shining constellations are in motion. This may happen because swift currents of ether ... whirl round and round and roll their fires at large across the nocturnal regions of the sky. Or an external current of air from some other quarter may whirl them along in their course. Or they may swim of their own accord, each responsive to the call of its own food, and feed their fiery bodies in the broad pastures of the sky. One of these causes must certainly operate in our world.... But to lay down which of them it is lies beyond the range of our stumbling progress. (Lucretius' Physics)

67
$t = \log(u)$ as the inverse of $u = \exp(t)$

> The spectacular thing about Johnny [von Neumann] was not his power as a mathematician, which was great, or his insight and his clarity, but his rapidity; he was very, very fast. And like the modern computer, which no longer bothers to retrieve the logarithm of 11 from its memory (but, instead, computes the logarithm of 11 each time it is needed), Johnny didn't bother to remember things. He computed them. You asked him a question, and if he didn't know the answer, he thought for three seconds and would produce and answer. (Paul R. Halmos)

The function $u = u(t) = \exp(t)$ satisfies

$$\frac{du}{dt} = u, \quad u(0) = 1, \tag{67.1}$$

which we can rewrite as

$$\frac{dt}{du} = \frac{1}{u}, \tag{67.2}$$

that is, recalling (201.2), $t = \log(u)$, since $t = 0$ for $u = 1$. In other words.

$$u = \exp(t) \quad \text{if and only if } t = \log(u) \tag{67.3}$$

which means that $u = \exp(t)$ and $t = \log(u)$ are *inverse functions*. We thus have for $t, u > 0$

$$\log(\exp(t)) = t, \quad \exp(\log(u)) = u. \tag{67.4}$$

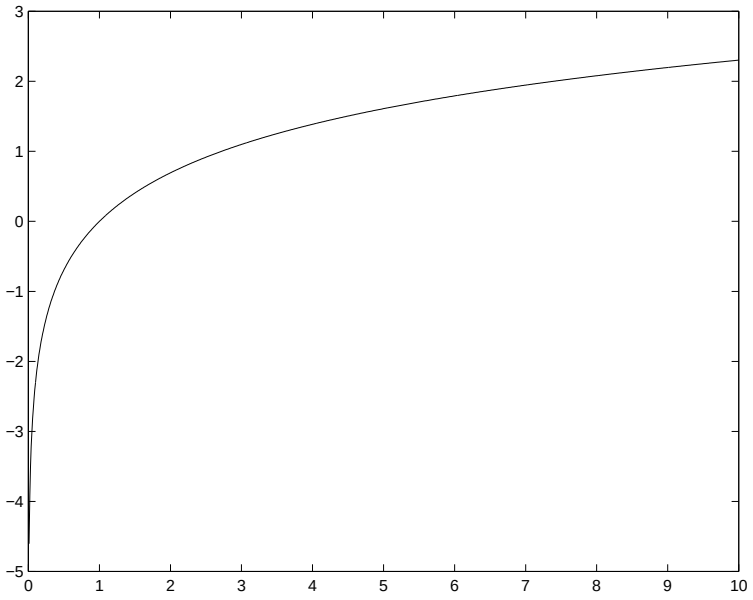


FIGURE 67.1. The logarithm $\log(x)$ for $x > 0$.

Writing

$$e^t = \exp(t) \tag{67.5}$$

we refer to $e = e^1 = \exp(1)$ as the *base of the natural logarithm* with exponent $t = \log(e^t)$.

Since by definition

$$\frac{d}{dx}\log(x) = x^{-1} \quad \text{for } x > 0, \tag{67.6}$$

the logarihm $\log(x)$ fills in the missing value $p = 0$ in the list (65.8 of derivatives of x^p:

$$\frac{d}{dt}x^p = x^{p-1} \quad \text{for } p = \pm 1, \pm 2, \pm 3.., \tag{67.7}$$

where we changed the name of the variable from t to x.

Note that with $p = 0$, $x^p = x^0 1$, and $\frac{d}{dt}x^0 = 0 \neq x^{-1}$.

67.1 Domain and Range of $\log(x)$

Since the domain of the function $x = \exp(t) > 0$ is $\mathbb{R}$ and range $\mathbb{R}+$, the domain of the inverse function $t = \log(x)$ is $\mathbb{R}+$ and range $\mathbb{R}$. In particular,

$$-\log(x) = -\int_1^x \frac{1}{y}\,dy = \int_x^1 \frac{1}{y}\,dy \tag{67.8}$$

increases without bound as $x > 0$ approaches 0.

67.2 To Think About

- What was the use of Napier's logarithms? Are they still used?
- What is a slide rule and how does it work?
- What are the basic rules for computing with logarithms?

67.3 Read More

- The Logarithm.

In particular you here proofs of the basic properties of $\log(x)$:

$$\log(ab) = \log(a) + \log(b), \quad \log(a^r) = r\log(a). \tag{67.9}$$

67.4 Watch

- John Napier 1
- John Naper 2
- Napier's Bones

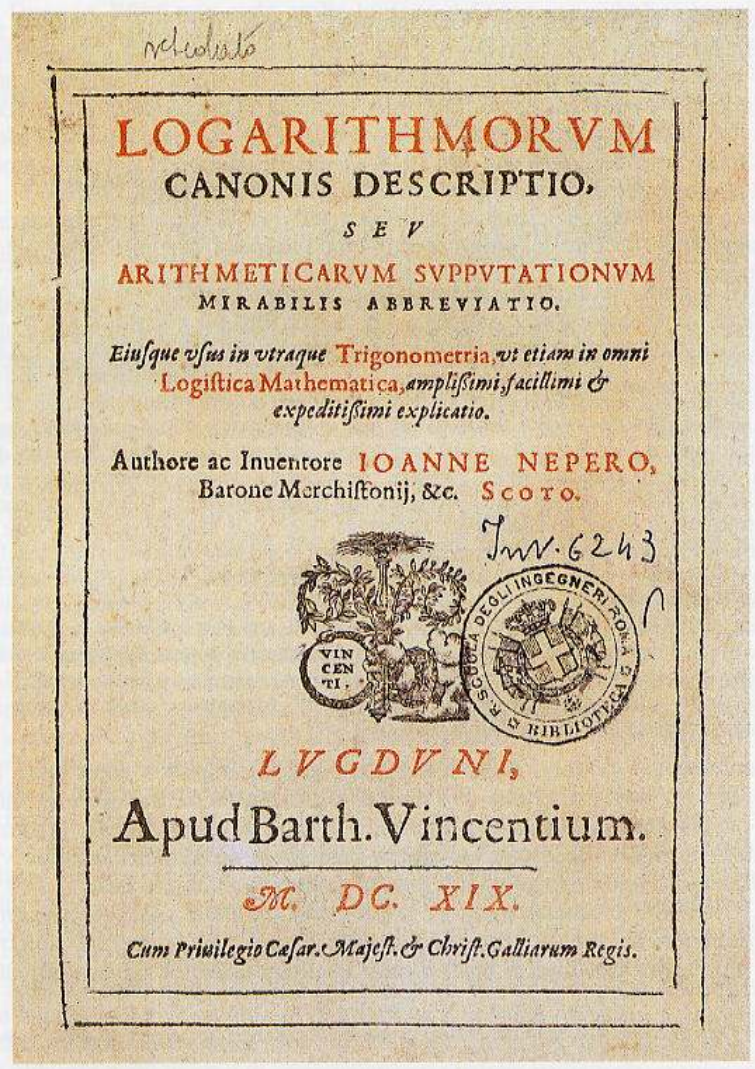


FIGURE 67.2. Title page of John Napier's logarithm tables.

68
Elementary Functions

> Elementary, my dear Watson, elementary!...It was easier to know it than to explain why I know it. If you were asked to prove that two and two made four, you might find some difficulty, and yet you are quite sure of the fact...In solving a problem of this sort, the grand thing is to be able to reason backward. That is a very useful accomplishment, and a very easy one, but people do not practise it much. In the everyday affairs of life it is more useful to reason forward, and so the other comes to be neglected. There are fifty who can reason synthetically for one who can reason analytically...How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?...I never guess. It is a shocking habit, destructive to the logical faculty... You know my methods. Apply them. (Sherlock Holmes)

In general, so called *elementary functions*, such as the exponential function, are defined as solutions of certain (basic elementary) differential equations, can be computed, decimal by decimal, by time stepping with smaller and smaller time step. There are lots ofelementary functions, since there are lots of possible more or less elementary differential equations, many of them named after famous mathematicians, including

- Bessel functions
- Legendre polynomials
- Jacobi functio

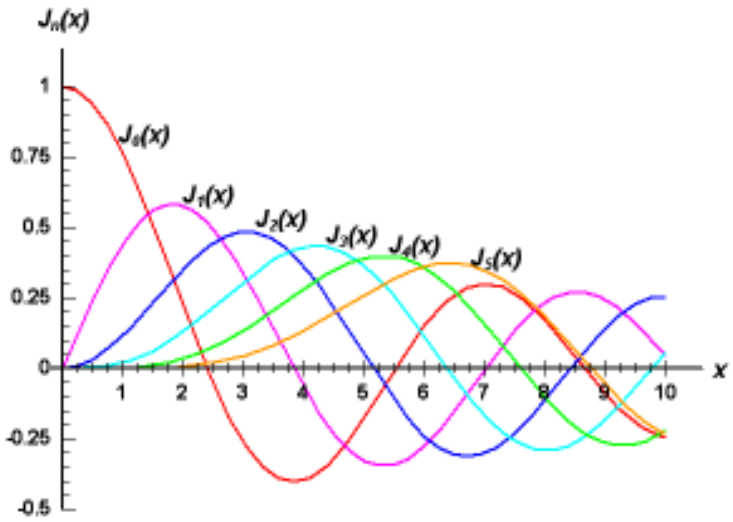


FIGURE 68.1. Bessel functions J_α.

- Hermite functions
- Laguerre functions
- Hankel functions.

For example, the Bessel functions are solutions $x(t)$ to the differential equation

$$t^2\ddot{x} + t\dot{x} + (t^2 - \alpha^2)x = 0 \tag{68.1}$$

where α is a constant, which arises for in problems with cylindrical or spherical symmetry.

The time step required to reach a certain precision or number of decimals, can vary from one differential equation and elementary function to the other. Below we shall study this problem, that is the dependence on the solution of differential equation on the time step used to compute it.

We can only compute a finite number of decimals of $\exp(t)$, as many as our computational resources allows, but we can never list all of the decimals of $\exp(t)$, expect for specific values such as $\exp(0) = 1$. We can think of $\exp(t)$ as unique number, with a possibly never repeating decimal expansion, but we should be aware of the fact that this a kind illusion because we can never pin down exactly what $\exp(t)$ is, except in the implicit form of saying that it is the function $u(t)$ with the property to solve $\dot{u}(t) = u(t)$ for $t > 0$ with $u(0) = 1$. In old times, the values of elementary functions were listed in printed mathematical tables obtained by time-stepping the

FIGURE 68.2. Hermite crab function.

corresponding differential equations. In a computer, the values of elemntary functins are not stored in tables but are recomputed every time a valu is requested.

68.1 To Think About

- How are elementary functions computed by your computer?

69

Trigonometric Functions: $\cos(t)$, $\sin(t)$

> The integrals which we have obtained are not only general expressions which satisfy the differential equation, they represent in the most distinct manner the natural effect which is the object of the phenomenon... when this condition is fulfilled, the integral is, properly speaking, the equation of the phenomenon; it expresses clearly the character and progress of it, in the same manner as the finite equation of a line or curved surface makes known all the properties of those forms. (Fourier)
>
> God does not care about our mathematical difficulties. He integrates empirically. (Einstein)

69.1 Defining Differential Equation

The *trigonometric functions* $\sin(t)$ and $\cos(t)$ are elementary functions defined as solution to $\frac{dx}{dt} = v$ and $\frac{dv}{dt} = -x$ with $x(0) = 0$ and $v(0) = 1$, or the system

$$\begin{aligned} &\dot{x}(t) = v(t), \quad \dot{v}(t) = -x(t) \quad \text{for } t > 0, \\ &x(0) = 0, \quad v(0) = 1, \end{aligned} \tag{69.1}$$

is the *trigonometric functions* $x(t) = \sin(t)$ and $v(t) = \cos(t)$. These functions can be extended to $t < 0$ as solutions to the differential equations for $t < 0$.

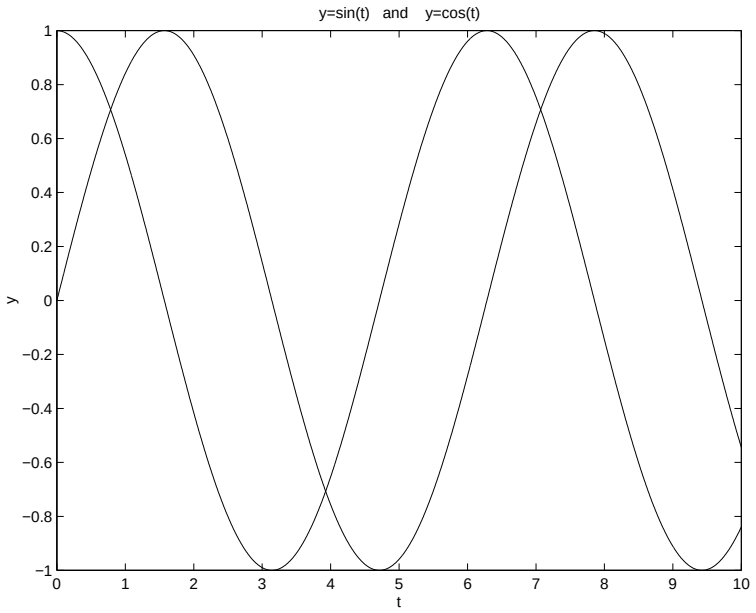


FIGURE 69.1. The trigonometric functions $\sin(t)$ and $\cos(t)$.

69.2 Properties of Trigonometric Functions

By definition, we have

$$\frac{d}{dt}\sin(t) = \cos(t), \quad \frac{d}{dt}\cos(t) = -\sin(t). \tag{69.2}$$

Further, we have with $x(t) = \sin(t)$ and $v(t) = \cos(t)$,

$$\frac{d}{dt}(x^2 + v^2) = 2(x\dot{x} + v\dot{v}) = 2(xv - xv) = 0 \tag{69.3}$$

and thus $(x^2(t) + v^2(t))$ is constant in time, and since $(x^2(0) + v^2(0)) = 1$, we have for all t

$$\sin^2(t) + \cos^2(t) = 1. \tag{69.4}$$

69.3 Geometric Interpretation

We shall now give an interpretation of $\sin(t)$ and $\cos(t)$ in the plane with usual coordinate system $x = (x_1, x_2)$. If we write denote $x_1(t) = \cos(t)$ and $x_2(t) = \sin(t)$, then the defining differential equation is written

$$\dot{x}_1(t) = -x_2(t), \quad \dot{x}_2(t) = x_1(t), \tag{69.5}$$

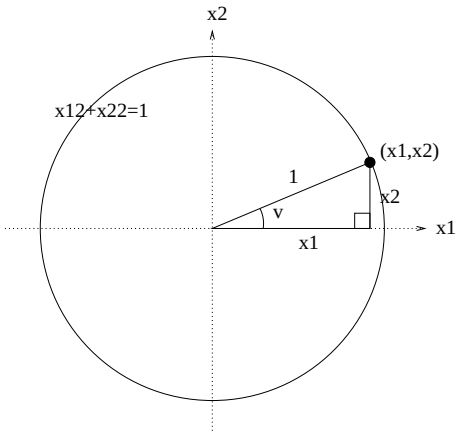


FIGURE 69.2. Geometric interpretation of trigonometric functions $\sin(t)$ and $\cos(t)$.

from which follows that

$$\dot{x}(t) \cdot x(t) = -x_2(t)x_1(t) + x_1(t)x_2(t) = 0 \tag{69.6}$$

which means that the *velocity vector* $\dot{x}(t)$ is perpendicular to the vector x connecting the origin with the point x. Recalling that

$$x_1^2(t) + x_2^2(t) = \dot{x}_2^1(t) + \dot{x}_1^2(t) = 1 \tag{69.7}$$

it follows that as t varies the point $(x_1(t), x_2(t)) = (\cos(t), \sin(t))$ moves along a unit circle centered at the origin with unit speed, as illustrated in Fig. 69.2. We can choose time t to be a measure of the angle, from the horisontal.

Note that we can interprete (69.7) as Pythagoras Theorem. We have thus given a proof of Pythagoras theorem which is different from that suggested in Fig. **??** based on similarity.

69.4 Measuring Angles in Radians

We have seen that $(\cos(t), \sin(t))$ are the coordinates of a point moving counterclockwise on the unit circle with unit velocity starting at $(1, 0)$ for $t = 0$. Let us denote by $\frac{\varphi}{2}$ the first smallest t for which $\cos(t)) = 0$, and by (69.7) $\sin(t) = 1$. By periodicity it follows that for $t = 2\pi$ the point will be back to $(1, 0)$ and thus the length of the circumference is equal to of a unit circle is equal to 2π. If we agree to measure the angle formed by the line from the origin to the point $(\cos(t), \sin(t)$ by $t =$ the length of the circle arc from $(1, 0)$ to $(\cos(t), \sin(t)$ then we measure the angle in the unit of

radians. One revolution will then correspond to 2π radians. In other words $360^o = 360$ degrees $= 2\pi$ radians, or

$$\text{one degree} = 1^o = \frac{\pi}{180} \quad \text{radians} \tag{69.8}$$

We shall see below that the choice of $y = (0, 1)$ in fact covers the general case (since a general vetor y of length 1 can be rotated to $(0, 1)$ by an orthogonal transformation which does not change the scalar product).

69.5 Angle vs Scalar Product

Let $x = (x_1, x_2)$ and $y = (1, 0)$ be two points in the plane with corresponding vectors (or arrows) from the origin also denoted by $x = (x_1, x_2)$ and $y = (1, 0)$. Since

$$x \cdot y = x_1 = |x| \cos(\theta) \tag{69.9}$$

where θ is the angle in radians between x and y. This formula extends to any two vectors x and y:

$$x \cdot y = |x||y| \cos(\theta), \tag{69.10}$$

where θ is the angle between the vectors.

69.6 Read More

- Pythagoras.
- Trigonometric functions.
- Geometry in $\mathbb{R}^2$.
- Complex numbers.

69.7 To Think About

- How are trigonomteric functions defined in standard Calculus texts?

x			sin(x)	cos(x)	tan(x)	cot(x)	csc(x)	sec(x)
deg	rad							
0	0		0.0000	1.0000	0.0000			1.0000
1	0.0175		0.0175	0.9998	0.0175	57.2900	57.2987	1.0002
2	0.0349		0.0349	0.9994	0.0349	28.6363	28.6537	1.0006
3	0.0524		0.0523	0.9986	0.0524	19.0811	19.1073	1.0014
4	0.0698		0.0698	0.9976	0.0699	14.3007	14.3356	1.0024
5	0.0873		0.0872	0.9962	0.0875	11.4301	11.4737	1.0038
6	0.1047		0.1045	0.9945	0.1051	9.5144	9.5668	1.0055
7	0.1222		0.1219	0.9925	0.1228	8.1443	8.2055	1.0075
8	0.1396		0.1392	0.9903	0.1405	7.1154	7.1853	1.0098
9	0.1571		0.1564	0.9877	0.1584	6.3138	6.3925	1.0125
10	0.1745		0.1736	0.9848	0.1763	5.6713	5.7588	1.0154
11	0.192		0.1908	0.9816	0.1944	5.1446	5.2408	1.0187
12	0.2094		0.2079	0.9781	0.2126	4.7046	4.8097	1.0223
13	0.2269		0.2250	0.9744	0.2309	4.3315	4.4454	1.0263
14	0.2443		0.2419	0.9703	0.2493	4.0108	4.1336	1.0306
15	0.2618		0.2588	0.9659	0.2679	3.7321	3.8637	1.0353
16	0.2793		0.2756	0.9613	0.2867	3.4874	3.6280	1.0403
17	0.2967		0.2924	0.9563	0.3057	3.2709	3.4203	1.0457
18	0.3142		0.3090	0.9511	0.3249	3.0777	3.2361	1.0515
19	0.3316		0.3256	0.9455	0.3443	2.9042	3.0716	1.0576
20	0.3491		0.3420	0.9397	0.3640	2.7475	2.9238	1.0642
21	0.3665		0.3584	0.9336	0.3839	2.6051	2.7904	1.0711
22	0.384		0.3746	0.9272	0.4040	2.4751	2.6695	1.0785
23	0.4014		0.3907	0.9205	0.4245	2.3559	2.5593	1.0864
24	0.4189		0.4067	0.9135	0.4452	2.2460	2.4586	1.0946
25	0.4363		0.4226	0.9063	0.4663	2.1445	2.3662	1.1034
26	0.4538		0.4384	0.8988	0.4877	2.0503	2.2812	1.1126
27	0.4712		0.4540	0.8910	0.5095	1.9626	2.2027	1.1223
28	0.4887		0.4695	0.8829	0.5317	1.8807	2.1301	1.1326
29	0.5061		0.4848	0.8746	0.5543	1.8040	2.0627	1.1434
30	0.5236		0.5000	0.8660	0.5774	1.7321	2.0000	1.1547
31	0.5411		0.5150	0.8572	0.6009	1.6643	1.9416	1.1666
32	0.5585		0.5299	0.8480	0.6249	1.6003	1.8871	1.1792
33	0.576		0.5446	0.8387	0.6494	1.5399	1.8361	1.1924
34	0.5934		0.5592	0.8290	0.6745	1.4826	1.7883	1.2062
35	0.6109		0.5736	0.8192	0.7002	1.4281	1.7434	1.2208
36	0.6283		0.5878	0.8090	0.7265	1.3764	1.7013	1.2361
37	0.6458		0.6018	0.7986	0.7536	1.3270	1.6616	1.2521
38	0.6632		0.6157	0.7880	0.7813	1.2799	1.6243	1.2690
39	0.6807		0.6293	0.7771	0.8098	1.2349	1.5890	1.2868
40	0.6981		0.6428	0.7660	0.8391	1.1918	1.5557	1.3054
41	0.7156		0.6561	0.7547	0.8693	1.1504	1.5243	1.3250
42	0.733		0.6691	0.7431	0.9004	1.1106	1.4945	1.3456
43	0.7505		0.6820	0.7314	0.9325	1.0724	1.4663	1.3673
44	0.7679		0.6947	0.7193	0.9657	1.0355	1.4396	1.3902
45	0.7854		0.7071	0.7071	1.0000	1.0000	1.4142	1.4142

www,analyzemath.com

FIGURE 69.3. Table of values of trigonometric functions.

FIGURE 69.4. Triangulations by Olle Baertling.

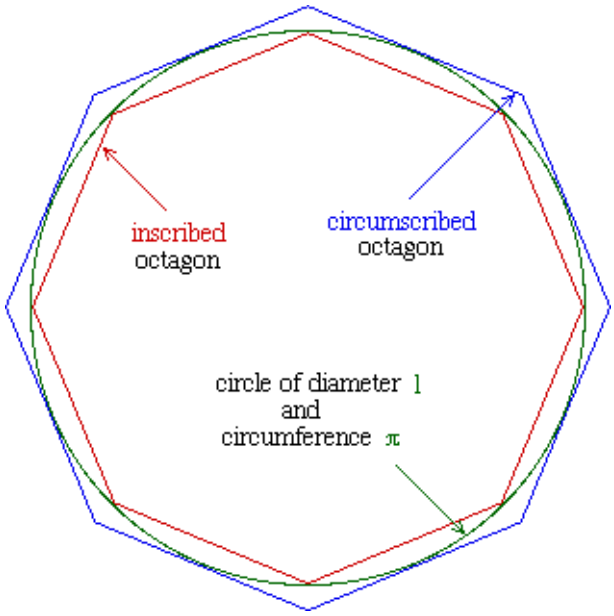


FIGURE 69.5. Archimedes computed the value of π by inscribing and circumscribing polygons (octagons) to a circle. What value did he obtain?

70
Lipschitz Continuity

> All my life as an artist I have asked myself: What pushes me continually to make sculpture? I have found the answer. art is an action against death. It is a denial of death... Imagination is a very precise thing, you know-it is not fantasy; the man who invented the wheel while he was observing another man walking-that is imagination!... I am the most curious of all to see what will be the next thing that I will do. (Jacques Lipchitz)

The Cubist sculptor Jacques Lipchitz (1891-1973) was not at all related to the German mathematician Rudolf Lipschitz (1832-1903), who is remembered for requiring the function $f(x)$ in the IVP

$$\dot{x}(t) = f(x(t)) \quad \text{for } t > 0,\ x(0) = x^0, \tag{70.1}$$

to be *Lipschitz continuous* in order to guarantee that a unique solution exists.

Definition 70.1 *A function $u : \mathbb{Q} \to \mathbb{Q}$ is said to be* Lipschitz continuous *if there is a constant L, called the* Lipchitz constant*, such that*

$$|u(t + dt) - u(t)| \leq L|dt| \quad \textit{for all } t,\ dt \in \mathbb{Q}, \tag{70.2}$$

where we here allow dt to also be negative.

For a Lipschitz continuous function $u(t)$ the difference $du = u(t + dt) - u(t)|$ is small if $|dt|$ is small, up to the constant L, in the sense that

$$|du| \leq L|dt|. \tag{70.3}$$

FIGURE 70.1. The Lipchitz sculpture *Happiness of Living* is Lipschitz continuous.

We may say that a Lipschitz continuous function $u(t)$ is locally close to a constant value in the sense that $u(t+dt)$ deviates from $u(t)$ less than $L|dt|$. By this we don't mean that $u(t)$ is close to a constant over its entire span, just locally.

One can relax the concept of Lipschitz continuity to *Hölder continuity* requiring instead for some fixed constant $0 < \alpha < 1$

$$|u(t+dt) - u(t)| \leq L|dt|^\alpha \quad \text{for all } t,\, dt. \tag{70.4}$$

Hölder continuity also expresses local constancy, but with a different measure.

In order for a function value $u(t)$ to be well defined for a given argument t, it is necessary that u is (Lipschitz or Hölder) continuous at t. This is because if t is an irrational number, then t it is not known exactly to all its decimals, and thus $u(t)$ has to be replaced by the value $u(t+dt)$ with $t+dt$ a finite decimal approximation of t, and in order for the replacement to make sense we must be able to guarantee that $u(t+dt)$ is close to $u(t)$ if dt is small.

70.1 Extension of a Lipschitz Continous Function

The previous argument can be used to show that a Lipschitz continuous function $f : \mathbb{Q} \to \mathbb{Q}$ can uniquely be extended to a Lipschitz function $f : \mathbb{R} \to \mathbb{R}$. Can you write down the argument? Hint: Is it sufficient to show that $u(t+dt)$ is close to $u(t)$ if dt is small?

The concept of Lipschitz continuity is naturally extended to a function $u : I \to \mathbb{R}$ defined on some interval or union of intervals I: A function $u : I \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L if

$$|u(t+dt) - u(t)| \leq L|dt| \quad \text{for } t, t+dt \in \mathbb{Q}, \tag{70.5}$$

Two Lipschitz continuos functions defined on two adjoining intervals $[a,b]$ and $[b,c]$ can be defined as one function defined on the union of intervals $[a,c]$ with possibly a jump discontinuity at the common point b. We can thus speak about such *piecewise Lipschitz continuous functions*, but not about discontinuos functions in general.

70.2 Extension to a Function $u(x)$

The concept of Lipschitz continuity is naturally extended to a function $u(x)$ where x is a space variable, or soem other variable, as follows:

Definition 70.2 *A function $u : \Omega \to \mathbb{R}$ where Ω is a domain in $\mathbb{R}^d$, is said to be Lipschitz continuous with Lipschitz constant L, if*

$$|u(x+dx) - u(x)| \leq L|dx| \quad \textit{for all } x,\, x+dx \in \Omega. \tag{70.6}$$

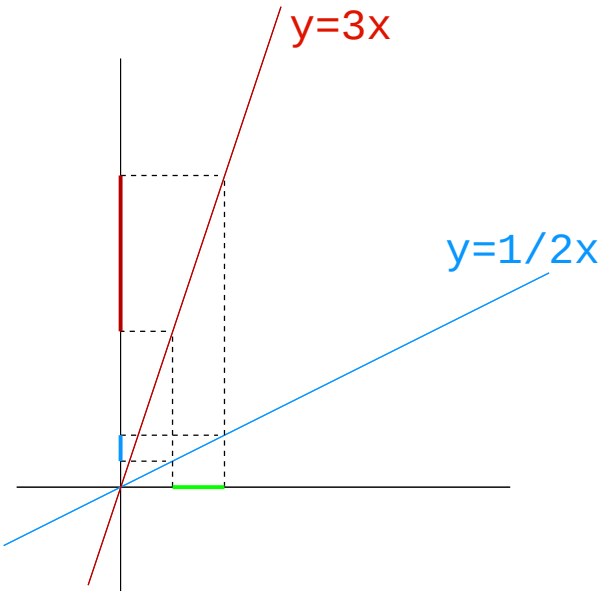


FIGURE 70.2. Two linear functions with different slope and Lipschitz constant.

As above we understand that a Lipschitz continuous function $u : \mathbb{Q}^d \to \mathbb{Q}^d$ can uniquely be extended to Lipschitz continuous function $u : \mathbb{R}^d \to \mathbb{R}^d$.

70.3 A Horrible Function which is a Non-Function

In math books you can find the following specification of values $x(t)$

- x(t)=1 if t is rational
- x(t)=0 if t is irrational.

Obviously, this is not a Lipschitz or Hölder continuous function, and in fact it is not a function at all, because you can not in general tell if a given argument t has a finite/periodic decimal expansion or not.

70.4 To Think About

- Is the sum of two Lipschitz continuous functions, Lipschitz continuous?
- What about other combinations of functions?

A Lipschitz continuous function does not change more rapidly than a linear function.

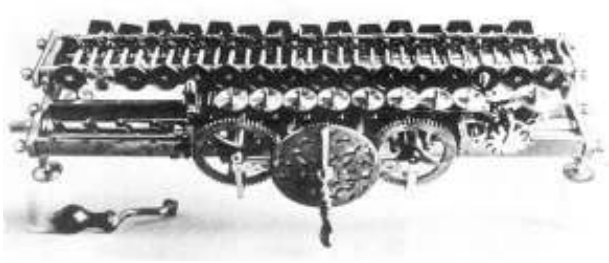


FIGURE 70.3. Leibniz' desktop computer.

70.5 Read More

- Lipschitz continuity.

70.6 Qualitative Definition of Continuity

The concept of Lipschitz continuity is a quantitative concept of continuity, with the Lipschitz constant giving quantitative control of a difference in function value du in terms of difference in argument dt say, as $|du| \leq L|dt|$, which in quantitative form expresses that

- $|du|$ is small if $|dt|$ is small.

In most standard Calculus books such a purely qualitative concept of continuity is used, which has advantage of being "more general" but at the cost of being "less precise", since it does not connect smallness of dt to smallness of du in any quantitative form. Hölder continuity is also quantitative while allowing almost any full generality.

The reason we use Lipschitz/Hölder continuity is that it is more precise than a purely qualitative concept and therefore easier to both understand and use, while the loss of generality does not have any real cost.

70.7 To Think About

- Can you give an example of a function which is not Lipschitz continuous?
- What about $u(x) = \frac{1}{x}$ defined for $0 < x \leq 1$.

71
Derivative with respect to t

Derivatives and Bank Collapse.

> But just as much as it is easy to find the derivative of a given quantity, so it is difficult to find the integral of a given derivative. Moreover, sometimes we cannot say with certainty whether the integral of a given quantity can be found or not. (Johann Bernoulli)
>
> Among all of the mathematical disciplines the theory of differential equations is the most important... It furnishes the explanation of all those elementary manifestations of nature which involve time. (Sophus Lie)

We are now ready to give a formal definition of the derivate $\dot{u}(t)$ of a function $u(t)$ depending on time t.

Definition 71.1 *A function $u : I \to \mathbb{R}$ defined on an interval $I = (a, b)$, is said to be* differentiable *in I with* derivative $\dot{u} : I \to \mathbb{R}$ *if for some positive constant C_u*

$$|u(t + dt) - u(t) - \dot{u}(t)\, dt| \leq C_u |dt|^2 \quad \text{for } t, t + dt \in I. \tag{71.1}$$

A differentiable function $u(t)$ is locally close to a linear function in t in the sense that $u(t + dt) \approx u(t) + \dot{u}(t)dt$ up to a quadratic term in $|dt|$.

A differentiable function is Lipschitz continuous, since it is locally close to a linear function and a linear function is Lipschitz continuous.

We note the following connection between Lipschitz continuity and the size the derivative.

Theorem 71.2 *If $u : I \to \mathbb{R}$ is differentiable with $|\dot{u}(t)| \leq L$ for $t \in I$, then $u : I \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L.*

Proof: This result should be intuitively obvious: It is like saying that if your velocity is never bigger than 1 km/hour, then it is impossible to travel a distance longer than 1 km in an hour. Right? If you hesitate, consider the follwoing formal proof: Given $t, s \in I$, we are supposed to show that

$$|u(t) - u(s)| \leq L|t - s|. \tag{71.2}$$

To this end let us note that it is sufficient to prove that for any given $\epsilon > 0$,

$$|u(t) - u(s)| \leq (L + \epsilon)|t - s|. \tag{71.3}$$

Now choose first $\epsilon > 0$ and then dt such that $C_u dt \leq \epsilon$ and $(t-s) = ndt$ for some natural number n, assuming $t > s$. We then have splitting the interval (s, t) into n subintervals of length dt: $(s, s + dt)$, $(s + dt, s + 2dt)$,....$(s + (n - 1)dt, s + ndt)$, and using the definition of $\dot{u}$ on each subinterval:

$$|u(s+(m+1)dt)-u(s+mdt)-\dot{u}(s+mdt)dt| \leq C_u dt^2 \quad \text{for } m = 0, ..., n-1, \tag{71.4}$$

that is, using that $|\dot{u}(s + mdt)| \leq L$,

$$|u(s + (m + 1)dt) - u(s + mdt)| \leq Ldt + C_u dt^2. \tag{71.5}$$

By the triangle inequality we now have since $ndt = t - s = |t - s|$,

$$\begin{aligned} &|u(t) - u(s)| \\ &\leq |u(s + dt) - u(s)| + |u(s + 2dt) - u(s + dt)| + ...|u(s + (n - 1)dt - u(t)| \\ &\leq n(Ldt + C_u dt^2) \leq L|t - s| + C_u dt|t - s| = (L + \epsilon)|t - s|, \end{aligned} \tag{71.6}$$

which we wanted to show. ■

EXAMPLE 71.1. The function $u(t) = |t|$ is Lipschitz continous (in particular at $t = 0$), but is not differentiable at $t = 0$ because it is not close to a linear function for t close to 0 (since it has a kink).

71.1 Read More

- Derivative.
- Rules of Differentiation.

72
Derivative with respect to x

> In the fall of 1972 President Nixon announced that the rate of increase of inflation was decreasing. This was the first time a sitting president used the third derivative to advance his case for reelection. (Hugo Rossi)
>
> Calculus required continuity, and continuity was supposed to require the infinitely little; but nobody could discover what the infinitely little might be. (Bertrand Russell)
>
> Who has not been amazed to learn that the function $u(t) = e^t$, like a phoenix rising from its own ashes, is its own derivative? (Francois le Lionnais)
>
> I recoil with dismay and horror at this lamentable plague of functions which do not have derivatives. (Charles Hermite)
>
> Senate Panel Approves Tougher Rules on Derivatives.

We now extend to a function $u(x)$ depending on position $x = (x_1, x_2, x_3)$ instead of time t.

Definition 72.1 *A function $u : \Omega \to \mathbb{R}$ where $\Omega \subset \mathbb{R}^3$, said to be* differentiable *in Ω with* derivative *or* gradient

$$\nabla u(x) = (\frac{\partial u}{\partial x_1}(x), \frac{\partial u}{\partial x_2}(x), \frac{\partial u}{\partial x_3}(x)), \tag{72.1}$$

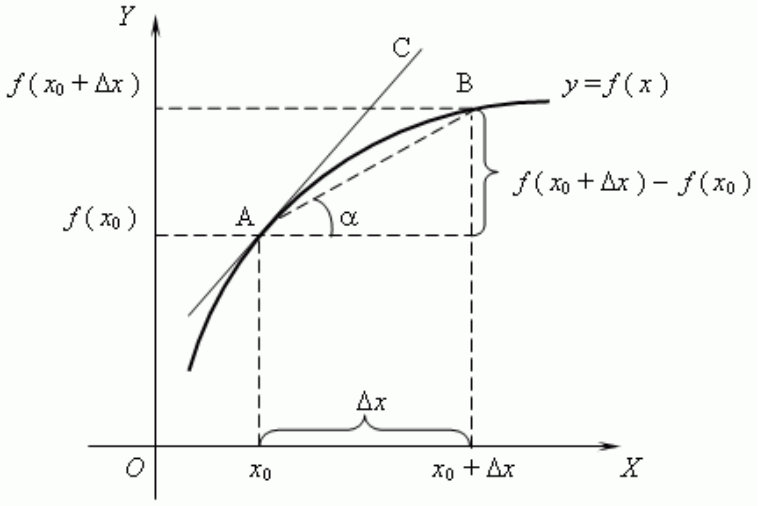


FIGURE 72.1. Seeking the derivative as the slope of the tangent for a function $f(x)$ of one variable $x \in \mathbb{R}$.

if for some positive constant C_u *and* $x \in \Omega$,

$$|u(x+dx) - u(x) - \nabla u(x) \cdot dx| \leq C_u|dx|^2 \quad \textit{for } |dx| \quad \textit{small.} \tag{72.2}$$

In particular, choosing $dx = (dx_1, 0, 0)$, we have

$$|u(x_1+dx_1, x_2, x_3) - u(x_1, x_2, x_3) - \frac{\partial u}{\partial x_1}(x)dx_1| \leq C_u|dx_1|^2 \quad \text{for } |dx_1 \quad \text{small.} \tag{72.3}$$

which means that $\frac{\partial u}{\partial x_1}(x)$ is the derivative of $f(x)$ with respect to x_1, with x_2 and x_3 kept constant, referred to as the *partial derivative* with respect to x_1.

The definition directly generalizes to real-valued function $u(x)$ of d-vector variable $x = (x_1, x_2, ..., x_d)$, where the variable components can have have some other meaning than position. In the case $d = 1$, that is with $u(x)$ a function of one variable $x \in \mathbb{R}$, we often use $u'(x)$ to denote the derivative, thus with the defining relation

$$|u(x+dx) - u(x) - u'(x)dx| \leq C_u|dx|^2 \quad \text{for } |dx| \quad \text{small.} \tag{72.4}$$

72.1 Vector-valued function of vector variable

The definition of derivative directly generalize to an m-vector-valued function $u(x) = (u_1(x), ..., u_m(x))$ of an n-vector variable $x = (x_1, x_2, ..., x_n)$:

Definition 72.2 *A function $u : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable with derivative $u'(x)$ if for som positive constant C_u*

$$|u(x + dx) - u(x) - u'(x)dx| \leq C_u|dx|^2 \quad \text{for } x,\ x + dx \in \mathbb{R}^n. \tag{72.5}$$

Here the derivative $u'(x)$ is an $m \times n$ matrix.

We shall use this derivative below when solving an equation $u(x) = 0$ where $u : \mathbb{R}^n \to \mathbb{R}^n$ is a differentiable function with non-singular derivative $u'(x)$ using Newton's method.

72.2 Read More

- Calculus of Several Variables.

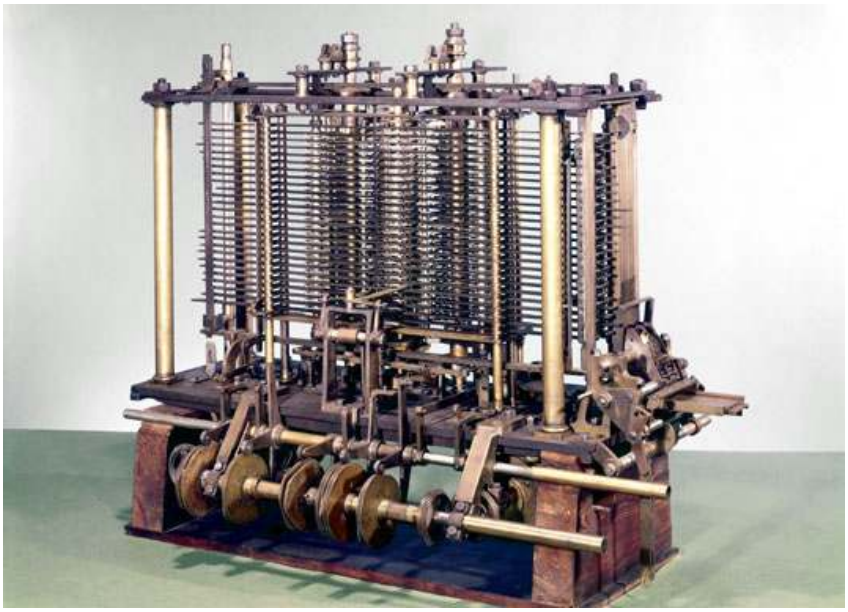


FIGURE 72.2. Charles Babbage's Analytical Engine 1871.

73
Rules of Differentiation

The ultimate reason of things must lie in a necessary substance, in which the differentiation of the changes only exists eminently as in their source; and this is what we call God. (Leibniz)

The freedom of thought is a sacred right of every individual man, and diversity will continue to increase with the progress, refinement, and differentiation of the human intellect. (Felix Adler)

How did Biot arrive at the partial differential equation? [the heat conduction equation] . . . Perhaps Laplace gave Biot the equation and left him to sink or swim for a few years in trying to derive it. That would have been merely an instance of the way great mathematicians since the very beginnings of mathematical research have effortlessly maintained their superiority over ordinary mortals. (Clifford Truesdell)

Common integration is only the memory of differentiation. (Augustus De Morgan)

73.1 Derivative of a Linear Combination

We have directly from the definition

$$\frac{d}{dt}(u+v) = \frac{d(u+v)}{dt} = \frac{du}{dt} + \frac{dv}{dt} \tag{73.1}$$

and if α is a constant

$$\frac{d}{dt}(\alpha u) = \frac{d(\alpha u)}{dt} = \alpha \frac{du}{dt}. \tag{73.2}$$

Combining these results we have

$$\frac{d}{dt}(\alpha u + \beta v) = \alpha \frac{du}{dt} + \beta \frac{dv}{dt} \tag{73.3}$$

where α and β are constants.

73.2 Derivative of Product

If $u(t)$ and $v(t)$ are real-valued differentiable functions of t, then

$$d(uv) = udv + vdu \tag{73.4}$$

or

$$\frac{d}{dt}(uv) = \frac{du}{dt}v + u\frac{dv}{dt} = \dot{u}v + u\dot{v}, \tag{73.5}$$

because

$$\begin{aligned}
&|d(uv) - udv - vdu| = |u(t+dt)v(t+dt) - u(t)v(t) - u(t)\dot{v}dt - v\dot{u}dt| \\
&= |u(t+dt)(v(t+dt) - v(t)) - udv + v(t)(u(t+dt) - u(t)) - vdu| \\
&\leq |u(t+dt)(v(t+dt) - v(t)) - udv| + |v(t)(u(t+dt) - u(t)) - vdu| \\
&\leq C|dt|^2.
\end{aligned} \tag{73.6}$$

73.3 Derivative of a Quotient

We compute the derivative of the quotient $\frac{1}{v(t)}$ assuming $v(t) \neq 0$ is differentiable:

$$\frac{1}{v(t+dt)} - \frac{1}{v(t)} = -\frac{v(t+dt) - v(t)}{v(t+dt)v(t)} \approx -\frac{\dot{v}(t)}{u(t)^2}dt \tag{73.7}$$

up to a quadratic deviation in $|dt|$. Thus, by combination with the previous result,

$$\frac{d}{dt}\frac{u}{v} = \frac{\dot{u}v - u\dot{v}}{v^2}. \tag{73.8}$$

73.4 The Chain Rule

If $v : \mathbb{R} \to \mathbb{R}$ and $w : \mathbb{R} \to \mathbb{R}$ are two differentiable functions, then the composite function $u(x) = v(w(x))$ is differentiable with derivative

$$u'(x) = v'(w(x))w'(x) \quad \text{or} \quad \frac{du}{dx} = \frac{dv}{dw}\frac{dw}{dx}. \tag{73.9}$$

To see this we estimate

$$\begin{aligned} u(x+dx) - u(x) = v(w(x+dx) - v(w(x)) &\approx v'(w(x))(w(x+dx) - w(x)) \\ &\approx v'(v(x))w'(x)dx \end{aligned} \tag{73.10}$$

up to terms of order $|dx|^2$.

73.5 Read More

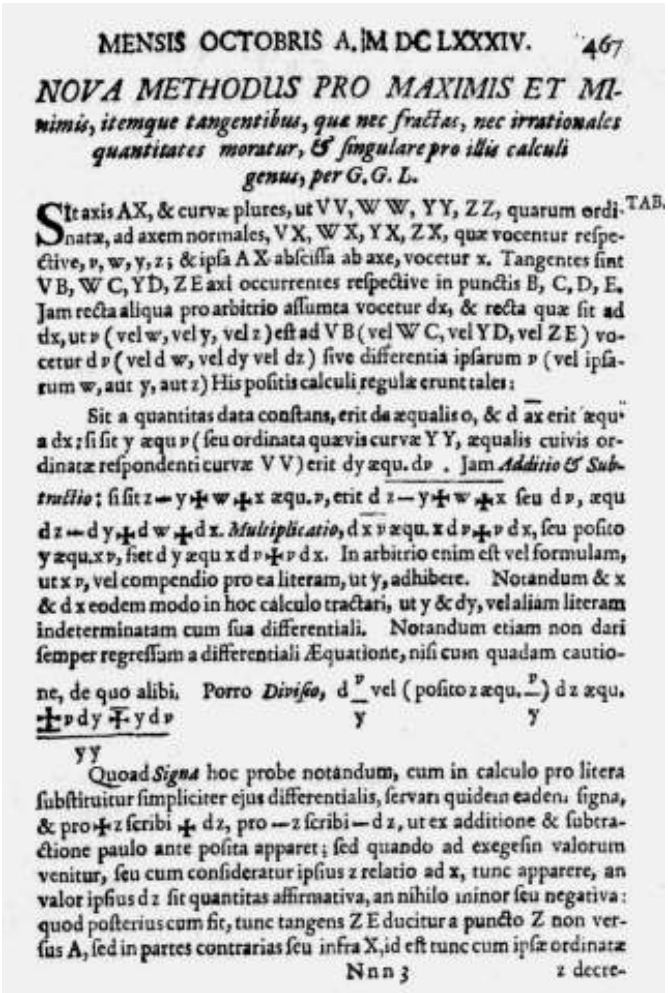
- Rules of Differentiation.
- Rules of Integration
- Differentiation Rules

73.6 To Think About

- How sensitive is differentiation to perturbations of function values?
- How to compute thye derivate of a function if analytic differentiation is not an option, because it is too difficult or the function is not given by an analytic expression?

73.7 Watch

- A no-where differentiable Weierstrass function
- Weierstrass rap
- Perelman million dollar Poincaré rap
- Math professor



MENSIS OCTOBRIS A. M DC LXXXIV. 467

NOVA METHODUS PRO MAXIMIS ET MINIMIS, itemque tangentibus, quæ nec fractas, nec irrationales quantitates moratur, & singulare pro illis calculi genus, per G. G. L.

Sit axis AX, & curvæ plures, ut V V, W W, Y Y, Z Z, quarum ordinatæ, ad axem normales, V X, W X, Y X, Z X, quæ vocentur respective, v, w, y, z; & ipsa A X abscissa ab axe, vocetur x. Tangentes sint V B, W C, Y D, Z E axi occurrentes respective in punctis B, C, D, E. Jam recta aliqua pro arbitrio assumta vocetur dx, & recta quæ sit ad dx, ut v (vel w, vel y, vel z) est ad V B (vel W C, vel Y D, vel Z E) vocetur d v (vel d w, vel dy vel dz) sive differentia ipsarum v (vel ipsarum w, aut y, aut z) His positis calculi regulæ erunt tales: TAB.

Sit a quantitas data constans, erit da æqualis o, & d $\overline{ax}$ erit æqu. a dx; si sit y æqu v (seu ordinata quævis curvæ Y Y, æqualis cuivis ordinatæ respondenti curvæ V V) erit dy æqu. dv. Jam *Additio & Subtractio*; si sit z — y + w + x æqu. v, erit d $\overline{z - y + w + x}$ seu d v, æqu dz — dy + dw + dx. *Multiplicatio*, d x v æqu. x d v + v dx, seu posito y æqu. x v, fiet d y æqu x d v + v dx. In arbitrio enim est vel formulam, ut x v, vel compendio pro ea literam, ut y, adhibere. Notandum & x & dx eodem modo in hoc calculo tractari, ut y & dy, vel aliam literam indeterminatam cum sua differentiali. Notandum etiam non dari semper regressum a differentiali Æquatione, nisi cum quadam cautione, de quo alibi. Porro *Divisio*, d $\frac{v}{y}$ vel (posito z æqu. $\frac{v}{y}$) dz æqu. $\frac{+v\,dy - y\,dv}{yy}$

Quoad *Signa* hoc probe notandum, cum in calculo pro litera substituitur simpliciter ejus differentialis, servari quidem eadem signa, & pro + z scribi + dz, pro — z scribi — dz, ut ex additione & subtractione paulo ante posita apparet; sed quando ad exegesin valorum venitur, seu cum consideratur ipsius z relatio ad x, tunc apparere, an valor ipsius dz sit quantitas affirmativa, an nihilo minor seu negativa: quod posterius cum fit, tunc tangens Z E ducitur a puncto Z non versus A, sed in partes contrarias seu infra X, id est tunc cum ipsæ ordinatæ

FIGURE 73.1. Leibniz's first paper on calculus, Acta Eruditorum, 1684, with the above rules for differentation.

74
Rules of Integration

74.1 Linearity

Basic linearity properties of integrals $u(t) = \int_0^t f(s)\,ds$ follow directly from linearity of the underlying IVP $\dot{u} = f$, that is

$$\int_0^t (f(s) + g(s))\,ds = \int_0^t tf(s)\,ds + \int_0^t g(s)\,ds, \int_0^t \alpha f(s)\,ds = \alpha \int_0^t f(s)\,ds, \tag{74.1}$$

where $\alpha \in \mathbb{R}$ is a constant. Further, for $a < b < c$,

$$\int_a^b f(s)\,ds + \int_b^c f(s)\,ds = \int_a^c f(s)\,ds, \tag{74.2}$$

which is extended to arbitrary limits a, b and c, by defining for $b < a$

$$\int_a^b f(s)\,ds = -\int_b^a f(s)\,ds. \tag{74.3}$$

Alternatively, these rules are derived directly from the Riemann-sum representation of the integral.

74.2 Integration by Parts

By the Fundamental Theorem of Calculus, we have

$$u(t)v(t) - u(0)v(0) = \int_0^t \frac{d}{ds}(u(s)v(s))ds = \int_0^t (u\dot{v} + u\dot{v})ds. \tag{74.4}$$

FIGURE 74.1. On June 30 2007 The Swedish Parliament dismantled Integrationsverket, the Ministry for Integration, and replaced it by the Ministry for Time-Stepping.

which can be written

$$\int_0^t u\dot{v}\,ds = [uv]_0^s - \int_0^t u\dot{v}\,ds, \tag{74.5}$$

with $[uv]_0^s = u(t)v(t) - u(0)v(0)$. We see that we "can move the dot" from u to v if we change sign and take into account the difference of end-point values of uv.

74.3 Change of Integration Variable

If $w : [a, b] \to \mathbb{R}$ is differentiable and $v : [w(a), w(b)] \to \mathbb{R}$ is Lipschitz continuous, then

$$\int_{w(a)}^{w(b)} v(y)\,dy = \int_a^b v(w(x))w'(x)dx, \tag{74.6}$$

because with $y = w(x)$, we have $dy \equiv dw = w'dx$.

75
Proof of the Fundamental Theorem

> The quadrature of all figures follow from the inverse method of tangents, and thus the whole science of sums and quadratures can be reduced to analysis, a thing nobody even had any hopes of before. (Leibniz)
>
> Knowing thus the Algorithm of this calculus, which I call Differential Calculus, all differential equations can be solved by a common method. (Leibniz)

Let us now study the effect of the time step in solution of

$$\dot{u}(t) = f(t), \quad \text{for } t > 0, \quad u(0) = u^0, \tag{75.1}$$

by Forward Euler time stepping

$$u(ndt + dt) = u(ndt) + f(ndt)dt, \quad n = 0, 1, 2, ... \tag{75.2}$$

We compare taking one step with time step dt with two steps of time step $\frac{dt}{2}$, for a given n:

$$\begin{aligned} u(ndt + dt) - \bar{u}(ndt + dt) &= f(ndt)dt - (f(ndt) + f(ndt + \frac{dt}{2}))\frac{dt}{2} \\ &= (f(ndt) - f(ndt + \frac{dt}{2}))\frac{dt}{2}, \end{aligned} \tag{75.3}$$

where $\bar{u}$ is computed with time step $\frac{dt}{2}$, and we assume that the same intial value for $t = ndt$ is used so that $\bar{u}(ndt) = u(ndt)$. Assuming that $f(t)$ is

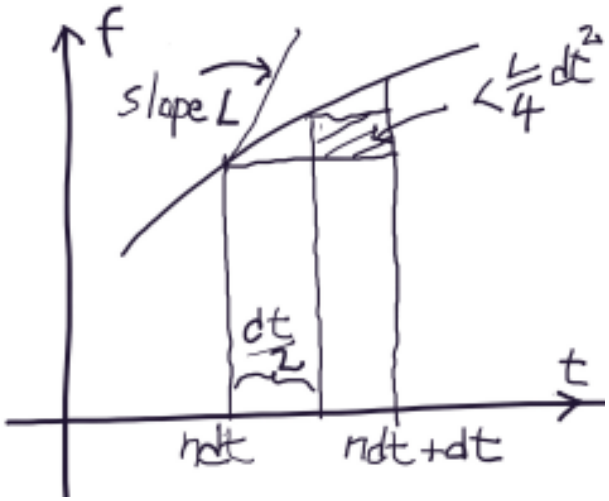


FIGURE 75.1. The fundamental step in the proof of the Fundamental Theorem.

Lipschitz continuous with Lipschitz constant L, we then find that

$$|u(ndt+dt)-\bar{u}(ndt+dt)| \leq \frac{L}{4}dt^2. \tag{75.4}$$

Summing now the contributions from all time steps with $n = 0, 1, 2, ..., N$, where $T = (N+1)dt$ is a final time, we get using that $\sum_{n=0}^{N} dt = T$,

$$|u(T)-\bar{u}(T)| \leq \frac{LT}{4}\,dt, \tag{75.5}$$

where thus $u(T)$ is computed with time step dt and $\bar{u}(T)$ with time step $\frac{dt}{2}$. Repeating the argument with successively refined times step $\frac{dt}{4}, \frac{dt}{8}, ...$, we get

$$|u(T)-\bar{u}(T)| \leq \frac{LT}{2}\,dt \tag{75.6}$$

for the difference between $u(T)$ computed with time step dt and $\bar{u}(T)$ computes with vanishingly small time step, since

$$\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+... < \frac{1}{2}. \tag{75.7}$$

We have now proved the Fundamental Theorem of Calculus:

Theorem 75.1 *If $f : [0, T] \to \mathbb{R}$ is Lipschitz continuous, then the function $u(t) = \int_0^t f(s)$ defined by Forward Euler time-stepping with vanishing time step, solves the IVP: $\dot{u}(t) = f(t)$ for $t \in (0, 1)$, $u(0) = 0$.*

The proof shows what it means *to understand* the Fundamental Theorem of Calculus, which means to realize that (letting k denote a finite time step and dt a vanishingly small step)

$$u(T) = \int_0^T f(t)\,dt \approx \sum_{n=0}^{N} f(nk)k \quad \text{if } T = (N+1)k, \tag{75.8}$$

as a consequence of

$$u((n+1)k) \approx u(nk) + f(nk)k, \quad \text{or} \quad \frac{u((nk+k) - u(nk)}{k} \approx f(nk), \tag{75.9}$$

where the sum is referred to as a *Riemann sum*, with the following bound for the difference

$$|\int_0^T f(t)\,dt - \sum_{n=0}^{N} f(nk)k| \leq \frac{LTk}{2} \quad \text{if } T = (N+1)k, \tag{75.10}$$

if $f : [0,T] \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L.

In other words, *understanding* the integral $u(t) = \int_0^t f(s)\,ds$ of a function $f : [0,T] \to \mathbb{R}$ means to understand that it is determined by Riemann sums with vanishingly small step size, as the solution to the IVP $\dot{u}(t) = f(t)$, $u(0) = 0$, and to understand that the difference between two Riemann sums with mesh size k and $\frac{k}{2}$, is bounded by Lk (or more precisely by $\frac{L}{4}k$).

75.1 Even Better Understanding

As a serious student, you now probably ask: In precisely what sense the differential equation $\dot{u}(t) = f(t)$ is satisfied by an Euler Forward solution $u(t)$ with time step k? It certainly is so constructed, but can we get a direct verification? One way to do this is to associate a continuous piecewise linear function determined by the values $u(nk)$ at the discrete time levels nk, again denoted by $u(t)$. We then have on each interval $(nk,(n+1)k)$, by the definition of $u(t)$:

$$\dot{u}(t) = \frac{u((n+1)k) - u(nk)}{k} = f(nk), \tag{75.11}$$

from which we conclude that

$$|\dot{u}(t) - f(t)| \leq |f(nk) - f(t)| \leq Lk \quad \text{for } t \in ((n+1)k, nk). \tag{75.12}$$

We can thus say that $u(t)$ satisfies the differential equation $\dot{u}(t) = f(t)$ for all t with a precision of Lk. In other words, the *residual* $\dot{u}(t) - f(t)$ is smaller than Lk. We have now understood the Fundamental Theorem even better, right?

FIGURE 75.2. The sad result of Archimedes mathematics.

We shall see below that extending a function defined on a discrete set of points to a continuous piecewise linear function, is a central aspect of approximation in general and of the finite element method in particular.

75.2 To Think About

- What is fundamental about the Fundamental Theorem?
- Why is $\frac{d}{dt}\int_0^t f(s)\,ds = f(t)$? (compare with last argument)
- What is the Riemann sum error using the Trapezoidal Rule (62.12)?

Hint: $\int_0^{t+dt} f(s)\,ds - \int_0^t f(s)\,ds = \int_t^{t+dt} f(s)\,ds = f(t)dt \pm \frac{L}{2}dt^2$.

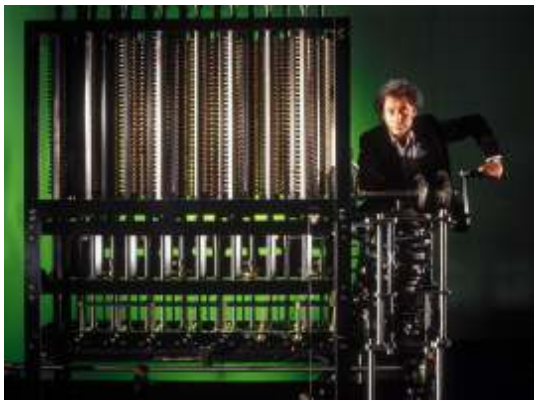


FIGURE 75.3. Babbage's Difference Engine No. 2 1847.

76
Contraction Mapping for $u = g(u)$

Give me a fixed point, and I will move the Earth. (Archimedes)

76.1 Solving $f(u) = 0$ by Time Stepping

To solve an equation $f(u) = (f_1(u), f_2(u), ..., f_N(u)) = 0$ of N equations $f_i(u) = 0$, $i = 1, ..., N$, in N unknowns $u = (u_1, u_2, ..., u_N)$, with thus $f : \mathbb{R}^n \to \mathbb{R}^n$, it is natural to connect to solution of the IVP: Find $u(t)$ such that

$$\dot{u}(t) + f(u(t)) = 0 \quad \text{for } t > 0, \quad u(0) = u^0, \tag{76.1}$$

with some given initial value u^0. If it turns out that as t increases, the function $u(t)$ tends to some value $\hat{u}$, then $\dot{u}(t)$ could be expected to become small, and if so, we would have

$$f(u(t)) \approx 0. \tag{76.2}$$

and we would be led to set $\hat{u} = u(t)$ for some large t and consider $\hat{u}$ to be an approximate solution of $f(u) = 0$ with small residual $f(\hat{u})$.

If $f(u)$ has several different solutions, which is often the case, then we could expect to capture different solutions by chosing different initial values u^0.

Computing $u(t)$ by Forward Euler with time step $dt = 1$, we would have

$$u^{n+1} = u^n - f(u^n), \quad \textit{for } n = 0, 1, 2, ..., \tag{76.3}$$

If $|u^{n+1} - u^n|$ would become small for increasing n, then $f(u^n)$ would become small and thus u^n would be an approximate solution of $f(u) = 0$ with small residual $f(u^n)$.

76.2 Solving $u = g(u)$

We are thus led to study the convergence of the iteration

$$u^{n+1} = g(u^n), \quad n = 0, 1, 2, ..., \tag{76.4}$$

where

$$g(u) = u - f(u). \tag{76.5}$$

To this end we take the difference of (76.4) for two consecutive steps to get

$$e^{n+1} \equiv u^{n+1} - u^n = g(u^n) - g(u^{n-1}). \tag{76.6}$$

If $g : \mathbb{R}^N \to \mathbb{R}^N$ is Lipschitz continuous with Lipschitz constant L, then

$$|e^{n+1}| = |g(u^n) - g(u^{n-1})| \leq L|e^n| \leq L^2|e^{n-1}| \leq L^n|u^1 - u^0| \tag{76.7}$$

We see that if $L < 1$, then $|e^n|$ becomes vanishingly small as n increases, which by (76.3) means that $f(u^n)$ becomes vanishingly small and thus u^n may be viewed as an approximate solution of $f(u)$ in the sense that the residual $f(u^n)$ is small. In the next chapter we also consider the error in the approximate root u^n.

We see that if $L << 1$ then the convergence is fast, and if $L \approx 1$ then the convergence is slow. If $L = \frac{1}{2}$ then the residual $|g(u^n) - u^n| = |u^{n+1} - u^n|$ is reduced with a factor 2 in each iteration step, that is with a binary digit per step.

If $L < 1$ then the mapping $u \to g(u)$ is said to be a *contraction*, because

$$|g(u) - g(v)| \leq L|u - v| < |u - v| \tag{76.8}$$

expressing that the distance between the images $|g(u) - g(v)|$ is smaller than the distance between the arguments $|u - v|$. We have just proved the famous

Contraction Mapping Theorem: If $g : \mathbb{R}^N \to \mathbb{R}^N$ is a contraction with Lipschitz constant $L < 1$, then the iteration $u^{n+1} = g(u^n)$ converges to a unique fixed point satisfying $u = g(u)$ at the rate L^n.

77
Newton's Method for $f(u) = 0$

> The sciences, are small power; because not eminent; and therefore, not acknowledged in any man; nor are at all, but in a few; and in them, but of few things. For science is of that nature, as none can understand it to be, but such as in a good measure have attained it. (Thomas Hobbes in Leviathan Chapter X 14.)
>
> Arts of public use, as fortifications, making of engines, and other instruments of war; because they confer to defence, and victory, are power: and though the true mother of them, be science, namely the mathematics; yet, because they are brought into the light, by hand of the artificer, they be esteemed (the midwife passing with vulgar for the mother,) as his issue. (Thomas Hobbes in Leviathan Chapter X 15.)

We now consider a variant of (76.3) for solving $f(u) = 0$ with faster convergence by invoking the (inverse of the) derivative $f'(u)$, referred to as *Newton's Method*.

Let us then start with $N = 1$ and let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function. Consider the following iteration:

$$u^{n+1} = u^n - \frac{f(u^n)}{f'(u^n)} = g(u^n) \tag{77.1}$$

with corresponding function

$$g(u) = u - \frac{f(u)}{f'(u)} \tag{77.2}$$

assuming that $f'(u) \neq 0$. Computing the derivative $g'(u)$, we get

$$g'(u) = 1 - \frac{f'(u)}{f'(u)} + \frac{f(u)f''(u)}{(f'(u))^2} = 0, \tag{77.3}$$

if $f(u) = 0$. Thus we may expect that $|g'(u)|$ is small, that is that $L << 1$ implying fast convcergence.

The iteration (258.1) is called *Newton's Method* for computing a solution of the equation $f(u) = 0$. Newton's method directly generalizes to $f : \mathbb{R}^N \to \mathbb{R}^N$ in the form

$$u^{n+1} = u^n - (f'(u^n))^{-1} f(u^n) \tag{77.4}$$

where $f'(u^n))^{-1}$ is the inverse of the $N \times N$ matrix $f'(u^n)$ (thus assuming that $f'(u^n)$ is non-singular). One can show that $|e^{n+1}| \sim |e^n|^2$, if the initial guess is close enough to the root, which means that the number of correct digits may double at iteration step.

77.1 Wellposed and Illposed Roots

Suppose u is an approximate solution with residual $f(u) \approx 0$, or approximate *root*, of an equation $f(u) = 0$ with exact root $\bar{u}$. We have for small $|u - \bar{u}|$

$$f(u) - f(\bar{u}) \approx f'(u)(u - \bar{u}), \tag{77.5}$$

(still assuming for simplicity $N = 1$. This shows that

$$|u - \bar{u}| \approx \frac{|f(u)|}{|f'(u)|} \tag{77.6}$$

indicating that the residual error $|f(u)|$ translates to the root error $|u - \bar{u}|$ with the *stability factor*

$$S = \frac{1}{|f'(u)|}, \tag{77.7}$$

that is

$$|u - \bar{u}| \approx S|f(u)| \tag{77.8}$$

In other words: If $|f'(u)|$ is not small so that S is not large, then the root is well defined or *wellposed*, while if $|f'(u)|$ is small so that S is large, then the root is *illposed* or not well defined.

For a wellposed root u the curve $x \to f(x)$ crosses the x-axis at $x = u$ with a definite slope, which makes the crossing point well determined. For an illposed root the curve is almost tangent to the x-axis which makes the crossing point difficult to pin down.

77.2 Newton's Method Requires Good Initial Guess

Newton's method converges very quickly towards a root, if the starting value is close enough to the root. If not, the iterations may diverge and then give rise complex fractal patterns as shown in the figure below showing big basins of convergence around roots separated by fractal boundary zones.

77.3 Learn More

- Fixed point iteration.
- Newton's method

77.4 To Think About

- How to compute $\sqrt{2}$? By Solving $x^2 = 2$? How?

77.5 Watch

- Newton's method fractal 1
- Newton's method fractal 2
- Newton fractals algorithm

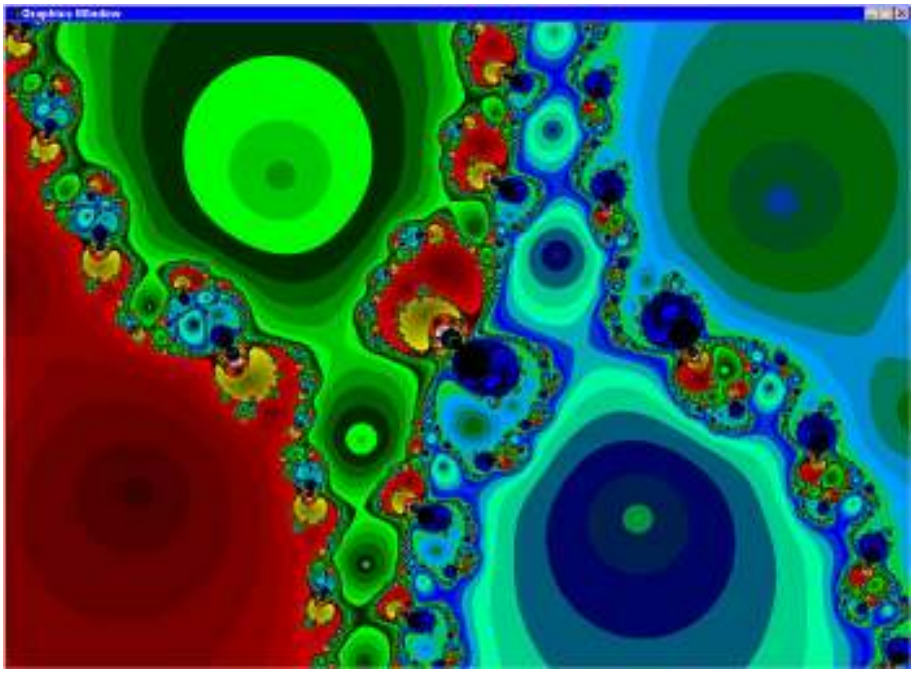


FIGURE 77.1. Fractals from iterations by Newton's method. Big basins show roots. Boundaries between basins show fractal complexity.

78
Generalized Fundamental Theorem

I believe in the fundamental Truth of all the great religions of the world. I believe that they are all God given. I came to the conclusion long ago... that all religions were true and also that all had some error in them. (Mahatma Gandhi)

The fairest thing we can experience is the mysterious. It is the fundamental emotion which stands at the cradle of true art and true science. He who know it not and can no longer wonder, no longer feel amazement, is as good as dead, a snuffed-out can (Einstein)

Most of the fundamental ideas of science are essentially simple, and may, as a rule, be expressed in a language comprehensible to everyone. (Einstein)

78.1 Time Stepping $\dot{u} = u$

The Fundamental Theorem concerns time-stepping of the IVP

$$\dot{u}(t) = f(t) \quad \text{for } t > 0, \quad u(0) = u^0, \tag{78.1}$$

where the Lipschitz continuous function $f(t)$ does not depend on the unknown u, only on the (independent) time variable t.

We now extend to allow f to depend also on u. Assuming for simplicity no explicit dependence on t, we thus consider the IVP:

$$\dot{u}(t) = f(u(t)) \quad \text{for } t > 0, \quad u(0) = u^0, \tag{78.2}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a given Lipschitz continuous function with Lipschitz constant L. The basic question is if the solution can be computed to arbitrary precision by time-stepping?

The basic case is $f(u) = u$ and $u^0 = 1$, that is the IVP:

$$\dot{u}(t) = u(t) \quad \text{for } t > 0, \quad u(0) = 1, \tag{78.3}$$

with $L = 1$ and the solution $u(t) = \exp(t)$ computed by Forward Euler:

$$\exp(t) \approx (1 + \frac{t}{n})^n \tag{78.4}$$

with time step $k = \frac{t}{n}$. We estimate the effect of dividing the time-step by a factor 2, using that $(t + dt)^n - t^n \approx nt^{n-1}dt$ (because $\frac{d}{dt}t^n = nt^{n-1}$):

$$\begin{aligned}(1 + \frac{t}{2n})^{2n} - (1 + \frac{t}{n})^n &= ((1 + \frac{t}{2n})(1 + \frac{t}{2n}))^n - (1 + \frac{t}{n})^n \\ &= ((1 + \frac{t}{n} + \frac{t^2}{4n^2})^n - (1 + \frac{t}{n})^n \approx n(1 + \frac{t}{n})^{n-1}\frac{t^2}{4n^2} \approx \frac{t}{n}\exp(t)\frac{t}{4}.\end{aligned}$$

We see that the difference is proportional to the time step $k = \frac{t}{n}$. As in the proof of the Fundamental Theorem of Calculus, we conclude that $(1 + \frac{t}{n})^n$ determines $\exp(t)$ with a precision proportional to the time step with a multiplicative factor $\approx \exp(t)\frac{t}{4} \sim \exp(t)$.

78.2 Time Stepping $\dot{u} = f(u)$

The above proof extends to an arbitrary Lipschitz continuous function $f(u)$ with the difference that the time-stepping error in computing $u(t)$, now is proportional to the time step with a multiplicative factor $\exp(Lt)$, where L is the Lipschitz constant of f. This extends to systems with $f : \mathbb{R}^d \to \mathbb{R}^d$ with $d > 1$.

It is natural to refer to this result as a *Generalized Fundamental Theorem*: The Fundamental Theorem concerns $\dot{u}(t) = f(t)$ and the Generalized Fundamental Theorem concerns $\dot{u}(t) = f(u(t))$. Calculus in a nutshell!

A proof of the Generalized Fundamental Theorem can be performed by combining the following two steps:

Step 1: Estimate the difference $u((n + 1)k) - \tilde{u}((n + 1)k)$ by taking one (Forward Euler) time step of length k and two time steps on length $\frac{k}{2}$, from

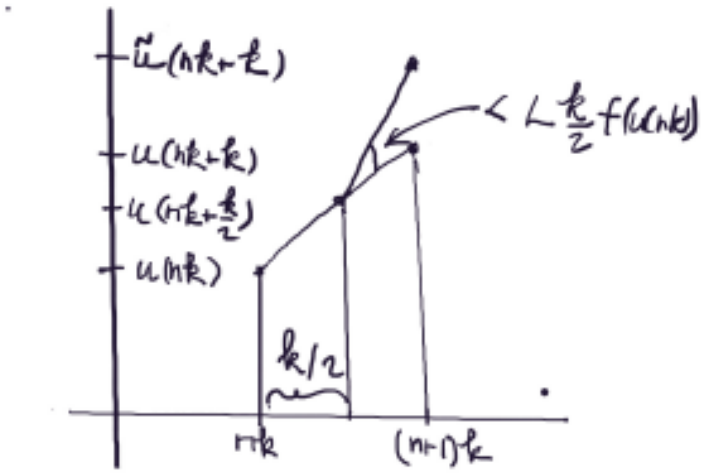


FIGURE 78.1. Fundamental Step 1 in the proof of the Generalized Fundamental Theorem.

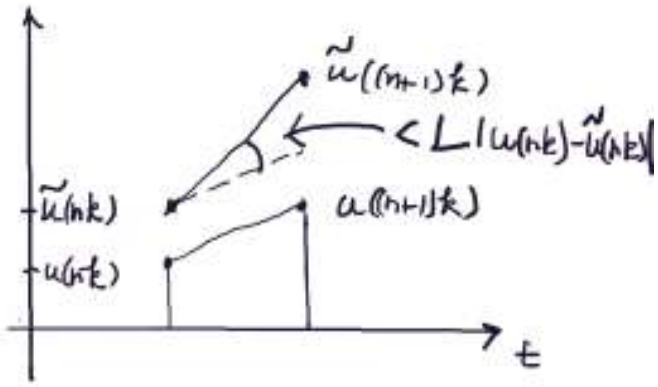


FIGURE 78.2. Fundamental Step 2 in the proof of the Generalized Fundamental Theorem.

the same initial value $u(nk)$:

$$\begin{aligned}
&|u((n+1)k) - \tilde{u}((n+1)k)| \\
&= |u(nk) + kf(u(nk)) - (u(nk) + \frac{k}{2}f(u(nk)) + \frac{k}{2}f(u(nk) + \frac{k}{2}f(u(nk)))| \\
&= \frac{k}{2}|f(u(nk) - f(u(nk) + \frac{k}{2}f(u(nk)))| \leq \frac{k}{2}L\frac{k}{2}|f(u(nk))|.
\end{aligned} \tag{78.5}$$

Step 2: Estimate the difference after one time step from different initial conditions $u(nk) - \tilde{u}(nk)$:

$$\begin{aligned}
|u((n+1)k) - \tilde{u}((n+1)k) &= |u(nk) + kf(u(nk)) - \tilde{u}(nk) + kf(\tilde{u}(nk))| \\
&\leq (1 + kL)|u(nk) - \tilde{u}(nk)|.
\end{aligned} \tag{78.6}$$

Combining Steps 1 and 2, we obtain a final error proportional to the time step k with a multiplicative factor $\exp(Lt)$, which we refer to as a *stability factor*. For details see Completion of the Proof.

To see the connection with the basic case $f(u) = u$, think of estimating a general function $f(u)$, assuming $f(0) = 0$ for simlipicity, by $f(u) \approx f'(0)u$, which suggests that the factor $\exp(t)$ for $f(u) = u$ should be replaced by $\exp(Lt)$ for a general $f(u)$ (because $|f'(0)| \leq L$).

Generalization to a vector valued function $f : \mathbb{R}^d \to \mathbb{R}^d$ with $d > 1$ is direct, and we have thus presented the essential steps of a proof of the following main result of Calculus:

Theorem 78.1 Generalized Fundamental Theorem of Calculus: *The solution $u(t)$ of the IVP $\dot{u}(t) = f(u(t))$ for $0 < t \leq T$ with $u(0)$ given, where $f : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous with Lipschitz constant L, is uniquely computable by Forward Euler time stepping with a precision proportional to the time step times a stability factor of size* $\exp(LT)$.

For a completion of the proof, see below and Special Case and General Case.

The proof is similar for the other methods we have so far encountered (with $u^n = u(nk)$):

$$\begin{aligned}
u^{n+1} &= u^n + kf(u^{n+1}) \quad \text{Backward Euler,} \\
u^{n+1} &= u^n + kf(\frac{u^n + u^{n+1}}{2}) \quad \text{Midpoint Euler,} \\
u^{n+1} &= u^n + \frac{k}{2}(f(u^n) + f(u^{n+1})) \quad \text{Trapezoidal Method.}
\end{aligned} \tag{78.7}$$

We see that if $f(u)$ is linear in u, then Midpoint Euler and the Trapezoidal Method coincide.

78.3 A Posteriori Error Control

For a more precise error control, based on computed solutions, see

- Time Stepping Error Analysis
- Time Stepping by FEM

78.4 The Illusion of an $\exp(LT)$ Bound

If $L = 10$ and $T = 30$, which looks pretty harmeless, then $\exp(LT) = \exp(300) >> 10^{100} =$ **googol**, an incredibly large number, much larger than the number of atoms in the Universe. A matching time step of 10^{-100} is beyond all rationale and thus computation of a solution of an IVP with moderate Lipschitz constant oevr a moderatley long time interval may be impossible. An example is the Lorenz system with

$$f(u) = (-10u_1 + 10u_2, 28u_1 - u_2 - u_1u_3, -\frac{8}{3}u_3 + u_1u_2), \tag{78.8}$$

for which computation on an interval of length T requires computation with about $T/2$ digits, see:

- BS The Lorenz System and the Essence of Chaos
- Long-Time Computability of the Lorenz System

78.5 Stiff IVPs

There is a class of IVPs with large or very large Lipschitz constants, which are computable on long time intervals, because the function $f(u)$ has a decay property (negative derivative) causing errors to decay rather than grow exponentially. Such problems are called *stiff problems* and may require implicit time stepping to avoid severe time step restrcitions in explicit methods. See Stiff Problems.

78.6 Wave Equations

IVPs with wavelike solutions, like the system

$$\dot{u}_1 = u_2 \quad \dot{u}_2 = -u_1 \tag{78.9}$$

with solutions being linear combinations of $\sin(t)$ and $\cos(t)$, have formally Lipschitz constants of size 1, can be integrated with error growth $\sim t$, instead of $\sim \exp(t)$ by the above (crude) estimate, if a proper time-stepping

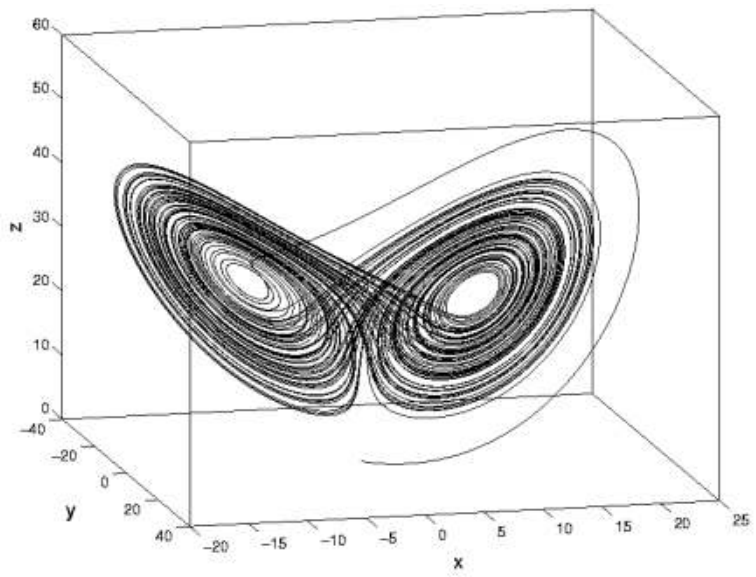


FIGURE 78.3. A Lorenz system solution trajectory.

method (like cG(1)) is used. This is due to error cancellation in wave motion.

For more complex wave problems, or problems with more or less periodic solutions, the stability factor can have a polynomial growth in time t, e.g. quadratic for simple planetary systems.

78.7 Summary: Time Stepping of IVP

The precision in time stepping the solution $u(t)$ of an IVP $\dot{u} = f(u)$ for $0 < t \leq T$, with first order method with time step k, can be estimated by $S(T)k$, where $S(t)$ acts as stability factor measuring error propagation and accumulation of size

- $S(T) \sim \exp(LT)$ (general), where L is the Lipschitz constant of f.
- $S(T) \sim 1$ (stiff: diffusion problems)
- $S(T) \sim T$ (wave problems), $\quad S(T) \sim T^2$ (planetary system).

78.8 Preparing for a More Precise Analysis

As a preparation for the more precise error analysis in Time Stepping Error Analysis and Time Stepping by FEM, we consider two solutions $u(t)$ and $\tilde{u}(t)$ com-

puted with the same time step k but different initial data u^0 and $\tilde{u}^0$. Subtracting the update formulas we have formally for the difference $e = u - \tilde{u}$:

$$\dot{e}(t) \approx f'(u(t))e(t) \quad \text{for } t > 0, \quad e(0) = e^0 \equiv u^0 - \tilde{u}^0 \tag{78.10}$$

showng that an initial error is propagated as a solution to a linearized IVP with coefficient $f'(u(t))$ depending on a computed solution $u(t)$. We shall see that by solving the linearized problem (or rather a closely related dual linearized problem), the stability factors $S(t)$ measuring error growth can be computed and the precision on the computation of $u(t)$ can be assessed.

The linearized problem (or its dual) thus gives the key to unlock time stepping precision.

Note that with $f(u) = u$, the linearized problem reads $\dot{e} = e$ with solution $e(t) = \exp(t)e^0$, showing exponential error growth, as expected.

78.9 Completion of the Proof

To complete the proof of the Generalized Fundmental Theorem we are to sum up the error contributions from each subinterval of length k, which according to Step 1 and Step 2 amounts to

$$\begin{aligned} \sum_{n=1}^{N}(1+kL)^n Lk^2M &\approx kML\sum_{n=0}^{N}\exp(Lnk)k \\ &\approx kML\int_0^T \exp(Ls)\,ds \approx kM\exp(LT), \end{aligned} \tag{78.11}$$

where $Nk = T$ and $M \geq \max_u |f(u)|$. Can you explain what is going on here? If not, take a look at:

78.10 Hint to Completion of the Proof

We can think of comparing computations with k and $\frac{k}{2}$ with corresponding solutions $u(t)$ and $\tilde{u}(t)$ in two ways depending on how we choose initial values on each time interval $(nk, (n+1)k)$:

1. Compute $u(t)$ and $\tilde{u}(t)$ independently with timestep k and $\frac{k}{2}$.
2. Assume that $\tilde{u}(nk) = u(nk)$ and account for the effect at final time of the difference $\tilde{u}(nk) - u(nk)$.

1. is the most direct from computational point of view and a corresponding proof is given in General Case.

Here we consider 2. because the proof is (maybe) simpler: The error from the first time step is bounded by Lk^2M assuming no error in initial data,

and is propagated with a factor bounded by $(1 + kL)$ for each time step, thus with a factor $(1 + kL)^N$ after N steps. Similarly, the error from the second time step is bounded by Lk^2M, again assuming no error in the corresponding initial value, and is propagated with a factor $(1 + kL)^{N-1}$, et cet. Summing we obtain a bound of the total error after N steps.

It is instructive to in illustrate 1. and 2. in a figure complementing Figs. 77.1-2.

78.11 Uniqueness of Solution

To prove that the solution of the IVP (**??**) is unique, we assume $v(t)$ is a possibly different solution also satisfying $\dot{v}(t) = f(v(t))$ for $t > 0$ and $v(0) = u^0$. Subtraction gives for the difference $w = u - v$

$$\dot{w} = f(u) - f(v) \tag{78.12}$$

and thus taking the scalar product with w and using Cauchy's inequality, we get

$$\frac{d}{dt}\frac{1}{2}|w|^2 = \frac{d}{dt}\frac{1}{2}w \cdot w = (f(u) - f(v)) \cdot w \leq L|w||w| \tag{78.13}$$

and thus for $W = |w|^2$

$$\dot{W} \leq 2LW, \tag{78.14}$$

which shows that (why?)

$$W(t) \leq W(0)\exp(2Lt) \quad \text{for } t > 0. \tag{78.15}$$

But $W(0) = |u(0) - v(0)| = 0$ and thus $W(t) = 0$ and $u(t) = v(t)$ for $t > 0$ and uniqueness follows. But to be scientifically honest, size of the exponential factor $\exp(Lt)$ is crucial. If $L10$ and $t = 30$, which does not look too frightening, then $\exp(Lt) = \exp(300) > 10^{100}$ =googol, which means that the argument the $\exp(300)W(0)$ is small (zer0) if $W(0)$ is small (zero), is questionable, very questionable, right?

78.12 How to Prove $\exp(t + s) = \exp(t)\exp(s)$?

To prove the basic law of the exponential $\exp(t + s) = \exp(t)\exp(s)$, note that the function $u(s) = \exp(t + s)$ satisfies $\frac{du}{ds} = u$ for $s > 0$ and $u(0) = \exp(t)$. But the function $v(s) = \exp(t)\exp(s)$ also satisfies $\frac{dv}{ds} = v$ for $s > 0$ and $v(0) = \exp(t)$, and by uniqueness $u(s) = v(s)$, that is $\exp(t + s) = \exp(t)\exp(s)$.

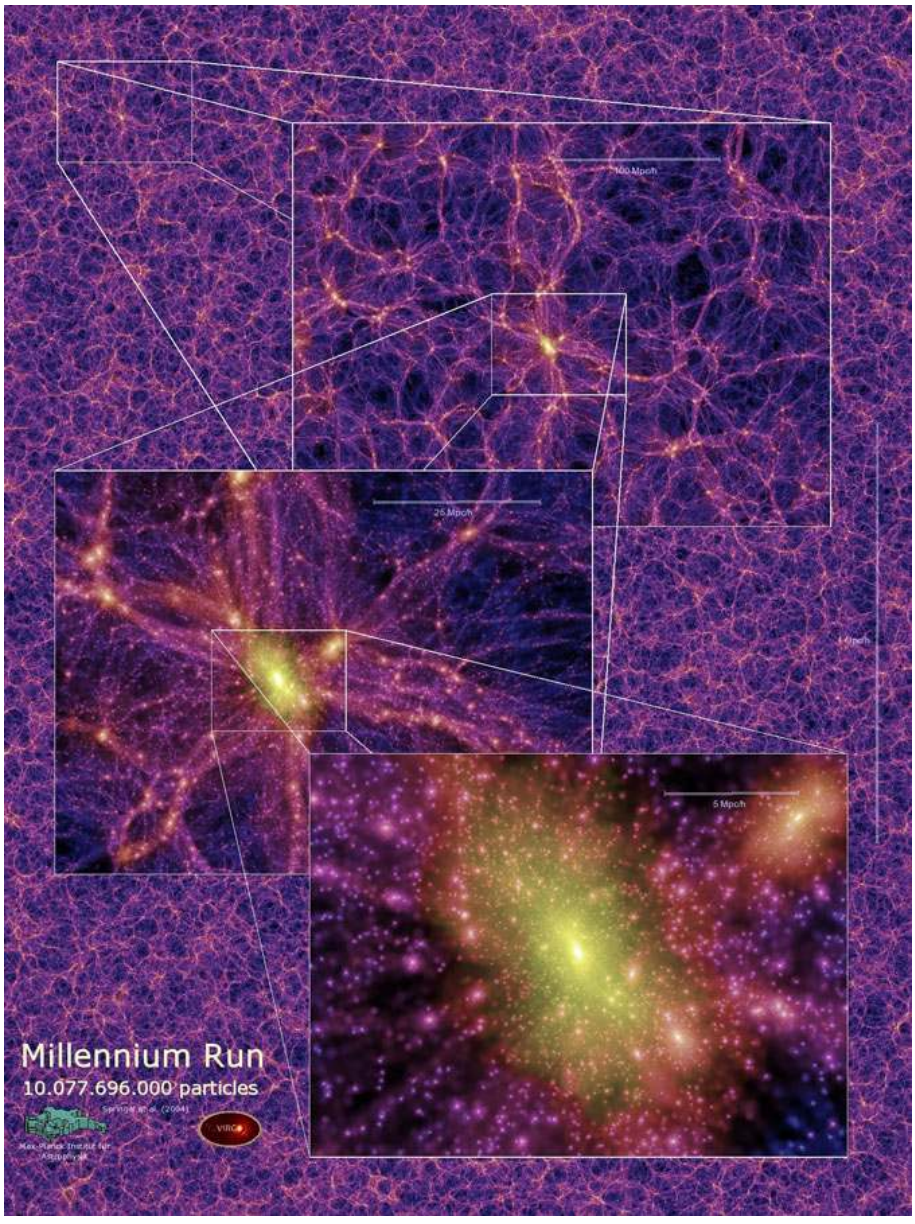


FIGURE 78.4. The Millennium Run: A large n-body computation with $n = 10.077.696.000$.

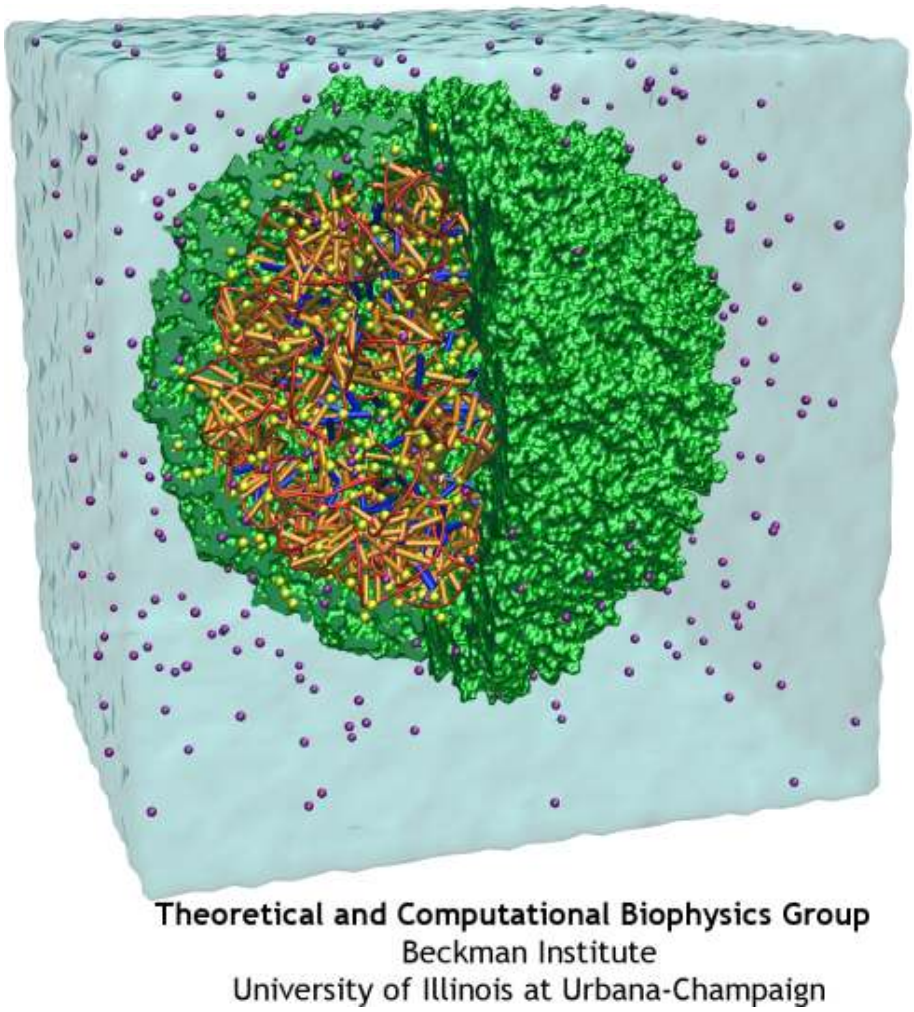


FIGURE 78.5. Model of a virus as a large molecular dynamics system of differential equations $\dot{u} = f(u)$.

79
Existence of World from $\dot{u} = f(u)$

> Dubito ergo cogito; cogito ergo sum. (I doubt, therefore I think; I think therefore I am). (Descartes)
>
> Man is nothing else but what he makes of himself. Such is the first principle of existentialism. (Jean-Paul Sarte)
>
> We should all be obliged to appear before a board every five years and justify our existence...on pain of liquidation. (George Bernard Shaw)
>
> The person lives most beautifully who does not reflect upon existence. (Friedrich Nietzsche)
>
> There is no place I know that compares to Pure Imagination. (Roal Dahl)

We can summarize all our studies as the study of the IVP

$$\dot{u}(t) = f(u(t)) \quad \text{for } t > 0, \quad u(0) = u^0, \tag{79.1}$$

where $f(u)$ a vector valued function depending on the vector function u and its derivatives, representing the *state* of a system.

The equation $\dot{u}(t) = f(u(t))$ connects the rate of change $\dot{u}(t)$ to the present state $u(t)$ through the function $f(u(t))$, from which the dynamics of the system can be computed by time-stepping. The model is the function $f(u)$.

FIGURE 79.1. A dot-model of the World.

The Generalized Fundamental Theorem shows that a World modeled by $\dot{u} = f(u)$ exists! Convinced? What is the weak point with this argument?

The model $\dot{u} = f(u)$ thus gives a very compact description of the World. Easy to remember!

79.1 Autonomous and Non-Autonomous IVPs

An IVP with a function $f(u(t), t)$ with an explicit dependence on t, referred to as a *non-autonomuous* IVP, can be rewritten on the *autonomuos* form (79.1) by introducing the new dependent variable $u_{d+1} = t$ to give $\hat{u} = (u_1, ..., u_d, u_{d+1})$ and adjoining the new equation $\dot{u}_{d+1} = f_{d+1} \equiv 1$ into an augumented $\dot{\hat{u}} = \hat{f}(\hat{u})$.

79.2 What Calculus is Most Useful?

The Egg of Calculus is the derivative and the integral is the Hen: First comes the derivative in the formulation of $\dot{u}(t) = f(t)$ and then comes the integral as the solution $u(t)$, by time-stepping in the same way as the Egg gradually develops into the Hen. Here $f(t)$ acts as the genetic code in

FIGURE 79.2. Summary of Calculus.

interplay with the environment from which the the solution the Hen as the solution $u(t)$.

Classical analytical Calculus of primitive functions concerns techniques for analytical solution of $\dot{u}(t) = f(t)$ as means to circumvent tedious laborious time-stepping: An analytical primitive function $u(t)$ can be seen as a shortcut replacing a tiresome step-by-step solution. In precomputer times such shortcuts were useful, and thus accordingly highly praised, but with the computer the original motivation has largely dissappeaed: Time-stepping solves $\dot{u}(t) = f(t)$ for any $f(t)$ at little computer cost, and thus in general is much more cost effective than tricky analytical shortcuts.

What is difficult in classical Calculus is analytical integration, since it consists of a bag of tricks with only limited power. Replacing analytical integration by time-stepping both simplifies Calculus and makes it more useful.

Human brains are good at formulating problems using principles, but cannot to massive computation and require a lot of training to handle bags of tricks.

The basic philosophy of BodyandSoul is to use the Soul/Brain to formulate equations like $\dot{u} = f(u)$ and then let the Body/Computer compute the solution by time-stepping. See also The Hen and the Egg of Gravitation.

80
Stability of Solutions to $\dot{u} = f(u)$

It is often in the name of cultural integrity as well as social stability and national security that democratic reforms based on human rights are resisted by authoritarian governments. (Aung San Suu Kyi)

It's well known I'm a Scientologist, and that has helped me to find that inner peace in my life and it's something that has given me great stability and tools that I use. (Tom Cruise)

A party of order or stability, and a party of progress or reform, are both necessary elements of a healthy state of political life. (John Stuart Mill)

80.1 Sensitivity to Perturbations

We return to the fundamental aspect of *stability* of solutions of an IVP:

$$\dot{u}(t) = f(u(t)) \quad \text{for } 0 < t \leq T,\ u(0) = u^0, \tag{80.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a given bounded Lipschitz continuous function, $u^0 \in \mathbb{R}^d$ is a given initial value and $[0, T]$ a given time interval.

We are thus interested in the *sensitivity* of a solution $u(t)$ with respect to *perturbations* of given data f, u^0 and T, that is the change of the solution $u(t)$ under small changes of data. If the solution changes a little under small

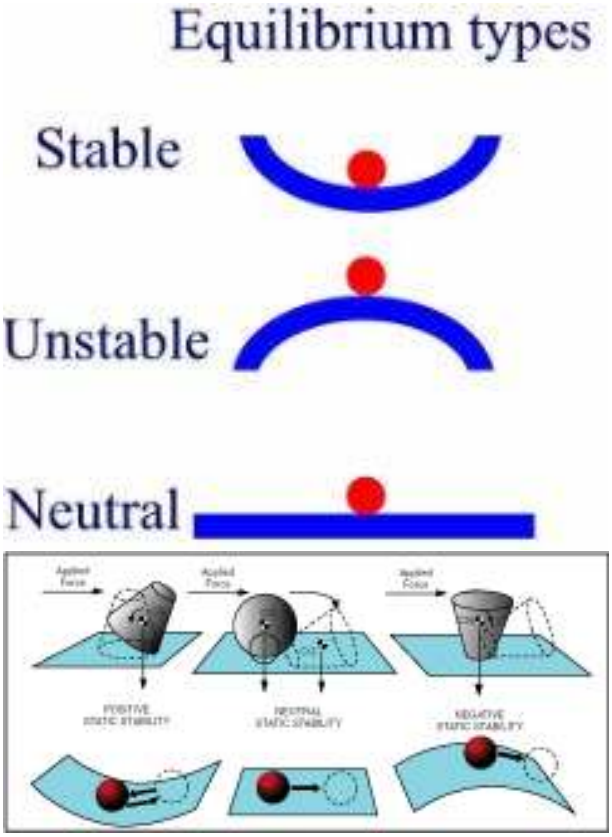


FIGURE 80.1. Stability in pictures.

changes of data, then the solution is *stable*, and if the solution changes a lot under small changes of data, the solution is *unstable*, qualitatively speaking. We can also say that the IPV is stable or unstable if this is a charateristic of all solutions. We focus here on perturbations of the initial data u^0.

We now turn to a more precise quantitaive analysis of stability, and we will then be led to the following *linearized problem*, assuming $f(u)$ is differentiable with respect to u with derivative $f'(u)$,

$$\dot{\varphi}(t) = f'(u(t))\varphi(t) \quad \text{for } 0 < t \leq T,\ \varphi(0) = \varphi^0, \tag{80.2}$$

where $u(t)$ is a given (computed) solution of (80.1) and the solution $\varphi(t)$ measures the effect on the solution $u(t)$ at time t of an initial perturbation φ^0 of initial data u^0. We see that the sign of the derivative $f'(u(t))$ determines if $\varphi(t)$ will be growing (if $f'(u(t)) > 0$) or decaying (if $f'(u(t)) < 0$), assuming that $\varphi^0 > 0$ with an analogous argument if $\varphi^0 < 0$.

80.2 Derivation of the Linearized Problem

We start considering a *stationary solution* $u(t) = \bar{u}$ for $0 \le t \le T$, where $\bar{u} \in \mathbb{R}$, that is a solution $u(t)$ of (80.1) that is independent of time t. Since $\dot{u}(t) = 0$ if $u(t)$ is independent of time, $\bar{u}$ is a solution of of the equation $f(\bar{u}) = 0$, studied in Newton's Method and Fixed Point Iteration.

We also refer to a stationary solution $u(t) = \bar{u}$ with $f(\bar{u}) = 0$ as an *equilibrium solution*.

We consider the initial value problem (80.1) with $u^0 = \bar{u} + \varphi^0$, where $\varphi^0 \in \mathbb{R}^d$ is a given small perturbation of the initial data $\bar{u}$. We denote the corresponding solution by $u(t)$ and focus attention on the corresponding perturbation in the solution, that is $\psi(t) = u(t) - \bar{u}(t) = u(t) - \bar{u}$. We want to derive a differential equation for the perturbation $\psi(t)$, and to this end we linearize f at $\bar{u}$ and write

$$f(u(t)) = f(\bar{u} + \psi(t)) = f(\bar{u}) + f'(\bar{u})\psi(t) + e(t),$$

where $f'(\bar{u})$ is the derivative of $f : \mathbb{R} \to \mathbb{R}$ at $\bar{u}$ and the error term $e(t)$ is quadratic in $\psi(t)$ (and thus is very small if $\psi(t)$ is small). Since $f(\bar{u}) = 0$ and $u(t)$ satisfies (219.1), we have

$$\dot{\psi}(t) = \frac{d}{dt}(\bar{u} + \psi(t))) = f(u(t)) = f'(\bar{u})\psi(t) + e(t).$$

Neglecting the quadratic term $e(t)$, we are led to a linear initial value problem,

$$\dot{\varphi}(t) = f'(\bar{u})\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0, \tag{80.3}$$

which is the linearized problem (80.2). Here $\varphi(t)$ is an approximation of the perturbation $\psi(t) = u(t) - \bar{u}$ up to a (small) second order term. The solution is given by

$$\varphi(t) = \exp(At)\varphi^0, \tag{80.4}$$

where $A = f'(\bar{u})$ is a constant.

We can immediately generalize to a time-dependent solution $u(t)$ with linearized problem

$$\dot{\varphi}(t) = f'(u(t))\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0, \tag{80.5}$$

with solution

$$\varphi(t) = \exp(A(t))\varphi^0, \tag{80.6}$$

where $A'(t) = f'(u(t))$ with $A(0) = 0$.

80.3 Stability Analysis

We consider the following cases

- $A < 0$: stable: perturbation decays: systems returns to initial state,
- $A > 0$: unstable: perturbation grows: system deviates from initial state,
- $A = i$: neutral: perturbation maintains size: system oscillates around initial state.

We can quantify stability in terms of the following stability factor with the subscript d referring to (initial) data:

$$\begin{aligned} S_d(f,T,\varphi^0) &= \max_{0\le t\le T} \frac{|\varphi(t)|}{|\varphi^0|}, \\ S_d \equiv S_d(f,T) &= \max_{\varphi^0\neq 0} S_d(f,T,\varphi^0). \end{aligned} \tag{80.7}$$

A large S_d will then indicate an unstable solution, and a moderate size or small S_d a stable or netural solution.

80.4 Dual Linearized Problem

$$S_d(u,T,\varphi^T) = \frac{|\psi(T)|}{|\varphi^T|} = \frac{|\varphi(0)|}{|\varphi^T|}, \quad S_d \equiv S_d(u,T) = \max_{\varphi^T\neq 0} S_d(u,T,\varphi^T). \tag{80.8}$$

80.5 Learn More

- Linearization and Stability of IVP
- The Crash Problem
- BMW stability control
- Unstable mathematics I
- Unstable mathematics II

> The existence of life must be considered as an elementary fact that can not be explained, but must be taken as a starting point in biology, in a similar way as the quantum of action, which appears as an irrational element from the point of view of classical mechanical physics, taken together with the existence of elementary particles, forms the foundation of atomic physics. The asserted impossibility of a physical or chemical explanation of the function peculiar to life would in this sense be analogous to the insufficiency of the mechanical analysis for the understanding of the stability of atoms. (Niels Bohr)

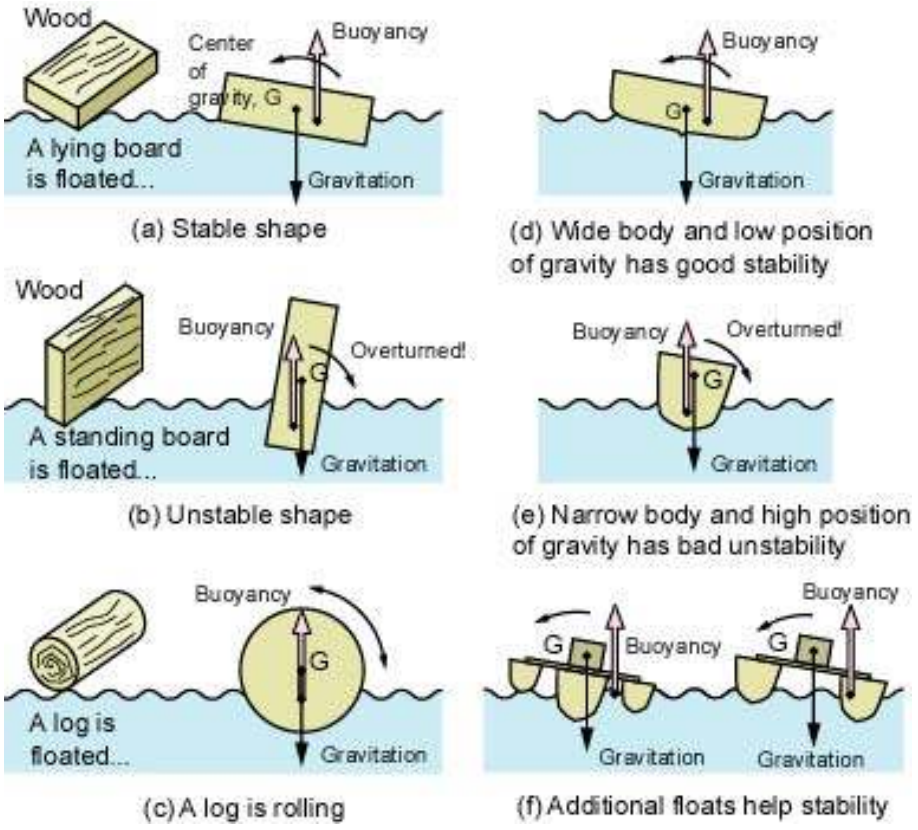


FIGURE 80.2. Stability of floating bodies, see Archimedes Prinicple

The stability of the atom is inexplicable. (Max Born)

81
What about Limits and Sequences?

If you always put limit on everything you do, physical or anything else. It will spread into your work and into your life. There are no limits. There are only plateaus, and you must not stay there, you must go beyond them. (Bruce Lee)

The mind is the limit. As long as the mind can envision the fact that you can do something, you can do it, as long as you really believe 100 percent. (Arnold Schwarzenegger)

81.1 Alternative Definitions of Continuity and Derivative

In Calculus books you usually find definitions of continuity and derivative based on the notion of *limit*: A function $x : \mathbb{R} \to \mathbb{R}$ is said to be continuous (at t) if

$$x(t) = \lim_{dt \to 0} x(t + dt), \tag{81.1}$$

and differentiable with derivative $\dot{x}(t)$ if

$$\dot{x}(t) = \lim_{dt \to 0} \frac{x(t + dt) - x(t)}{dt} \quad \text{where } dt \neq 0). \tag{81.2}$$

If you define continuity and derivative this way using limits, obviously there is a reason to confront the student with the (difficult) concept of limit.

FIGURE 81.1. Computational mathematics without limits.

The notion of limit relates to *converging sequences* with $\lim_{dt\to 0^+} x(t+dt) = x(t)$ expressing that for any decreasing sequence of positive time steps $dt_1 > dt_2 > .... > dt_n > dt_{n+1}, ...$ approaching 0, the difference $|x(t+dt_n) - x(t)|$ is smaller than any given positive number if only dt_n is small enough (but not zero).

The notion is extended to also $dt_n < 0$ with $|t_n|$ approaching 0. In mathematical terms this is usually expressed as: For any given $\epsilon > 0$ there is a $\delta > 0$ such that

$$|x(t+dt) - x(t)| < \epsilon \quad \text{if } |dt| < \delta. \tag{81.3}$$

This looks more precise or "mathematical", but if you do not relate δ to ϵ, then it is as vague as saying that $x(t+dt)$ is close to $x(t)$ if dt is small, which is a pure qualitative statement.

81.2 Quantitative vs Qualitative Definitions

On the other hand, in our definitions of Lipschitz continuity and differentiability, no limits are visible. You can argue that our definitions are more precise since they are quantitative, not just qualitative, as expressed through the constants L and C_u.

Is it good or bad? Are we missing something using this approach? Well, you may judge yourself? Does the notion of limit capture the essence of continuity and differentiability?

An answer may be suggested by Achilles and the tortoise: With the limit/sequence definition, Achilles can be seen approaching the tortoise in a seemingly neverending sequential limit process with always half of the rest remaining, which appears paradoxical. With the definition without

limit/sequence, Achilles will simply at a certain moment in time have traveled the same distance as the tortoise and thereafter be ahead: No paradox.

In fact, the limit/sequence definitions, seemingly requiring that mysterious inifinitely small yet nonzero quantity dt, have created a lot of confusion and trouble through the history of Calculus, for both teachers and students, trouble which serves no reasonable purpose. To require that something locally is close to a constant or linear function, which we do in our limit/sequence-less definitions of continuity and differentiability, does not invite to any paradoxes, real or imagined.

The limit/sequence definitions, commonly viewed to be too difficult for high-school, form the core of university Honors Calculus with the pretention of giving a deeper understanding.

81.3 Sequences from Computation

In our approach, sequences will naturally occur as the output of a computational algorithm which generates a sequence $u_1, u_2, ...$, such as fixed point iteration/Newton's method, or by successive reduction of the time step k with a factor 2. We will meet sequences satisfying $|u_n - u_{n+1}| \leq \frac{C)}{2}^{-n}$, which uniquely determine a decimal expansion of a unique number. But we see no reason to consider other sequences than those generated by such computational algoritms. This simplifies the mathematics without loss of anything essential.

81.4 To Think About

- What is the use of limits in a standard Calculus text?
- What sequences in a standard Calculus text arise naturally?

82
Time Stepping Error Analysis

Sapiens nihil affirmat quod non probat.

After experience had taught me that all the usual surroundings of social life are vain and futile; seeing that none of the objects of my fears contained in themselves anything either good or bad, except in so far as the mind is affected by them, I finally resolved to inquire whether there might be some real good having power to communicate itself, which would affect the mind singly, to the exclusion of all else: whether, in fact, there might be anything of which the discovery and attainment would enable me to enjoy continuous, supreme, and unending happiness. (Spinoza)

82.1 Midpoint Euler

Consider a linear scalar IVP of the form: Find $u(t)$ such that

$$\dot{u}(t) + Au(t) = F(t) \quad \text{for } t > 0, \quad u(0) = u^0, \tag{82.1}$$

where A is a constant and $F(t)$ a given functions, which has the standard form $\dot{u}(t) = f(t, u(t))$ with $f(t, u) = F(t) - Au$.

Compute an approximate solution $U(t)$ by time stepping according to the Trapezoidal Method: Find $U^n = U(nk)$ such that

$$U^{n+1} + \frac{k}{2}(AU^{n+1} + AU^n) = U^n + \int_{I_n} F(t)\,dt, \quad \text{for } n = 0, 1, 2, ..., \tag{82.2}$$

with $I_n = (nk, (n+1)k)$ and $U^0 \approx u^0$, assuming the integral of $F(t)$ can be evaluated analytically. We here think of $U(t)$ as a piecewise linear continuous function taking on the computed values U^n at the discrete time levels nk with time step k. If so the Trapezoidal Method and Midpoint Euler coincide. In particular, we then have

$$\int_{I_n} \dot{U}\,dt = U^{n+1} - U^n, \quad \int_{I_n} AU\,dt = \frac{k}{2}(AU^{n+1} + AU^n), \tag{82.3}$$

which shows that Trapezoidal/Midpoint Euler satisfies:

$$\int_{I_n} (\dot{U} + AU - F)\,dt = 0 \quad \text{for } n = 0, 1, 2, ... \tag{82.4}$$

In other words, the mean-value over each subinterval I_n of the *residual* $R(U) \equiv \dot{U} + AU - F$ vanishes.

We shall consider the following basic choices of A with different stability characteristics connecting back to Summary: Timestepping of IVP:

1. constant non-negative: $A \geq 0$,
2. constant imaginary: $A = i$,
3. constant negative: $A < 0$,
4. oscillating positive-negative: $A(t) = \sin(t)$,

where we also added a basic case with $A(t)$ depending on t. We shall below see that (82.1) can also be interpreted as a system of differential equations with A a square matrix with the following analogous stability characteristics:

1. A positive semi-definite (symmetric): diffusion problems
2. A anti-symmetric: wave problems
3. A negative definite (symmetric) or general matrix: inverted pendulum...
4. A oscillating with zero mean: turbulence...

For a general non-linear system $\dot{u}+f(u) = 0$, the matrix A then corresponds to the Jacobian $f'(u(t))$ as concerns stability.

82.2 Error Analysis of Midpoint Euler

We shall now estimate the error $|u(T) - U(T)|$ at a given time $T = Nk$ in terms of the time step k and relevant quantities to be defined, where

we assume the $u(t)$ is an exact solution satisfying (82.1) to high precision (computed with a (vanishingly) small time step).

We shall do this by representing the error in terms of the solution $\varphi(t)$ of the following *dual problem*:

$$\begin{cases} -\dot{\varphi} + A\varphi = 0 & \text{for } T > t \geq 0, \\ \varphi(T) = \pm 1 = \text{sign of } e(T), \end{cases} \tag{82.5}$$

where $e = u - U$. We note that that (82.5) runs "backwards in time" starting at time T, because the (initial) data is given at $t = T$, and that (accordingly) the time derivative term $\dot{\varphi}$ has a minus sign. We start from the identity

$$0 = \int_0^T e\,(-\dot{\varphi} + A\varphi)\,dt,$$

and integrate by parts to get the following error representation (since $|e(T)| = e(T)\varphi(T)$):

$$|e(T)| = \int_0^T (\dot{e} + Ae)\varphi\,dt + e(0)\varphi(0),$$

where we allow $U(0)$ to be different from $u(0)$, corresponding to an error $e(0)$ in the initial value $u(0)$. Since u solves the differential equation (220.10), that is $\dot{u} + Au = F$, we have

$$\dot{e} + Ae = \dot{u} + Au - F - (\dot{U} + AU - F) = F - \dot{U} - AU = -R(U),$$

and thus we obtain the following representation of the error $|e(T)|$ in terms of the residual $R(U) = \dot{U} + AU - F$ and the dual solution φ:

$$|e(T)| = -\int_0^T R(U)\varphi\,dt + e(0)\varphi(0). \tag{82.6}$$

Recalling (82.4) we have

$$\int_{I_n} R(U)\,dt = 0 \quad \text{for } n = 0, 1, 2, ...,$$

which allows us to rewrite (82.6) as

$$|e(T)| = -\int_0^T R(U)(\varphi - \bar{\varphi})\,dt + e(0)\varphi(0), \tag{82.7}$$

where $\bar{\varphi}$ is the mean-value of φ over each time interval I_n, that is

$$\bar{\varphi}(t) = \frac{1}{k}\int_{I_n} \varphi(s)\,ds \quad \text{for } t \in I_n.$$

We shall now use the fact that

$$\int_{I_n} |\varphi - \bar{\varphi}|\, dt \le k \int_{I_n} |\dot{\varphi}|\, dt,$$

which follows by integration from the facts that

$$\varphi(t) - \bar{\varphi}(t) = \frac{1}{k} \int_{I_n} (\varphi(t) - \varphi(s))\, ds,$$

and

$$|\varphi(t) - \varphi(s)| \le \int_s^t |\dot{\varphi}(\sigma)|\, d\sigma \le \int_{I_n} |\dot{\varphi}(\sigma)|\, d\sigma \quad \text{for } s, t \in I_n.$$

Thus, (220.13) implies

$$\begin{aligned} |e(T)| &\le \sum_{n=0}^{N-1} R_n \int_{I_n} |\varphi - \bar{\varphi}| dt + |e(0)||\varphi(0)| \\ &\le \sum_{n=0}^{N-1} k\, R_n \int_{I_n} |\dot{\varphi}| dt + |e(0)||\varphi(0)|, \end{aligned} \tag{82.8}$$

where

$$R_n(U) = \max_{t \in I_n} |R(U(t))|.$$

Bringing out the max of $k_n R_n$ over n, we get

$$|e(T)| \le \max_{0 \le n \le N-1} k R_n \int_0^T |\dot{\varphi}|\, dt + |e(0)||\varphi(0)|.$$

Defining the *stability factors* $S_c(T)$ and $S_d(T)$ by

$$S_c(T) = \int_0^T |\dot{\varphi}(s)|\, ds, \quad S_d(T) = |\varphi(0)|, \tag{82.9}$$

we get the following *a posteriori error estimate*:

Theorem 82.1 *The approximate solution $U(t)$ of the initial value problem (82.1) computed by Midpoint Euler with time step k over intervals I_n, satisfies for $T > 0$*

$$|u(T) - U(T)| \le S_c(T) \max_n k R_n(U) + S_d(T)|u(0) - U(0)|, \tag{82.10}$$

where $u(t)$ is the exact solution computed with vanishingly small time step, $R_n(U) = \max_{I_n} |\dot{U} + AU - F|$ measures the residual over I_n, and $S_c(T)$ and $S_c(T)$ are stability factors defined by (82.9) related to time-discretization and initial data.

The stability factors $S_c(T)$ and $S_d(T)$ measure the effects of the accumulation of error in the approximation. To give the analysis a quantitative meaning, we have to give a quantitative bound of these factors. In general the stablity factors are computed by computing the solution of the dual problem. In special cases the stability factors can be computed analytically, as we now show:

The following lemma gives an estimate for $S_c(T)$ and $S_d(T)$ depending on the nature of A, in particular the sign of A, with possibly vastly different stability factors. We notice that the solution $\varphi(t)$ of (220.11) if A is constant is given by the explicit formula

$$\varphi(t) = \pm \exp(-A(T-t)).$$

We see that if $A > 0$, then the solution $\varphi(t)$ decays as t decreases from T, and the case $A > 0$ is thus the "stable case". If $A < 0$ then the exponential factor $\exp(-AT)$ enters, and depending on the size of A this case is "unstable". More precisely, we conclude directly from the explicit solution formula that

Theorem 82.2 *The stability factors $S_c(T)$ and $S_d(T)$ satisfy if $A < 0$:*

$$S_d(T) \le \exp(|A|T), \quad S_c(T) \le \exp(|A|T), \tag{82.11}$$

if $A \ge 0$:

$$S_d(T) \le 1, \quad S_c(T) \le 1 \tag{82.12}$$

if $A = i$:

$$S_d(T) \le 1, \quad S_c(T) \le T, \tag{82.13}$$

if $A = \sin(t)$:

$$S_d(T) \le \exp(1), \quad S_c(T) \le exp(1)T. \tag{82.14}$$

Proof: Changing variables $T - t \to t$, we can write the dual equation as the forward-in-time problem $\dot\varphi = -A\varphi$ for $t > 0$, $\varphi(0) = 1$ with solution $\exp(-At)$, if A is constant. We note that if $A > 0$, then

$$\int_0^T |\dot\varphi(t)|\,dt = \int_0^T A\exp(-At)\,dt = -\int_0^T \frac{d}{dt}\exp(-At)\,dt = 1-\exp(-AT) < 1. \tag{82.15}$$

Further, if $A < 0$, then

$$\int_0^T |\dot\varphi(t)|\,dt = \int_0^T \frac{d}{dt}\exp(-At)\,dt = \exp(-AT) - 1 \approx \exp(|A|T) \tag{82.16}$$

If $A = i$, then

$$\int_0^T |\dot\varphi(t)|\,dt = \int_0^T |\frac{d}{dt}\exp(-it)|\,dt = \int_0^T 1\,dt = T. \tag{82.17}$$

Finally, if $A = \sin(t)$ then $\varphi(t) = \exp(\cos(t))$, and so

$$|\varphi(T)| \leq \exp(1), \quad \int_0^T |\dot{\varphi}(t)|\,dt \leq \exp(1)T. \tag{82.18}$$

■.

The size of the stability factors indicate the degree of stability of the solution $u(t)$ being computed. If the stability factors are large, the residuals $R(U(t))$ and $e(0)$ have to be made correspondingly smaller by choosing smaller time steps and the computational problem is more demanding.

82.3 A Priori Error Estimate

The a posteriori error estimate (82.19) estimates the error in terms of the computed solution $U(t)$. There is a corresponding *a priori error estimate* with $R(U)$ replaced by $R(u_k)$ where u_k is the piecewise linear interpolant of the exact solution $u(t)$ taking on the same values at the discrete time levels nk. In this case the stability factors measure the stability of a corresponding *discrete dual problem*.

How big is then $R(u_k)$? Well, with piecewise linear interpolation, we have $|\dot{u} - \dot{u}_k| \approx k|\ddot{u}|$, and thus the a priori estimate takes the form

$$|u(T) - U(T)| \leq S_c(T)C(u)k^2 + S_d(T)|e(0)|, \tag{82.19}$$

where $C(u) = max_t|\ddot{u}(t)|$. In short, Midpoint Euler is *second-order accurate* with error proportional to k^2. Backward Euler and Forward Euler are *first order accurate* with error proportional to k.

82.4 Generalization

The above error analysis extends to a general IVP $\dot{u}(t) = f(u(t))$ for $t > 0$, $u(0) = u^0$, as shown in Chapter (161).

82.5 To Think About

- Show that the a posteriori estimate (82.19) directly extends to variable time steps k_n with k_nR_n replacing kR_n.
- For a basic aspect of duality in error estimation, see Error Control by Duality.

83
Integration in Several Dimensions

> This is a tricky domain because, unlike simple arithmetic, to solve a calculus problem - and in particular to perform integration - you have to be smart about which integration technique should be used: integration by partial fractions, integration by parts, and so on. (Marvin Minsky)

Let $f(x) = f(x_1, x_2)$ by a Lipschitz continuous real-valued function of $x = (x_1, x_2)$ defined on the unit square $\Omega = \{x : 0 \leq x_1, x_2 \leq 2\}$, that is $f : \Omega \to \mathbb{R}$. We define by iterated 1d integration

$$\int_\Omega f(x)\,dx = \int_0^1 (\int_0^1 f(x_1, x_2)dx_1)\,dx_2 = \int_0^1 (\int_0^1 f(x_1, x_2)dx_2)\,dx_1, \tag{83.1}$$

where the order of 1d integration is irrelevant, since both integrals express the common Riemann sum

$$\sum_{i,j=1}^{N} f(ih, jh)h^2. \tag{83.2}$$

Generalization to any domain in 2d or 3d is direct, by iterated 1d integration. Linearity properties follow directly from the Riemann sum representation. By the triangle inequality it follows that

$$|\int_\Omega f(x)\,dx| \leq \int_\Omega |f(x)|\,dx. \tag{83.3}$$

Cauchy's inequality takes the form

$$|\int_\Omega f(x)g(x)\,dx| \leq (\int_\Omega f^2(x)\,dx)^{\frac{1}{2}}(\int_\Omega g^2(x)\,dx)^{\frac{1}{2}}. \qquad (83.4)$$

83.1 Learn More

- Double integrals
- Multiple integrals

83.2 To Think About

- Is there a Fundamental Theorem for integration in 3d?

84
The Divergence Theorem

> Many who have had an opportunity of knowing any more about mathematics confuse it with arithmetic, and consider it an arid science. In reality, however, it is a science which requires a great amount of imagination. — Say what you know, do what you must, come what may. — It is impossible to be a mathematician without being a poet in soul(Sophia Kovalevskaya)

If Ω is a volume with boundary Γ with outward unit normal n, then

$$\int_\Omega \nabla \cdot u \, dx = \int_\Gamma u \cdot n \, ds, \tag{84.1}$$

which is referred to as the *Divergence Theorem* or alternatively *Gauss' Theorem*. We can see this result as a multidimensional analog of the Fundamental Theorem of Calculus:

$$\int_0^1 u'(x)dx = \int_0^1 \frac{du}{dx}dx = u(1) - u(0). \tag{84.2}$$

In the case Ω is the unit square in 2d or unit cube in 3d, the Divergence Theorem follows directly from the Fundamental Theorem. Show this, by consider the special case

$$\int_\Omega \frac{\partial u_1}{\partial x_1} \, dx_1 dx_2 = \int_\Gamma u_1 n_1 \, dx_2. \tag{84.3}$$

FIGURE 84.1. Sonya (Sophia) Kovalevskaya (1850-1891) of Russian origin, full professor of mathematics at the University of Stockholm 1889-91, remembered by the Cauchy-Kovalevskaya theorem, first female professor in Northern Europe: *It is impossible to be a mathematician without being a poet in soul...Many who have had an opportunity of knowing any more about mathematics confuse it with arithmetic, and consider it an arid science. In reality, however, it is a science which requires a great amount of imagination...Say what you know, do what you must, come what may...It seems to me that the poet has only to perceive that which others do not perceive, to look deeper than others look. And the mathematician must do the same thing.* .

84.1 Learn More

- The Divergence Theorem

85
Green's and Stokes' Theorems

If Ω is a domain in $\mathbb{R}^3$ with boundary Γ with outward unit normal $n = (n_1, n_2, n_3)$, and $u : \Omega \to \mathbb{R}^3$ and $v, w : \Omega \to \mathbb{R}$, then we obtain applying the Divergence Theorem to the product vw,

$$\int_\Omega \frac{\partial v}{\partial x_i} w \, dx = \int_\Gamma vw \, n_i \, ds - \int_\Omega v \frac{\partial w}{\partial x_i} \, dx, \quad i = 1, 2, 3.$$

Further, similarly,

$$\int_\Omega \nabla v \cdot \nabla w \, dx = \int_\Gamma v \partial_n w \, ds - \int_\Omega v \Delta w \, dx,$$

and

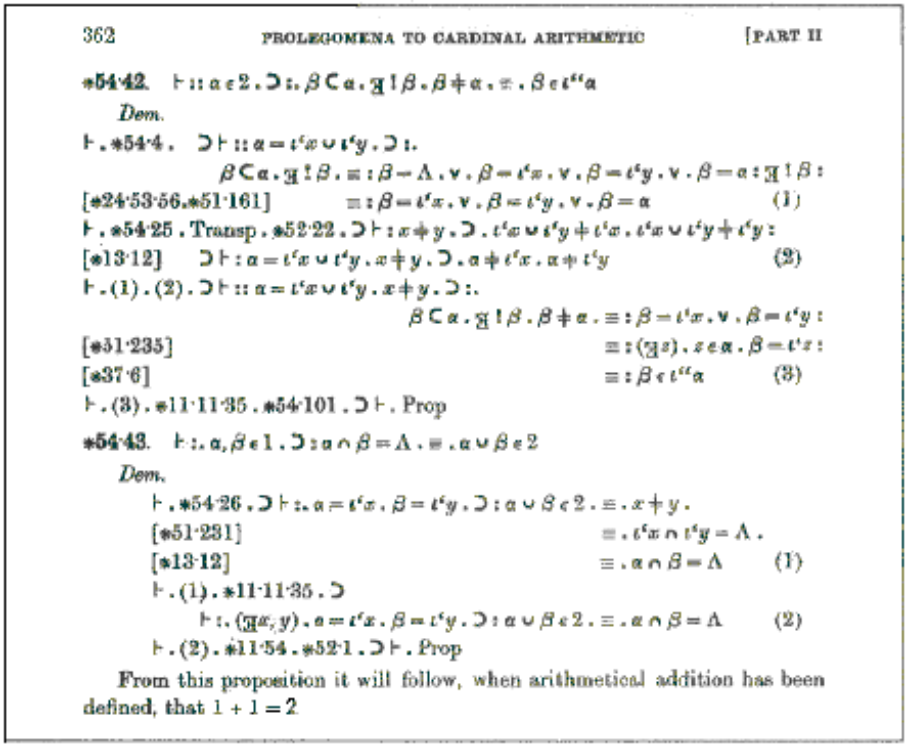
$$\int_\Omega v \Delta w \, dx - \int_\Omega \Delta v \, w \, dx = \int_\Gamma v \, \partial_n w \, ds - \int_\Gamma \partial_n v \, w \, ds.$$

These formulas are referred to as *Green's Formulas* and express 3d analogs to integration by parts in 1d.

If S is a surface in $\mathbb{R}^3$ bounded by a closed curve Γ, n is a unit normal to S, Γ is oriented in a clockwise direction following the positive direction of the normal n, and $u : \mathbb{R}^3 \to \mathbb{R}^3$ is differentiable, then

$$\int_S (\nabla \times u) \cdot n \, ds = \int_\Gamma u \cdot ds,$$

which is *Stokes' Theorem*.



362 PROLEGOMENA TO CARDINAL ARITHMETIC [PART II

***54·42.** $\vdash :: \alpha \in 2 . \supset : \beta \subset \alpha . \exists ! \beta . \beta \neq \alpha . \equiv . \beta \in \iota''\alpha$

Dem.

$\vdash . *54·4 . \quad \supset \vdash :: \alpha = \iota'x \cup \iota'y . \supset :.$

$\beta \subset \alpha . \exists ! \beta . \equiv : \beta = \Lambda . \vee . \beta = \iota'x . \vee . \beta = \iota'y . \vee . \beta = \alpha : \exists ! \beta :$

[*24·53·56.*51·161] $\equiv : \beta = \iota'x . \vee . \beta = \iota'y . \vee . \beta = \alpha$ (1)

$\vdash . *54·25 . \text{Transp} . *52·22 . \supset \vdash : x \neq y . \supset . \iota'x \cup \iota'y \neq \iota'x . \iota'x \cup \iota'y \neq \iota'y :$

[*13·12] $\supset \vdash : \alpha = \iota'x \cup \iota'y . x \neq y . \supset . \alpha \neq \iota'x . \alpha \neq \iota'y$ (2)

$\vdash . (1) . (2) . \supset \vdash :: \alpha = \iota'x \cup \iota'y . x \neq y . \supset :.$

$\beta \subset \alpha . \exists ! \beta . \beta \neq \alpha . \equiv : \beta = \iota'x . \vee . \beta = \iota'y :$

[*51·235] $\equiv : (\exists z) . z \in \alpha . \beta = \iota'z :$

[*37·6] $\equiv : \beta \in \iota''\alpha$ (3)

$\vdash . (3) . *11·11·35 . *54·101 . \supset \vdash . \text{Prop}$

***54·43.** $\vdash :. \alpha, \beta \in 1 . \supset : \alpha \cap \beta = \Lambda . \equiv . \alpha \cup \beta \in 2$

Dem.

$\vdash . *54·26 . \supset \vdash :. \alpha = \iota'x . \beta = \iota'y . \supset : \alpha \cup \beta \in 2 . \equiv . x \neq y .$

[*51·231] $\equiv . \iota'x \cap \iota'y = \Lambda .$

[*13·12] $\equiv . \alpha \cap \beta = \Lambda$ (1)

$\vdash . (1) . *11·11·35 . \supset$

$\vdash :. (\exists x, y) . \alpha = \iota'x . \beta = \iota'y . \supset : \alpha \cup \beta \in 2 . \equiv . \alpha \cap \beta = \Lambda$ (2)

$\vdash . (2) . *11·54 . *52·1 . \supset \vdash . \text{Prop}$

From this proposition it will follow, when arithmetical addition has been defined, that $1 + 1 = 2$

FIGURE 85.1. Proof that $1 + 1 = 2$ in Principia Mathematica

85.1 Read More

- The Divergence Theorem
- Stokes' Theorem

86
Who Invented Calculus?

Newton's Calculus is based on geometry and his notion of "fluxions", while Leibniz' Calculus is based on an algebra, a machine, for derivation and integration. Newton accussed Leibniz for plagiarism, and managed to get the Royal Society to set up a committee to pronounce on the priority dispute. The report of the committee, finding in favor of Newton, was written by Newton himself, while Leibniz was kept out to give his version of the events. Leibniz was defeated to death, but Leibniz' Calculus survived and is what we use today, while Newton's fluxions are forgotten. See

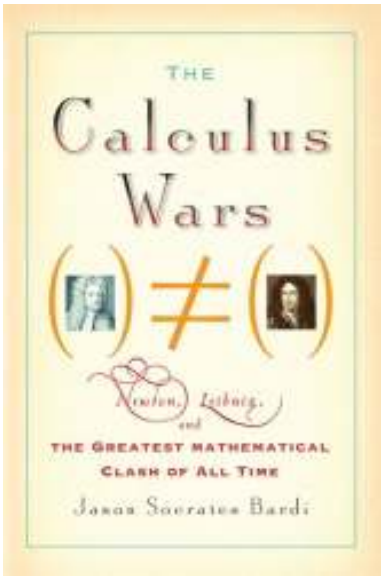
- The Newton-Leibniz Controversy
- The Baroque Cycle by Neal Stephenson.
- Letters between Newton and Leibniz in 6 minutes
- BBC Controversy Documentary (with my friend Erwing Stein)

We recall the beauty of Leibniz notation in his formulation of the Fundamental Theorem of Calculus:

$$\int_a^b \frac{du}{dt}dt = \int_a^b du = u(b) - u(a). \tag{86.1}$$

86.1 Watch

- Newton vs Leibniz



THE
METHOD of FLUXIONS
AND
INFINITE SERIES;
WITH ITS
Application to the Geometry of CURVE-LINES.

By the INVENTOR
Sir ISAAC NEWTON, *K^t.*
Late Preſident of the Royal Society.

Tranſlated from the AUTHOR's LATIN ORIGINAL
not yet made publick.

To which is fubjoin'd,
A PERPETUAL COMMENT upon the whole Work,
Conſiſting of
ANNOTATIONS, ILLUSTRATIONS, and SUPPLEMENTS,
In order to make this Treatiſe
A compleat Inſtitution for the uſe of LEARNERS.

By *JOHN COLSON*, M.A. and F.R.S.
Maſter of Sir *Joſeph Williamſon*'s free Mathematical-School at *Rochester*.

LONDON:
Printed by HENRY WOODFALL;
And Sold by JOHN NOURSE, at the *Lamb* without *Temple-Bar*.
M.DCC.XXXVI.

FIGURE 86.1. The Controversy.

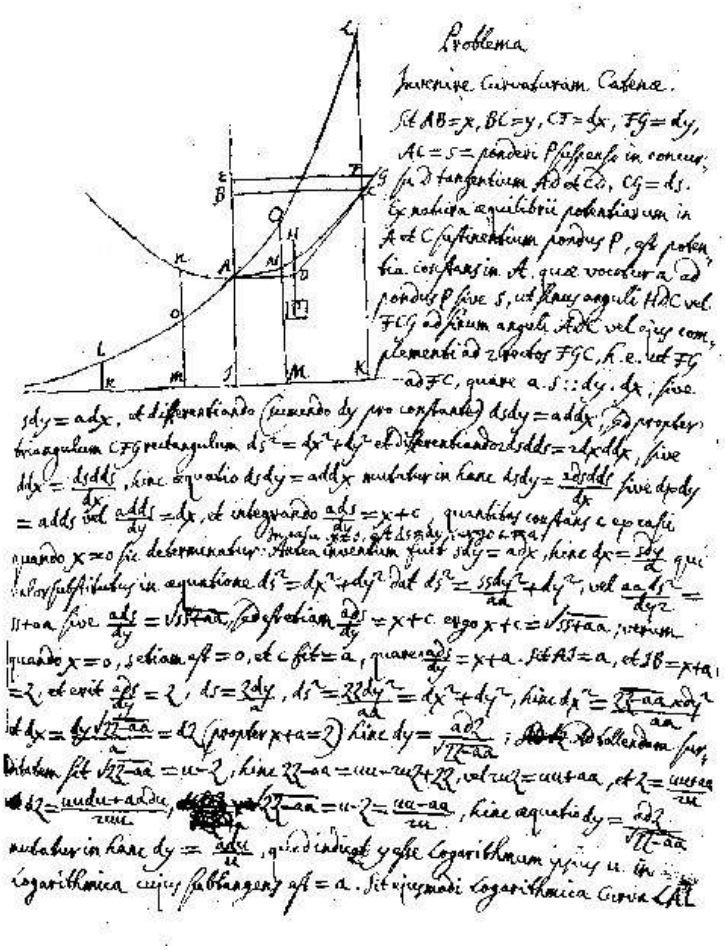


FIGURE 86.2. Problem formulated and solved by Leibniz Calculus.

87
Perspectives of Reformation

> The fewer the words, the better the prayer....When schools flourish, all flourishes... Music is the art of the prophets and the gift of God. (Martin Luther)

You have now met the basics of Calculus and you have seen that mathematical modeling involves (i) formulating and (ii) solving equations.

Traditional elementary school mathematics concerns a linear algebraic equation expressing proportionality:

$$ax + b = c, \tag{87.1}$$

with $a \neq 0$ and b and c given numbers, with solution

$$x = \frac{c - b}{a}. \tag{87.2}$$

To compute the solution of $ax + c = b$ both sides are subtracted by c and then multiplied by $\frac{1}{a}$ to give $x = \frac{1}{a}ax = \frac{c-b}{a}$. Not difficult.

I highschool mathematics the scope is extended to the quadratic algebraic equation

$$ax^2 + 2bx + c = 0 \tag{87.3}$$

with solution

$$x = -\frac{b}{a} \pm \sqrt{\frac{b^2}{a^2} - \frac{c}{a}}, \tag{87.4}$$

assuming $b^2 \geq ac$ so that the squareroot is a real number (it is imaginary if $b^2 < ac$). To solve for x, the equation is divided by a to give

$$x^2 + 2\frac{b}{a}x + \frac{b^2}{a^2} = \frac{b^2}{a^2} - c \tag{87.5}$$

obtained by adding $\frac{b^2}{a^2}$ to both sides (and moving c to the right hand side). This is referred to as "completing the square" on the left hand side, since

$$x^2 + 2\frac{b}{a}x + \frac{b^2}{a^2} = (x + \frac{b}{a})^2. \tag{87.6}$$

With this clever operation, the equation takes the form

$$(x + \frac{b}{a})^2 = \frac{b^2}{a^2} - \frac{c}{a} \tag{87.7}$$

from which follows that

$$x + \frac{b}{a} = \pm\sqrt{\frac{b^2}{a^2} - \frac{c}{a}} \tag{87.8}$$

which gives the solution formula (87.4).

In traditional school mathematics students are supposed to solve a large number of equations of the form (87.2) and (87.4) by using the analytical solution formulas. The objective of this activity, which fills a large part of the schedule, is to demonstate to students that analytical mathematics is powerful and can be used to solve many equations.

The trouble with this approach is that it does no longer work: Many students find it difficult to formulate linear or quadratic algebraic equations from some given information, and tend to mix up the solution formulas and don't see the beauty of "completing the square". On top of that there is a growing insight that analytical solution formulas are non-existing for more general problems: Already a 3rd order equation is tricky and a 5th order impossible to solve by taking roots (as shown by Galois shortly before his tragic death in a silly duel in 1832 at the age of 20).

On the other hand we have seen that just about any equation is solvable computationally, by Newton's method for algebraic equations, and time stepping for differential equations. The analytical solution formula (87.4) for a quadratic equation in fact involves a root, which will have to be computed somehow anyway, and thus instead of memorizing the formula with its trick of "completing the square", we may as well apply Newton's method to the original form of the equation and forget the trick.

In computational mathematics we can use **methods** with general applicability, while in analytical mathematics we are restricted to tricks for a few very special cases. We understand that there is a big difference between a general method and a bag of tricks, and that this difference requires a reformation of mathematics education...

FIGURE 87.1. Martin Luther posting his 95 Theses in Wittenberg in 1517 sparking the Protestant Reformation: *Who loves not wine, women and song, Remains a fool his whole life long... If you are not allowed to laugh in heaven, I don't want to go there.*

88
How to Learn and Use Calculus

> Among all of the mathematical disciplines the theory of differential equations is the most important... It furnishes the explanation of all those elementary manifestations of nature which involve time. (Sophus Lie, famous Norwegian mathematician)
>
> I recoil with dismay and horror at this lamentable plague of functions which do not have derivatives. (Charles Hermite)
>
> But just as much as it is easy to find the differential [derivative] of a given quantity, so it is difficult to find the integral of a given differential. Moreover, sometimes we cannot say with certainty whether the integral of a given quantity can be found or not. (Johann Bernoulli)

You have now been exposed to the basics of Calculus, the wonderful invention by Leibniz and Newton, and you have gathered some experience. You have seen that Calculus concerns

- functions together with derivates and integrals of functions,
- differential equations expressed in terms of derivatives functions,
- (there are also integral equations in terms of integrals of functions).

Further, you have

- discovered rules for differentiation and integration,

- met elementary functions as solutions to basic (elementary) differential equations,
- discovered their properties through their defining differential equations.

All of this can be captured in a set of formulas. You find in mathematics teaching and learning different attitudes to these formulas:

1. Learn formulas by heart and apply them to many similar problems.
2. Learn to prove a few basic formulas from basics and a few basic computational algorithms, implement them in computer code, and use them for a rich variety of problems.

Case 1 is very common in traditional mathematics focussed on problem solving, which is evidenced by large collections of similar problems to be solved, one after the other. There are endless variations of proportionality leading to the simple equation $ax = b$ in x, or the simple 2×2 system of equations $ax + by = c$ and $dx + ey = f$ in x and y.

We advocate case 2, where the essence is not to remember a certain formula or algorithm by heart, but instead to remember that you have once proved the formula and convergence of a computation algorithm, and implemented the formula and algorithm in computer code. With this approach you are able to solve e.g. an algebraic equation of any order by Newton's method, by using your computer.

The advantage of 2 is that you can use mathematics for a much richer variety of problems, by letting the computer do computational work:

- mathematics with computational turbo,
- mathematics as instruction telling the computer what to do,
- a combination of brains and pedals moving you forward.

FIGURE 88.1. Mathematics with Turbo: Brains and Pedals.

Part V

Descartes' World of Analytic Geometry

FIGURE 88.2. 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...As for me, I have never presumed my mind to be in any way better than the minds of people in general. As for reason or good sense, I am inclined to believe that it exists whole and complete in each of us, because it is the only thing that makes us men and distinguishes us from the lower animals.*

89
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)

89.1 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 num-

ber. 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.

89.2 Descartes, Inventor of Analytic Geometry

The foundation of modern science was laid by René Descartes (1596-1650) in *Discours de la méthod pour bien conduire sa raison et chercher la vérité dans les sciences* from 1637. *The Method* contained as an appendix *La Geometrie* with the first treatment of Analytic Geometry. Descartes believed that only mathematics may be certain, so all must be based on mathematics, the foundation of the *Cartesian* view of the World.

In 1649 Queen Christina of Sweden persuaded Descartes to go to Stockholm to teach her mathematics. However the Queen wanted to draw tangents at 5 a.m. and Descartes broke the habit of his lifetime of getting up at 11 o'clock, c.f. Fig. 88.2. After only a few months in the cold Northern climate, walking to the palace at 5 o'clock every morning, he died of pneumonia.

89.3 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. 89.1. 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 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.

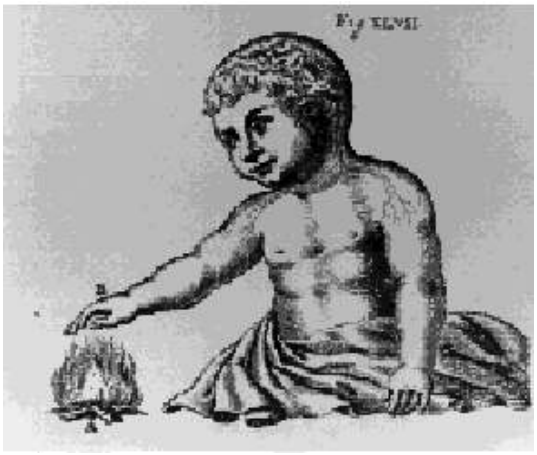


FIGURE 89.1. Automatic reaction in response to external stimulation from Descartes *De homine* 1662.

89.4 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. 89.2. The coordinates of the origin are $(0, 0)$.

In this way, we identify each point a in the plane with its coordinates (a_1, a_2), and we may thus represent the Euclidean plane as $\mathbb{R}^2$, where $\mathbb{R}^2$ is the set of ordered pairs (a_1, a_2) of real numbers a_1 and a_2. That is

$$\mathbb{R}^2 = \{(a_1, a_2) : a_1, a_2 \in \mathbb{R}\}.$$

We have already used $\mathbb{R}^2$ as a coordinate system above when plotting a function $f : \mathbb{R} \to \mathbb{R}$, where pairs of real numbers $(x, f(x))$ are represented as geometrical points in a Euclidean plane on a book-page.

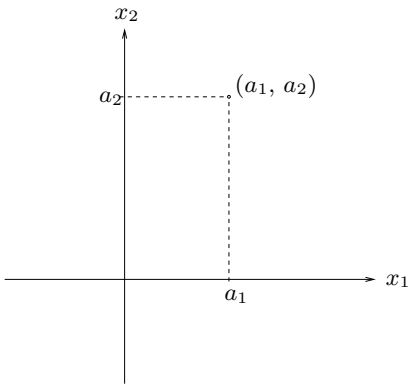


FIGURE 89.2. Coordinate system for $\mathbb{R}^2$

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. 89.2 or Fig. 89.3 displaying MATLAB's view of a coordinate system. The positive direction of each coordinate axis may be indicated by an arrow in the direction of increasing coordinates.

However, this is just one possibility. For example, to describe the position of points on a computer screen or a window on such a screen, it is not uncommon to use coordinate systems with the origin at the upper left corner and counting the a_2 coordinate positive down, negative up.

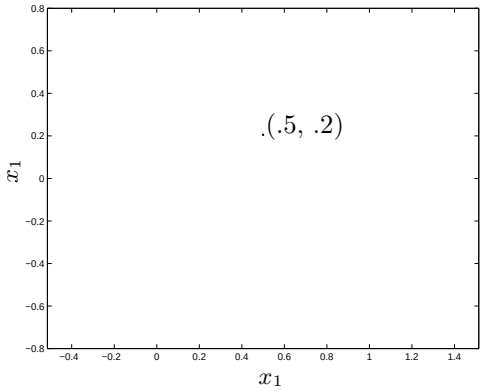


FIGURE 89.3. Matlabs way of visualizing a coordinate system for a plane.

89.5 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.

89.6 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.

89.7 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. 89.4. With the arrow interpretation of (a_1, a_2), we refer to (a_1, a_2) as a *vector*. Again, we agree to write $a = (a_1, a_2)$, and we say that a_1 and a_2 are the *components* of the arrow/vector $a = (a_1, a_2)$. We say that a_1 is the *first component*, occurring in the first place and a_2 the *second component* occurring in the second place.

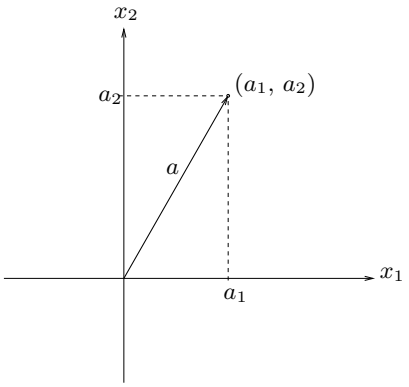


FIGURE 89.4. A vector with tail at the origin and the head at the point $a = (a_1, a_2)$

We thus may interpret an ordered pair (a_1, a_2) in $\mathbb{R}^2$ in two ways: as a point with coordinates (a_1, a_2), or as an arrow/vector with components (a_1, a_2) starting at the origin and ending at the point (a_1, a_2). Evidently, there is a very strong connection between the point and arrow interpretations, since the head of the arrow is located at the point (and assuming that the arrow tail is at the origin). In applications, *positions* will be connected to the point interpretation and *velocities* and *forces* will be connected to

the arrow/vector interpretation. We will below generalize the arrow/vector interpretation to include arrows with tails also at other points than the origin. The context will indicate which interpretation is most appropriate for a given situation. Often the interpretation of $a = (a_1, a_2)$ as a point or as an arrow, changes without notice. So we have to be flexible and use whatever interpretation is most convenient or appropriate. We will need even more fantasy when we go into applications to mechanics below.

Sometimes vectors like $a = (a_1, a_2)$ are marked by boldface or an arrow, like $\mathbf{a}$ or $\vec{a}$ or $\underline{a}$, or double script or some other notation. We prefer not to use this more elaborate notation, which makes the writing simpler, but requires fantasy from the user to make the proper interpretation of for example the letter a as a scalar number or vector $a = (a_1, a_2)$ or something else.

89.8 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

$$(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 + b_2), \tag{89.1}$$

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 89.1. We have $(2, 5)+(7, 1) = (9, 6)$ and $(2.1, 5.3)+(7.6, 1.9) = (9.7, 7.2)$.

89.9 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. 89.5. 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. 89.5.

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. 89.5.

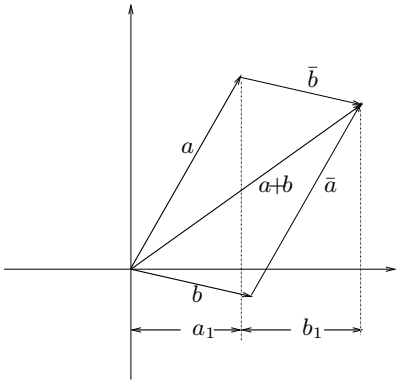


FIGURE 89.5. Vector addition using the Parallelogram Law

We sum up in the following theorem:

Theorem 89.1 *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. 89.6.

$$(a_1, a_2) = (a_1, 0) + (0, a_2). \tag{89.2}$$

89.10 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

$$\lambda a = \lambda(a_1, a_2) = (\lambda a_1, \lambda a_2). \tag{89.3}$$

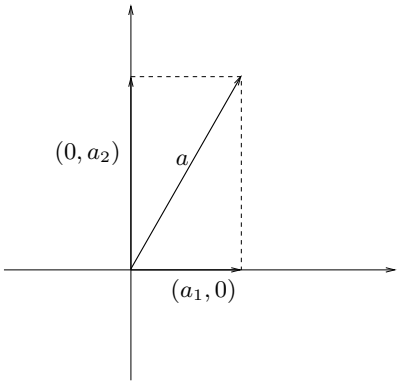


FIGURE 89.6. A vector a represented as the sum of two vectors parallel with the coordinate axes.

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. 89.7.

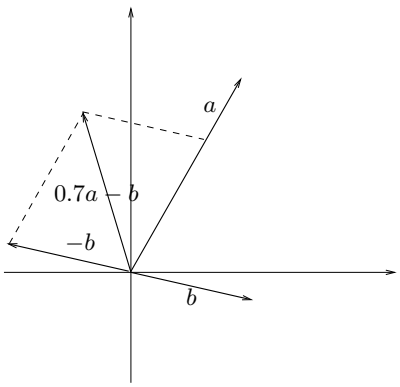


FIGURE 89.7. The sum $0.7a - b$ of the multiples $0.7a$ and $(-1)b$ of a and b.

89.11 The Norm of a Vector

We define the *Euclidean norm* $|a|$ of a vector $a = (a_1, a_2) \in \mathbb{R}^2$ as

$$|a| = (a_1^2 + a_2^2)^{1/2}. \tag{89.4}$$

By Pythagoras theorem and Fig. 89.8, the Euclidean norm $|a|$ of the vector $a = (a_1, a_2)$ is equal to the length of the hypothenuse of the right angled triangle with sides a_1 and a_2. In other words, the Euclidean norm of the vector $a = (a_1, a_2)$ is equal to the distance from the origin to the point $a = (a_1, a_2)$, or simply the length of the arrow (a_1, a_2). We have $|\lambda a| = |\lambda||a|$ if $\lambda \in \mathbb{R}$ and $a \in \mathbb{R}^2$; multiplying a vector by the real number λ changes the norm of the vector by the factor $|\lambda|$. The zero vector $(0, 0)$ has Euclidean norm 0 and if a vector has Euclidean norm 0 then it must be the zero vector.

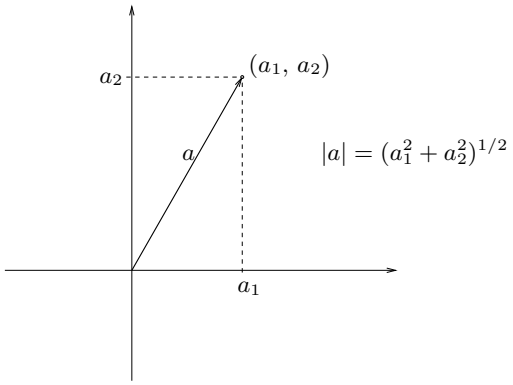


FIGURE 89.8. The norm $|a|$ of a vector $a = (a_1, a_2)$ is $|a| = (a_1^2 + a_2^2)^{1/2}$.

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 89.2. If $a = (3, 4)$ then $|a| = \sqrt{9 + 16} = 5$, and $|2a| = 10$.

89.12 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. 89.9.

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. 89.9

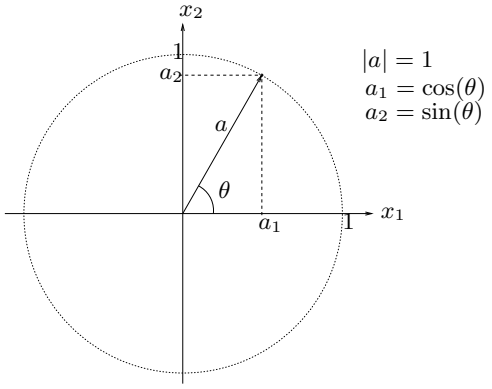


FIGURE 89.9. Vectors of length one are given by $(\cos(\theta), \sin(\theta))$

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

$$a = |a|\hat{a} = r(\cos(\theta), \sin(\theta)), \tag{89.5}$$

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. 89.10. We call (89.5) the *polar representation* of a. We call θ the direction of a and r the length of a, see Fig. 89.10.

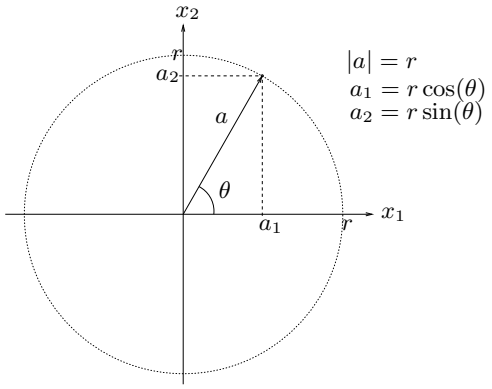


FIGURE 89.10. Vectors of length r are given by $a = r(\cos(\theta), \sin(\theta)) = (r\cos(\theta), r\sin(\theta))$ where $r = |a|$.

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 89.3. We have

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

89.13 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

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

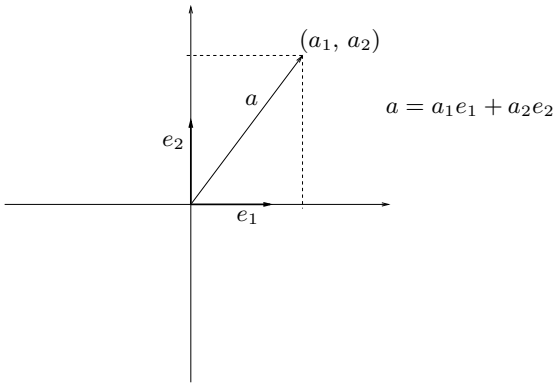


FIGURE 89.11. The standard basis vectors e_1 and e_2 and a linear combination $a = (a_1, a_2) = a_1 e_1 + a_2 e_2$ of e_1 and e_2.

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. 89.11.

EXAMPLE 89.4. We have $(3,7) = 3\,(1,0) + 7\,(0,1) = 3e_1 + 7e_2$.

89.14 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

$$a \cdot b = a_1 b_1 + a_2 b_2. \tag{89.6}$$

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:

$$|a| = (a \cdot a)^{\frac{1}{2}}. \tag{89.7}$$

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 89.5. We have $(3, 7)\cdot(5, 2) = 15+14 = 29$, and $(3, 7)\cdot(3, 7) = 9 + 49 = 58$ so that $|(3, 7)| = \sqrt{58}$.

89.15 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 (89.6). For example, we have

$$\begin{aligned} a \cdot (b + c) &= a_1(b_1 + c_1) + a_2(b_2 + c_2) \\ &= a_1 b_1 + a_2 b_2 + a_1 c_1 + a_2 c_2 = a \cdot b + a \cdot c. \end{aligned}$$

Using the notation $f(a, b) = a \cdot b$, the linearity properties may be written as

$$\begin{aligned} f(a, b + c) = f(a, b) + f(a, c), \qquad f(a + b, c) = f(a, c) + f(b, c), \\ f(\lambda a, b) = \lambda f(a, b) \qquad f(a, \lambda b) = \lambda f(a, b). \end{aligned}$$

We also say that the scalar product $a \cdot b = f(a, b)$ is a *bilinear form* on $\mathbb{R}^2 \times \mathbb{R}^2$, that is a function from $\mathbb{R}^2 \times \mathbb{R}^2$ to $\mathbb{R}$, since $a \cdot b = f(a, b)$ is a real number for each pair of vectors a and b in $\mathbb{R}^2$ and $a \cdot b = f(a, b)$ is linear both in the variable (or argument) a and the variable b. Furthermore, the scalar product $a \cdot b = f(a, b)$ is *symmetric* in the sense that

$$a \cdot b = b \cdot a \quad \text{or} \quad f(a, b) = f(b, a),$$

and *positive definite*, that is

$$a \cdot a = |a|^2 > 0 \quad \text{for } a \neq 0 = (0, 0).$$

We may summarize by saying that the scalar product $a \cdot b = f(a, b)$ is a *bilinear symmetric positive definite form on* $\mathbb{R}^2 \times \mathbb{R}^2$.

We notice that for the basis vectors $e_1 = (1, 0)$ and $e_2 = (0, 1)$, we have

$$e_1 \cdot e_2 = 0, \quad e_1 \cdot e_1 = 1, \quad e_2 \cdot e_2 = 1.$$

Using these relations, we can compute the scalar product of two arbitrary vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^2$ using the linearity as follows:

$$\begin{aligned} a \cdot b &= (a_1 e_1 + a_2 e_2) \cdot (b_1 e_1 + b_2 e_2) \\ &= a_1 b_1\, e_1 \cdot e_1 + a_1 b_2\, e_1 \cdot e_2 + a_2 b_1\, e_2 \cdot e_1 + a_2 b_2\, e_2 \cdot e_2 = a_1 b_1 + a_2 b_2. \end{aligned}$$

We may thus define the scalar product by its action on the basis vectors and then extend it to arbitrary vectors using the linearity in each variable.

89.16 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

$$a \cdot b = |a||b| \cos(\theta), \tag{89.8}$$

where θ is the angle between the vectors a and b, see Fig. 89.12. This formula has a geometric interpretation. Assuming that $|\theta| \leq 90^\circ$ so that $\cos(\theta)$ is positive, consider the right-angled triangle OAC shown in Fig. 89.12. The length of the side OC is $|a|\cos(\theta)$ and thus $a \cdot b$ is equal to the product of the lengths of sides OC and OB. We will refer to OC as the *projection* of OA onto OB, considered as vectors, and thus we may say that $a \cdot b$ is equal to the product of the length of the projection of OA onto OB and the length of OB. Because of the symmetry, we may also relate $a \cdot b$ to the projection of OB onto OA, and conclude that $a \cdot b$ is also equal to the product of the length of the projection of OB onto OA and the length of OA.

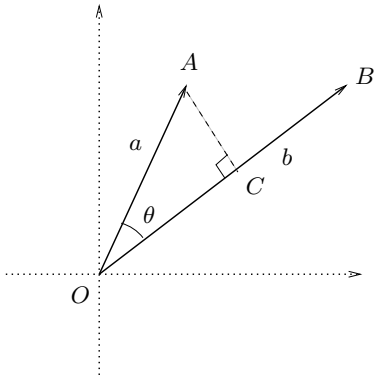


FIGURE 89.12. $a \cdot b = |a|\,|b|\cos(\theta)$.

To prove (89.8), 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 $\beta - \alpha$.

89.17 Orthogonality and Scalar Product

We say that two non-zero vectors a and b in $\mathbb{R}^2$ are *geometrically orthogonal*, which we write as $a \perp b$, if the angle between the vectors is 90° or 270°,

see Fig. 89.13. The basis vectors e_1 and e_2 are examples of geometrically orthogonal vectors, see Fig. 89.11.

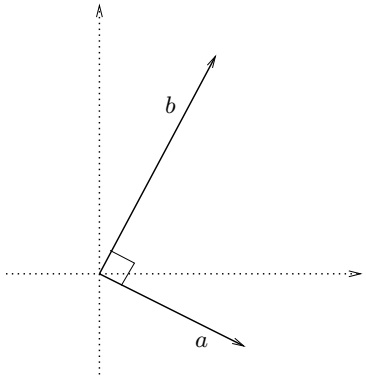


FIGURE 89.13. Orthogonal vectors a and b.

Let a and b be two non-zero vectors making an angle θ. From (89.8), 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 $\theta = 90^\circ$ or $\theta = 270^\circ$. We have now proved the following basic result, which we state as a theorem.

Theorem 89.2 *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

$$a_1 b_1 + a_2 b_2 = 0.$$

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 $a \cdot b = a_1 b_1 + a_2 b_2 = 0$.

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.

89.18 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.

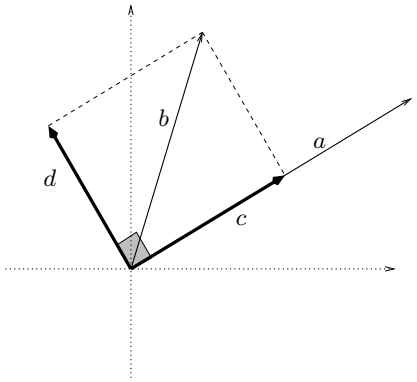


FIGURE 89.14. Orthogonal decomposition of b

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. 89.14. 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

$$P_a(b) = \frac{b \cdot a}{|a|^2}\, a. \tag{89.9}$$

We compute the length of $P_a(b)$ as

$$|P_a(b)| = \frac{|a \cdot b|}{|a|^2}|a| = \frac{|a|\,|b|\,|\cos(\theta)|}{|a|} = |b||\cos(\theta)|, \tag{89.10}$$

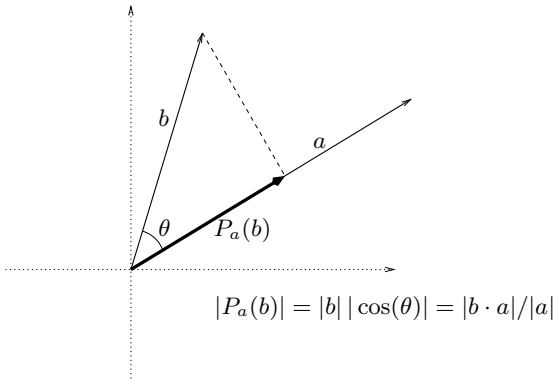


FIGURE 89.15. The projection $P_a(b)$ of b onto a.

where θ is the angle between a and b, and we use (89.8). We note that

$$|a \cdot b| = |a|\,|Pb|, \tag{89.11}$$

which conforms with our experience with the scalar product in Section 89.16, see also Fig. 20.15.

We can view the projection $P_a(b)$ of the vector b onto the vector a as a transformation of $\mathbb{R}^2$ into $\mathbb{R}^2$: given the vector $b \in \mathbb{R}^2$, we define the vector $P_a(b) \in \mathbb{R}^2$ by the formula

$$P_a(b) = \frac{b \cdot a}{|a|^2}\, a. \tag{89.12}$$

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

$$P(x + y) = Px + Py, \quad P(\lambda x) = \lambda Px, \tag{89.13}$$

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. 89.16.

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 89.3 *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 89.6. If $a = (1, 3)$ and $b = (5, 2)$, then $P_a(b) = \frac{(1,3)\cdot(5,2)}{1+3^2}(1, 3) = (1.1, 3.3)$.

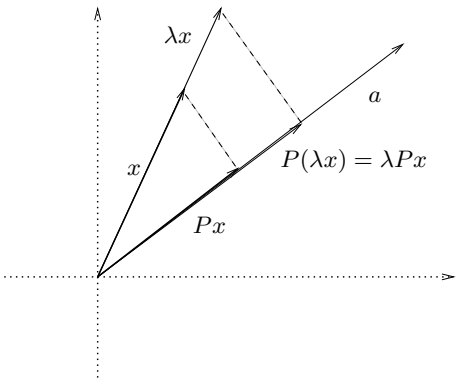


FIGURE 89.16. $P(\lambda x) = \lambda P x$.

89.19 Rotation by 90°

We saw above that to find the orthogonal decomposition $b = c + d$ with c parallel to a given vector a, it suffices to find c because $d = b - c$. Alternatively, we could seek to first compute d from the requirement that it should be orthogonal to a. We are thus led to the problem of finding a direction orthogonal to a given direction, that is the problem of rotating a given vector by 90°, which we now address.

Given a vector $a = (a_1, a_2)$ in $\mathbb{R}^2$, a quick computation shows that the vector $(-a_2, a_1)$ has the desired property, because computing its scalar product with $a = (a_1, a_2)$ gives

$$(-a_2, a_1) \cdot (a_1, a_2) = (-a_2)a_1 + a_1 a_2 = 0,$$

and thus $(-a_2, a_1)$ is orthogonal to (a_1, a_2). Further, it follows directly that the vector $(-a_2, a_1)$ has the same length as a.

Assuming that $a = |a|(\cos(\alpha), \sin(\alpha))$ and using the facts that $-\sin(\alpha) = \cos(\alpha + 90°)$ and $\cos(\alpha) = \sin(\alpha + 90°)$, we see that the vector $(-a_2, a_1) = |a|(\cos(\alpha + 90°), \sin(\alpha + 90°))$ is obtained by rotating the vector (a_1, a_2) counter-clockwise 90°, see Fig. 89.17. Similarly, the vector $(a_2, -a_1) = -(-a_2, a_1)$ is obtained by clockwise rotation of (a_1, a_2) by 90°.

We may view the counter clockwise rotation of a vector by 90° as a *transformation* of vectors: given a vector $a = (a_1, a_2)$, we obtain another vector $a^\perp = f(a)$ through the formula

$$a^\perp = f(a) = (-a_2, a_1),$$

where we denoted the image of the vector a by both $a^\perp$ and $f(a)$. The transformation $a \to a^\perp = f(a)$ defines a linear function $f : \mathbb{R}^2 \to \mathbb{R}^2$ since

$$f(a + b) = (-(a_2 + b_2), a_1 + b_1) = (-a_2, a_1) + (-b_2, b_1) = f(a) + f(b),$$
$$f(\lambda a) = (-\lambda a_2, \lambda a_1) = \lambda(-a_2, a_1) = \lambda f(a).$$

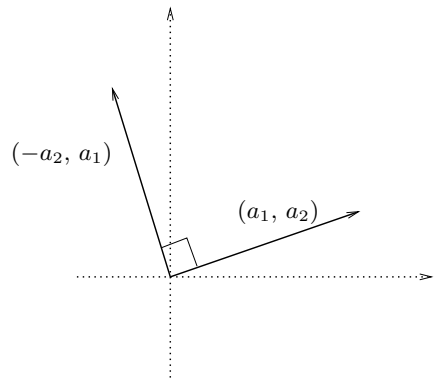


FIGURE 89.17. Counter-clockwise rotation of $a = (a_1, a_2)$ by 90°.

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 89.7. Rotating the vector $(1,2)$ the angle 90° counter-clockwise, we get the vector $(-2,1)$.

89.20 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

$$R_\theta(a) = (a_1\cos(\theta) - a_2\sin(\theta), a_1\sin(\theta) + a_2\cos(\theta)). \tag{89.14}$$

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 (89.14) 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 89.8. Rotating the vector $(1, 2)$ the angle $30°$, we obtain the vector $(\cos(30°) - 2\sin(30°), \sin(30°) + 2\cos(30°)) = (\frac{\sqrt{3}}{2} - 1, \frac{1}{2} + \sqrt{3})$.

89.21 Rotation by θ Again!

We present yet another way to arrive at (89.14) based on the idea that the transformation $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ of counter-clockwise rotation by θ is defined by the following properties,

$$\text{(i)} \quad |R_\theta(a)| = |a|, \quad \text{and} \quad \text{(ii)} \quad R_\theta(a) \cdot a = \cos(\theta)|a|^2. \tag{89.15}$$

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 (89.14).

89.22 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

$$\begin{aligned} a_1 &= \cos(\theta)\hat{a}_1 - \sin(\theta)\hat{a}_2, \\ a_2 &= \sin(\theta)\hat{a}_1 + \cos(\theta)\hat{a}_2. \end{aligned} \tag{89.16}$$

Since e_1 and e_2 are obtained by rotating $\hat{e}_1$ and $\hat{e}_2$ clockwise by the angle θ,

$$\begin{aligned} \hat{a}_1 &= \cos(\theta)a_1 + \sin(\theta)a_2, \\ \hat{a}_2 &= -\sin(\theta)a_1 + \cos(\theta)a_2. \end{aligned} \tag{89.17}$$

The connection between the coordinates with respect to the two coordinate systems is thus given by (89.16) and (89.17).

EXAMPLE 89.9. Rotating 45° counter-clockwise gives the following relation between new and old coordinates

$$\hat{a}_1 = \frac{1}{\sqrt{2}}(a_1 + a_2), \quad \hat{a}_2 = \frac{1}{\sqrt{2}}(-a_1 + a_2).$$

89.23 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

$$a \times b = a_1b_2 - a_2b_1. \tag{89.18}$$

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 $\mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$. This function is bilinear as is easy to verify, and *anti-symmetric*, that is

$$a \times b = -b \times a, \tag{89.19}$$

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,

$$e_1 \times e_1 = 0,\ e_2 \times e_2 = 0,\ e_1 \times e_2 = 1,\ e_2 \times e_1 = -1. \tag{89.20}$$

Using these relations,

$$a \times b = (a_1 e_1 + a_2 e_2) \times (b_1 e_1 + b_2 e_2) = a_1 b_2 e_1 \times e_2 + a_2 b_1 e_2 \times e_1 = a_1 b_2 - a_2 b_1 e_2.$$

We next show that the properties of bilinearity and anti-symmetry in fact determine the vector product in $\mathbb{R}^2$ up to a constant. First note that anti-symmetry and bilinearity imply

$$\begin{aligned} e_1 \times e_1 + e_1 \times e_2 = e_1 \times (e_1 + e_2) = -(e_1 + e_2) \times e_1 \\ = -e_1 \times e_1 - e_2 \times e_1. \end{aligned}$$

Since $e_1 \times e_2 = -e_2 \times e_1$, we have $e_1 \times e_1 = 0$. Similarly, we see that $e_2 \times e_2 = 0$. We conclude that the action of the vector product on the basis vectors is indeed specified according to (89.20) up to a constant.

We next observe that

$$a \times b = (-a_2, a_1) \cdot (b_1, b_2) = a_1 b_2 - a_2 b_1,$$

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 89.4 *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 $\theta + 90^\circ$:

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

where we use $|a^\perp| = |a|$ and $|\cos(\theta \pm \pi/2)| = |\sin(\theta)|$. Therefore,

$$|a \times b| = |a||b||\sin(\theta)|, \tag{89.21}$$

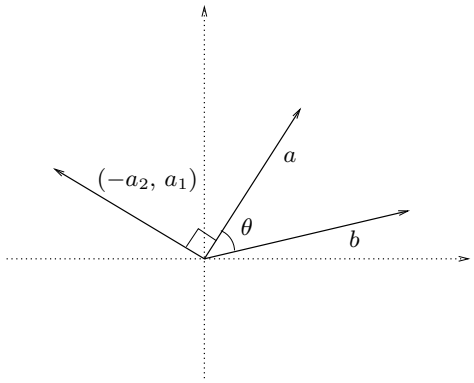


FIGURE 89.18. Why $|a \times b| = |a|\,|b|\,|\sin(\theta)|$.

where $\theta = \alpha - \beta$ is the angle between a and b, see Fig. 89.18.

We can make the formula (89.21) 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 (89.21), which we state as a theorem, see Fig. 89.19.

Theorem 89.5 *For two non-zero vectors a and b,*

$$a \times b = |a||b|\sin(\theta), \tag{89.22}$$

where θ is the angle between a and b counted positive counter-clockwise and negative clockwise starting from a.

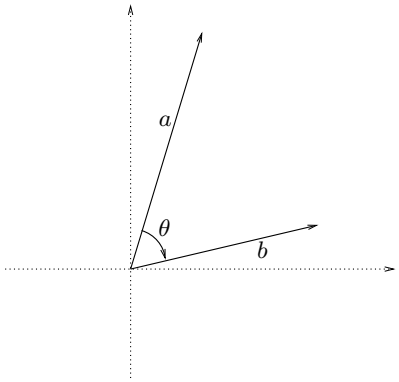


FIGURE 89.19. $a \times b = |a|\,|b|\sin(\theta)$ is negative here because the angle θ is negative.

89.24 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. 89.20. 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. 89.20. Recalling (89.21), we conclude

Theorem 89.6

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

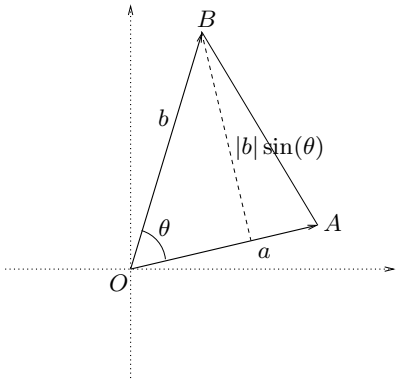


FIGURE 89.20. The vectors a and b span a triangle with area $\frac{1}{2}|a \times b|$.

The area of the triangle OAB can be computed using the vector product in $\mathbb{R}^2$.

89.25 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. 89.21.

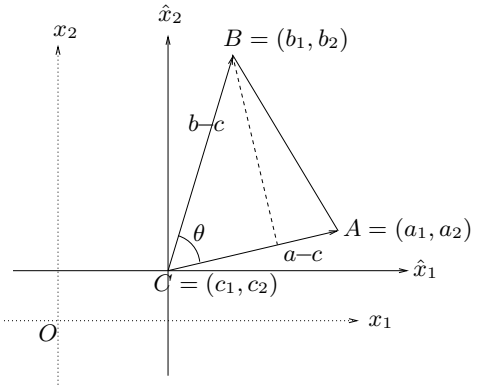


FIGURE 89.21. Vectors $a-c$ and $b-c$ span triangle with area $\frac{1}{2}|(a-c)\times(b-c)|$.

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

$$\text{Area}(CAB) = \frac{1}{2}|(a-c)\times(b-c)|. \tag{89.23}$$

EXAMPLE 89.10. The area of the triangle with coordinates at $A = (2,3)$, $B = (-2,2)$ and $C = (1,1)$, is given by $\text{Area}(CAB) = \frac{1}{2}|(1,2)\times(-3,1)| = \frac{7}{2}$.

89.26 The Area of a Parallelogram Spanned by Two Vectors

The area of the parallelogram spanned by a and b, as shown in Fig. 89.22, is equal to $|a\times b|$ since the area of the parallelogram is twice the area of the triangle spanned by a and b. Denoting the area of the parallelogram spanned by the vectors a and b by $V(a,b)$, we thus have the formula

$$V(a,b) = |a\times b|. \tag{89.24}$$

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

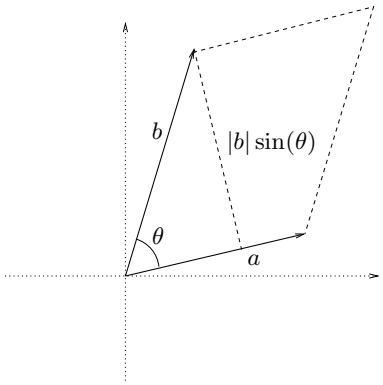


FIGURE 89.22. The vectors a and b span a rectangle with area $|a \times b| = |a|\,|b|\,\sin(\theta)|$.

89.27 Straight Lines

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

$$n_1 x_1 + n_2 x_2 = n \cdot x = 0, \tag{89.25}$$

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. 89.23. 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. 89.23. We state this insight as a theorem because of its importance.

Theorem 89.7 *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

$$n_1 x_1 + n_2 x_2 = d, \tag{89.26}$$

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

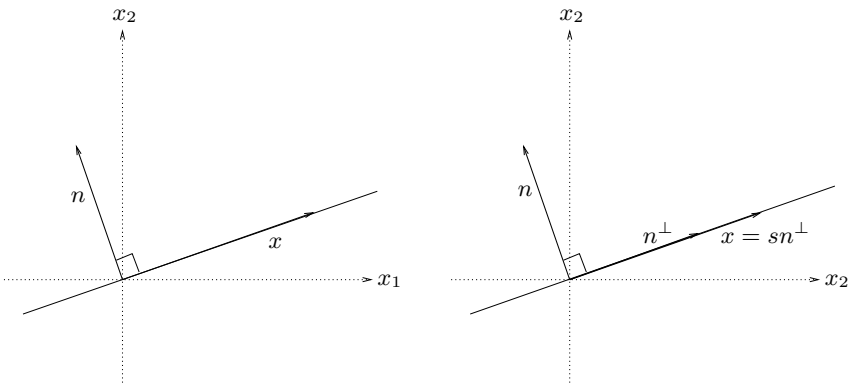


FIGURE 89.23. Vectors $x = sa$ with b orthogonal to a given vector n generate a line through the origin with normal a.

The line in Fig. 89.23 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 89.8 *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.*

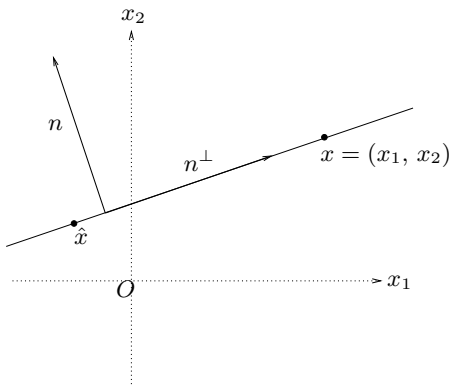


FIGURE 89.24. The line through the point $\hat{x}$ with normal n generated by directional vector a.

EXAMPLE 89.11. The line $x_1 + 2x_2 = 3$ can alternatively be expressed as the set of points $x = (1,1) + s(-2,1)$ with $s \in \mathbb{R}$.

89.28 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. 89.27. This is called the *projection* of the point b onto the line. Equivalently, we seek a point Pb on the line such that $b - Pb$ is orthogonal to the line, that is we seek a point Pb such that

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

We conclude that $Pb = b - \lambda n$ and the equation $n \cdot Pb = d$ thus gives $n \cdot (b - \lambda n) = d$, that is $\lambda = \frac{b \cdot n - d}{|n|^2}$ and so

$$Pb = b - \frac{b \cdot n - d}{|n|^2}\, n. \tag{89.27}$$

If $d = 0$, that is the line $n \cdot x = d = 0$ passes through the origin, then (see Problem 89.26)

$$Pb = b - \frac{b \cdot n}{|n|^2}\, n. \tag{89.28}$$

89.29 When Are Two Lines Parallel?

Let

$$a_{11}x_1 + a_{12}x_2 = b_1,$$
$$a_{21}x_1 + a_{22}x_2 = b_2,$$

be two straight lines in $\mathbb{R}^2$ with normals (a_{11}, a_{12}) and (a_{21}, a_{22}). How do we know if the lines are parallel? Of course, the lines are parallel if and only if their normals are parallel. From above, we know the normals are parallel if and only if

$$(a_{11}, a_{12}) \times (a_{21}, a_{22}) = a_{11}a_{22} - a_{12}a_{21} = 0,$$

and consequently *non*-parallel (and consequently intersecting at some point) if and only if

$$(a_{11}, a_{12}) \times (a_{21}, a_{22}) = a_{11}a_{22} - a_{12}a_{21} \neq 0, \tag{89.29}$$

EXAMPLE 89.12. The two lines $2x_1 + 3x_2 = 1$ and $3x_1 + 4x_2 = 1$ are non-parallel because $2 \cdot 4 - 3 \cdot 3 = 8 - 9 = -1 \neq 0$.

89.30 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*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1, \\ a_{21}x_1 + a_{22}x_2 &= b_2, \end{aligned} \tag{89.30}$$

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 (89.29), $a_{11}a_{22} - a_{12}a_{21} \neq 0$, then the lines must intersect and thus the system (89.30) 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

$$x_1 = (a_{22}b_1 - a_{12}b_2)(a_{11}a_{22} - a_{12}a_{21})^{-1}, \tag{89.31a}$$

$$x_2 = (a_{11}b_2 - a_{21}b_1)(a_{11}a_{22} - a_{12}a_{21})^{-1}. \tag{89.31b}$$

This formula gives the unique solution of (89.30) under the condition $a_{11}a_{22} - a_{12}a_{21} \neq 0$.

We can derive the solution formula (89.31) 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 (89.30) in vector form as

$$x_1 a_1 + x_2 a_2 = b. \tag{89.32}$$

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

$$x_1 a_1 \times a_2 = b \times a_2, \quad x_2 a_2 \times a_1 = b \times a_1.$$

Since $a_1 \times a_2 \neq 0$,

$$x_1 = \frac{b \times a_2}{a_1 \times a_2}, \quad x_2 = \frac{b \times a_1}{a_2 \times a_1} = -\frac{b \times a_1}{a_1 \times a_2}, \tag{89.33}$$

which agrees with the formula (89.31) 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 (89.30) 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 (89.30).

We summarize our experience from this section on systems of 2 linear equations in 2 unknowns as follows:

Theorem 89.9 *The system of linear equations $x_1 a_1 + x_2 a_2 = b$, where a_1, a_2 and b are given vectors in $\mathbb{R}^2$, has a unique solution (x_1, x_2) given by (89.33) 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 89.13. The solution to the system

$$\begin{aligned} x_1 + 2x_2 &= 3, \\ 4x_1 + 5x_2 &= 6, \end{aligned}$$

is given by

$$x_1 = \frac{(3,6) \times (2,5)}{(1,4) \times (2,5)} = \frac{3}{-3} = -1, \quad x_2 = -\frac{(3,6) \times (1,4)}{(1,4) \times (2,5)} = -\frac{6}{-3} = 2.$$

89.31 Linear Independence and Basis

We saw above that the system (89.30) 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 (89.30) 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 (89.33) 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 89.10 *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 89.14. 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 (89.33) we find that $x_1 = 1$ and $x_2 = 2$. The coordinates of b with respect to the basis $\{a_1, a_2\}$ are thus $(1, 2)$, while the coordinates of b with respect to the standard basis are $(5, 4)$.

We next introduce the concept of *linear independence*, which will play an important role below. We say that a set $\{a_1, a_2\}$ of two non-zero vectors a_1 and a_2 two non-zero vectors a_1 and a_2 in $\mathbb{R}^2$ is *linearly independent* if the system of equations

$$x_1 a_1 + x_2 a_2 = 0$$

has the unique solution $x_1 = x_2 = 0$. We just saw that if $a_1 \times a_2 \neq 0$, then a_1 and a_2 are linearly independent (because $b = (0, 0)$ implies $x_1 = x_2 = 0$). Conversely if $a_1 \times a_2 = 0$, then a_1 and a_2 are parallel so that $a_1 = \lambda a_2$ for some $\lambda \neq 0$, and then there are many possible choices of x_1 and x_2, not both equal to zero, such that $x_1 a_1 + x_2 a_2 = 0$, for example $x_1 = -1$ and $x_2 = \lambda$. We have thus proved:

Theorem 89.11 *The set $\{a_1, a_2\}$ of non-zero vectors a_1 and a_2 is linearly independent if and only if $a_1 \times a_2 \neq 0$.*

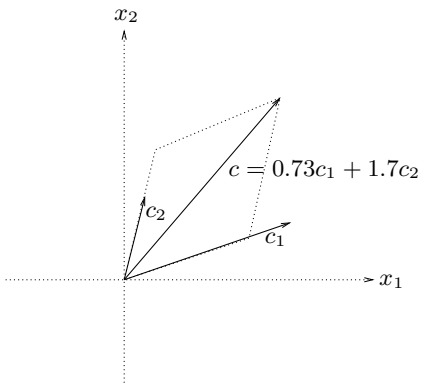


FIGURE 89.25. Linear combination c of two linearly independent vectors c_1 and c_2

89.32 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.

89.33 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 $\mathbb{R}$,

$$f(x + y) = f(x) + f(y) \quad \text{and} \quad f(\lambda x) = \lambda f(x). \tag{89.34}$$

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

89.34 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

$$f_1(x) = a_{11}x_1 + a_{12}x_2, \tag{89.35a}$$
$$f_2(x) = a_{21}x_1 + a_{22}x_2, \tag{89.35b}$$

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. 89.26.

EXAMPLE 89.16. The function $f(x_1, x_2) = (x_1 + 3x_2, 2x_1 - x_3)$ defines a linear mapping $\mathbb{R}^2 \to \mathbb{R}^2$.

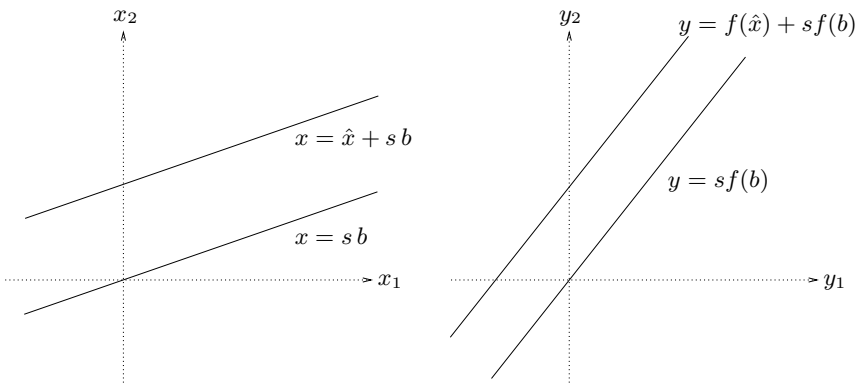


FIGURE 89.26. A linear mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$ maps (parallel) lines to (parallel) lines, and consequently parallelograms to parallelograms.

89.35 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 (89.35), we seek $x \in \mathbb{R}^2$ satisfying the 2×2 linear system of equations

$$a_{11}x_1 + a_{12}x_2 = b_1, \tag{89.36a}$$
$$a_{21}x_1 + a_{22}x_2 = b_2, \tag{89.36b}$$

where the coefficients a_{ij} and the coordinates b_i of the right hand side are given.

89.36 A First Encounter with Matrices

We write the left hand side of (89.36) as follows:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{pmatrix}. \tag{89.37}$$

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 (89.37) defines the *product* of a 2×2 matrix and a 2×1 matrix or a 2-column vector. The product can be interpreted as

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} c_1 \cdot x \\ c_2 \cdot x \end{pmatrix} \tag{89.38}$$

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

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{89.39}$$

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

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{and } x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{and } b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \tag{89.40}$$

we can phrase the system of equations (89.36) in condensed form as the following *matrix equation:*

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

We have now got a first glimpse of matrices including the basic operation of multiplication of a 2×2-matrix with a 2×1 matrix or 2-column vector. Below we will generalize to a calculus for matrices including addition of matrices, multiplication of matrices with a real number, and multiplication of matrices. We will also discover a form of matrix division referred to as inversion of matrices allowing us to express the solution of the system $Ax = b$ as $x = A^{-1}b$, under the condition that the columns (or equivalently, the rows) of A are linearly independent.

89.37 First Applications of Matrix Notation

To show the usefulness of the matrix notation just introduced, we rewrite some of the linear systems of equations and transformations which we have met above.

Rotation by θ

The mapping $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ corresponding to rotation of a vector by an angle θ is given by (89.14), that is

$$R_\theta(x) = (x_1\cos(\theta) - x_2\sin(\theta), x_1\sin(\theta) + x_2\cos(\theta)). \tag{89.41}$$

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

$$A = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}. \tag{89.42}$$

Projection Onto a Vector a

The projection $P_a(x) = \frac{x\cdot a}{|a|^2}\, a$ given by (89.9) 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

$$A = \begin{pmatrix} \frac{a_1^2}{|a|^2} & \frac{a_1 a_2}{|a|^2} \\ \frac{a_1 a_2}{|a|^2} & \frac{a_2^2}{|a|^2} \end{pmatrix}. \tag{89.43}$$

Change of Basis

The linear system (89.17) 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.

89.38 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:

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

89.39 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:

$$\lambda A = \lambda \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} \\ \lambda a_{21} & \lambda a_{22} \end{pmatrix} = C.$$

89.40 Multiplication of Two Matrices

Given two 2×2 matrices $A = (a_{ij})$ and $B = (b_{ij})$ with elements, we define the product $C = AB$ as the matrix with elements c_{ij} given by

$$c_{ij} = \sum_{k=1}^{2} a_{ik} b_{kj}.$$

Writing out the sum, we have

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

In other words, to get the element c_{ij} of the product $C = AB$, we take the scalar product of row i of A with column j of B.

The matrix product is generally non-commutative so that $AB \neq BA$ most of the time.

We say that in the product AB the matrix A multiplies the matrix B from the left and that B multiplies the matrix A from the right. Non-commutativity of matrix multiplication means that multiplication from right or left may give different results.

EXAMPLE 89.17. 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 89.18. We compute $BB = B^2$, where B is the projection matrix given by (89.43), 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 89.19. As another application we compute the product of two matrices corresponding to two rotations with angles α and β:

$$A = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} \cos(\beta) & -\sin(\beta) \\ \sin(\beta) & \cos(\beta) \end{pmatrix}. \quad (89.44)$$

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 $\alpha+\beta$.

89.41 The Transpose of a Matrix

Given a 2×2 matrix A with elements a_{ij}, we define the *transpose* of A denoted by $A^\top$ as the matrix $C = A^\top$ with elements $c_{11} = a_{11}$, $c_{12} = a_{21}$, $c_{21} = a_{12}$, $c_{22} = a_{22}$. In other words, the rows of A are the columns of $A^\top$ and vice versa. For example

$$\text{if} \quad A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad \text{then} \quad A^\top = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}.$$

Of course $(A^\top)^\top = A$. Transposing twice brings back the original matrix. We can directly check the validity of the following rules for computing with the transpose:

$$(A + B)^\top = A^\top + B^\top, \quad (\lambda A)^\top = \lambda A^\top,$$
$$(AB)^\top = B^\top A^\top.$$

89.42 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.

89.43 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.

89.44 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

$$A^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}. \tag{89.45}$$

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 (89.31) with $b = (1, 0)$ and the second column choosing $b = (0, 1)$.

The solution to the system of equations $Ax = b$ can be written as $x = A^{-1}b$, which we obtain by multiplying $Ax = b$ from the left by A^{-1}.

We can directly check the validity of the following rules for computing with the inverse:

$$(\lambda A)^{-1} = \lambda A^{-1}$$
$$(AB)^{-1} = B^{-1}A^{-1}.$$

89.45 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:

$$R_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \tag{89.46}$$

We have also seen that two successive rotations by angles α and β can be written as

$$y = R_\beta R_\alpha x, \tag{89.47}$$

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 (89.45),

$$R_\theta^{-1} = \frac{1}{\cos(\theta)^2 + \sin(\theta)^2} \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} = \begin{pmatrix} \cos(-\theta) & -\sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{pmatrix} \tag{89.48}$$

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

$$R_\theta R_\theta^\top = I. \tag{89.49}$$

We use this fact to prove that the length of a vector is not changed by rotation. If $y = R_\theta x$, then

$$|y|^2 = y^T y = (R_\theta x)^\top (R_\theta x) = x^\top R_\theta^\top R_\theta x = x^\top x = |x|^2. \tag{89.50}$$

More generally, the scalar product is preserved after the rotation. If $y = R_\theta x$ and $\hat{y} = R_\theta \hat{x}$, then

$$y \cdot \hat{y} = (R_\theta x)^\top (R_\theta \hat{x}) = x^\top R_\theta^\top R_\theta \hat{x} = x \cdot \hat{x}. \tag{89.51}$$

The relation (89.49) says that the matrix R_θ is *orthogonal*. Orthogonal matrices play an important role, and we will return to this topic below.

89.46 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

$$2P - I = \frac{2}{|a|^2} \begin{pmatrix} a_1^2 - 1 & a_1 a_2 \\ a_2 a_1 & a_2^2 - 1 \end{pmatrix}.$$

After some reflection(!), looking at Fig. 89.27, 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.

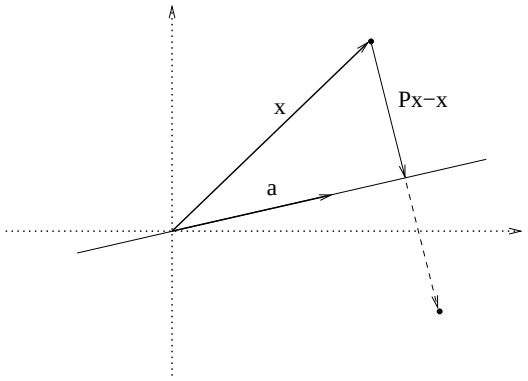


FIGURE 89.27. The mapping $2P - I$ maps points to its mirror point relative to the given line.

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

$$y \cdot \hat{y} = ((2P - I)x)^\top (2P - I)\hat{x} = x^\top (2P^\top - I)(2P - I)\hat{x} = \tag{89.52}$$

$$x^\top (4P^\top P - 2P^\top I - 2PI + I)\hat{x} = x^\top (4P - 4P + I)\hat{x} = x \cdot \hat{x}, \tag{89.53}$$

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

89.47 Change of Basis Again!

Let $\{a_1, a_2\}$ and $\{\hat{a}_1, \hat{a}_2\}$ be two different bases in $\mathbb{R}^2$. We then express any given $b \in \mathbb{R}^2$ as

$$b = x_1 a_1 + x_2 a_2 = \hat{x}_1 \hat{a}_1 + \hat{x}_2 \hat{a}_2, \tag{89.54}$$

with certain coordinates (x_1, x_2) with respect to $\{a_1, a_2\}$ and some other coordinates $(\hat{x}_1, \hat{x}_2)$ with respect $\{\hat{a}_1, \hat{a}_2\}$.

To connect (x_1, x_2) to $(\hat{x}_1, \hat{x}_2)$, we express the basis vectors $\{\hat{a}_1, \hat{a}_2\}$ in terms of the basis $\{a_1, a_2\}$:

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

with certain coefficients c_{ij}. Inserting this into (89.54), we get

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

Reordering terms,

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

We conclude by uniqueness that

$$x_1 = c_{11}\hat{x}_1 + c_{12}\hat{x}_2, \tag{89.55}$$

$$x_2 = c_{21}\hat{x}_1 + c_{22}\hat{x}_2, \tag{89.56}$$

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

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

89.48 Queen Christina

Queen Christina of Sweden (1626-1689), daughter of Gustaf Vasa King of Sweden 1611-1632, crowned to Queen at the age 5, officially coronated 1644, abdicated 1652, converted to Catholicism and moved to Rome 1655.

Throughout her life, Christina had a passion for the arts and for learning, and surrounded herself with musicians, writers, artists and also philosophers, theologians, scientists and mathematicians. Christina had an impressive collection of sculpture and paintings, and was highly respected for

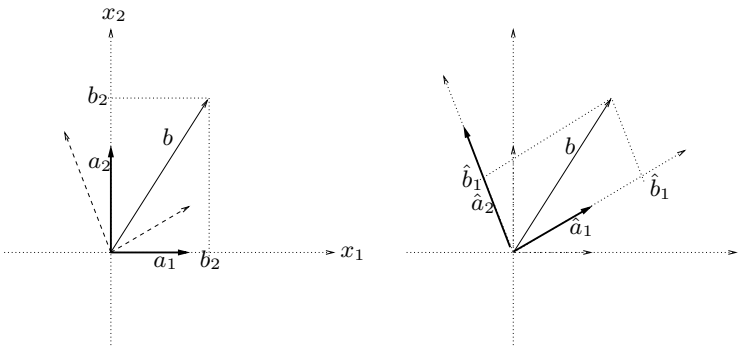


FIGURE 89.28. A vector b may be expressed in terms of the basis $\{a_1, a_2\}$ or the basis $\{\hat{a}_1, \hat{a}_2\}$.

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 89 Problems

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

$$a+b=b+a, \quad (a+b)+c=a+(b+c), \quad a+(-a)=0$$
$$a+0=a, \quad 3a=a+a+a, \quad \lambda(\mu a)=(\lambda\mu)a,$$
$$(\lambda+\mu)a=\lambda a+\mu a, \quad \lambda(a+b)=\lambda a+\lambda b, \quad |\lambda a|=|\lambda||a|.$$

Try to give both analytical and geometrical proofs.

89.2. 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.

89.3. Given $a = (3, 2)$ and $b = (1, 4)$, compute (i) $|a|$, (ii) $|b|$, (iii) $|a + b|$, (iv)) $|a - b|$, (v) $a/|a|$, (vi) $b/|b|$.

FIGURE 89.29. 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".

89.4. Show that the norm of $a/|a|$ with $a \in \mathbb{R}^2$, $a \neq 0$, is equal to 1.

89.5. Given $a, b \in \mathbb{R}^2$ prove the following inequalities a) $|a+b| \leq |a| + |b|$, b) $a \cdot b \leq |a||b|$.

89.6. Compute $a \cdot b$ with

(i) $a = (1,2), b = (3,2)$ (ii) $a = (10,27), b = (14,-5)$

89.7. 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) $|a \cdot b|$.

89.8. What is the angle between $a = (1,1)$ and $b = (3,7)$?

89.9. 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.

89.10. Find the projection of a onto b onto $(1,2)$ with (i) $a = (1,2)$, (ii) $a = (-2,1)$, (iii) $a = (2,2)$, (iv) $a = (\sqrt{2}, \sqrt{2})$.

89.11. 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.

89.12. 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.

89.13. 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.

89.14. 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) $x^T = (1,1)$

89.15. 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) $A^{-1}A$.

89.16. Show that $(AB)^T = B^T A^T$ and that $(Ax)^T = x^T A^T$.

89.17. What can be said about A if: a) $A = A^T$ b) $AB = I$?

89.18. 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 $P = P^T$.

89.19. 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.

89.20. Express the linear transformation of rotating a vector a certain given angle as a matrix vector product.

89.21. 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

$$c \cdot d = \bar{c} \cdot \bar{d}.$$

89.22. Compute $a \times b$ and $b \times a$ with (i) $a = (1, 2), b = (3, 2)$, (ii) $a = (1, 2), b = (3, 6)$, (iii) $a = (2, -1), b = (2, 4)$.

89.23. 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.

89.24. Check the answers to the above problems using Matlab.

89.25. Verify that the projection $Px = P_a(x)$ is linear in x. Is it linear also in a? Illustrate, as in Fig. 89.16, that $P_a(x + y) = P_a(x) + P_a(y)$.

89.26. Prove that the formula (90.29) for the projection of a point onto a line through the origin, coincides with the formula (89.9) for the projection of the vector b on the direction of the line.

89.27. 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.

90
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)

90.1 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. 90.1.

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

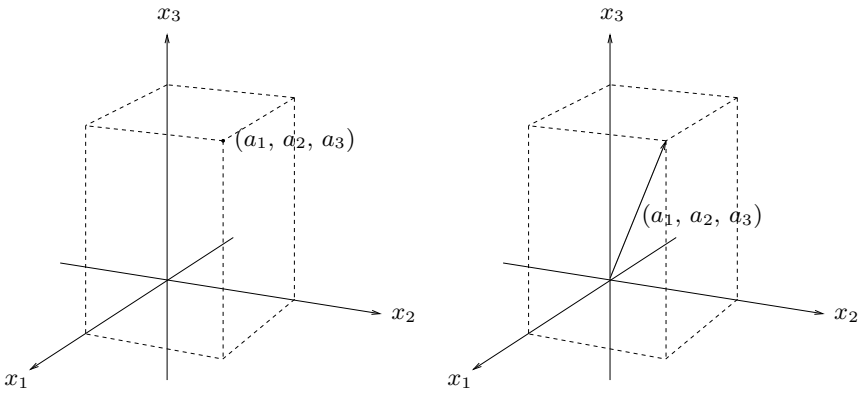


FIGURE 90.1. Coordinate system for $\mathbb{R}^3$

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. 90.2. 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.

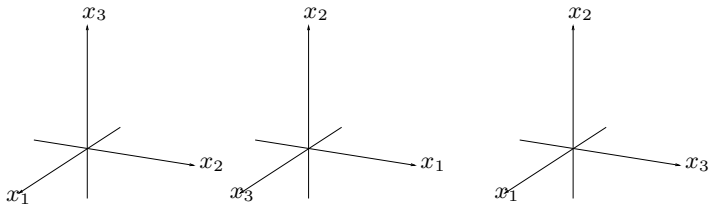


FIGURE 90.2. Two "right" coordinate systems and one "left", where the vertical coordinate of the view point is assumed positive, that is, the horizontal plane is seen from above. What happens if the vertical coordinate of the view point is assumed negative?

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. 90.1. 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.

90.2 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. 90.1.

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 $e_3 = (0, 0, 1)$.

90.3 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

$$a \cdot b = \sum_{i=1}^{3} a_i b_i = a_1 b_1 + a_2 b_2 + a_3 b_3. \tag{90.1}$$

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

$$|a| = (a \cdot a)^{\frac{1}{2}} = (\sum_{i=1}^{3} a_i^2)^{\frac{1}{2}}, \tag{90.2}$$

which expresses Pythagoras theorem in 3d, and which we may obtain by using the usual 2d Pythagoras theorem twice. The distance $|a-b|$ between two points $a=(a_1,a_2,a_3)$ and $b=(b_1,b_2,b_3)$ is equal to

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

Cauchy's inequality states that for any vectors a and b in $\mathbb{R}^3$,

$$|a\cdot b| \le |a|\,|b|. \tag{90.3}$$

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.

90.4 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

$$Pb = P_a(b) = \frac{a\cdot b}{a\cdot a}\,a = \frac{a\cdot b}{|a|^2}\,a. \tag{90.4}$$

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. 90.3, 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 (90.4) with $\lambda = \frac{a\cdot b}{|a|^2}$.

The transformation $P:\mathbb{R}^3\to\mathbb{R}^3$ is linear, that is for any b and $c\in\mathbb{R}^3$ and $\lambda\in\mathbb{R}$,

$$P(b+c) = Pb + Pc, \quad P(\lambda b) = \lambda Pb,$$

and $PP = P$.

90.5 The Angle Between Two Vectors

We define the angle θ between non-zero vectors a and b in $\mathbb{R}^3$ by

$$\cos(\theta) = \frac{a\cdot b}{|a|\,|b|}, \tag{90.5}$$

where we may assume that $0 \le \theta \le 180°$. By Cauchy's inequality (90.3), $|a \cdot b| \le |a|\,|b|$. Thus, there is an angle θ satisfying (90.5) that is uniquely defined if we require $0 \le \theta \le 180°$. We may write (90.5) in the form

$$a \cdot b = |a||b|\cos(\theta), \tag{90.6}$$

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 (90.5), 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 (90.5) of angle between two vectors and the usual geometric definition coincide. In particular, two non-zero vectors are geometrically orthogonal in the sense that the geometric angle θ between the vectors satisfies $\cos(\theta) = 0$ if and only if $a \cdot b = |a||b|\cos(\theta) = 0$.

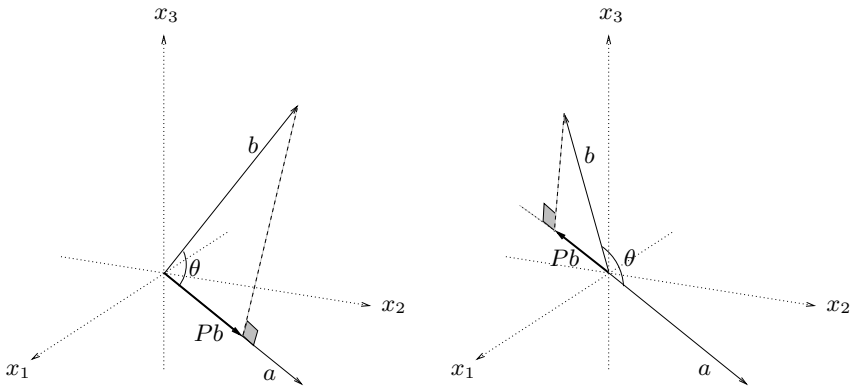


FIGURE 90.3. Projection Pb of a vector b onto a vector a.

90.6 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

$$a \times b = (a_2b_3 - a_3b_2, a_3b_1 - a_1b_3, a_1b_2 - a_2b_1). \tag{90.7}$$

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 90.1. 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 (90.7) 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

$$a \times (b + c) = a \times b + a \times c, \quad (a + b) \times c = a \times c + b \times c, \tag{90.8a}$$

$$(\lambda a) \times b = \lambda\, a \times b, \quad a \times (\lambda b) = \lambda\, a \times b, \tag{90.8b}$$

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

$$a \times b = -\, b \times a. \tag{90.9}$$

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!)

$$e_i \times e_i = 0,\ i = 1,2,3, \tag{90.10a}$$

$$e_1 \times e_2 = e_3, \quad e_2 \times e_3 = e_1, \quad e_3 \times e_1 = e_2, \tag{90.10b}$$

$$e_2 \times e_1 = -e_3, \quad e_3 \times e_2 = -e_1, \quad e_1 \times e_3 = -e_2. \tag{90.10c}$$

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

$$a \cdot (a \times b) = a_1(a_2b_3 - a_3b_2) + a_2(a_3b_1 - a_1b_3) + a_3(a_1b_2 - a_2b_1) = 0, \tag{90.11}$$

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 (90.10) as follows,

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

which conforms with (90.7).

90.7 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 (89.21) for the vector product in $\mathbb{R}^2$, we have

$$|a \times b| = |a||b||\sin(\theta)|, \tag{90.12}$$

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 (90.6), we obtain

$$|a|^2 |b|^2 \sin^2(\theta) = |a|^2 |b|^2 - (a \cdot b)^2. \tag{90.13}$$

Finally, a direct (but somewhat lengthy) computation shows that

$$|a \times b|^2 = |a|^2 |b|^2 - (a \cdot b)^2,$$

which proves (90.12). We summarize in the following theorem.

Theorem 90.1 *The vector product $a \times b$ of two non-zero vectors a and b in $\mathbb{R}^3$ is orthogonal to both a and b and $|a \times b| = |a||b||\sin(\theta)|$, where θ is the angle between a and b. In particular, a and b are parallel if and only if $a \times b = 0$.*

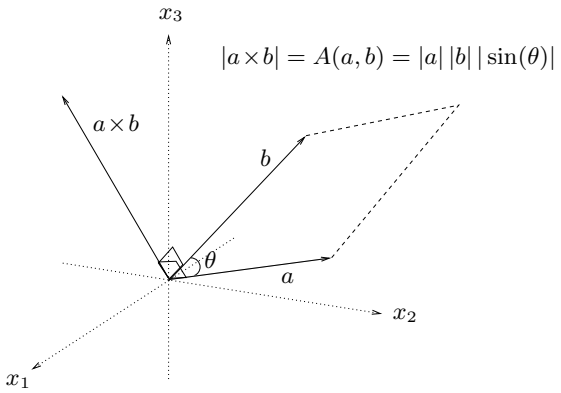


FIGURE 90.4. Geometric interpretation of the vector product.

We can make the theorem more precise by adding the following sign rule: The vector $a \times b$ is pointing in the direction of a standard screw turning the vector a into the vector b the shortest route.

90.8 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

$$a \times b = (0, 0, a_1b_2 - a_2b_1). \tag{90.14}$$

The previous formula $a \times b = a_1b_2 - a_2b_1$ 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 $a \times b = (0, 0, a_1b_2 - a_2b_1)$ for $a = (a_1, a_2, 0)$ and $b = (b_1, b_2, 0)$, with $a_1b_2 - a_2b_1$ 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 $a_1b_2 - a_2b_1 \geq 0$, 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 $\sin(\theta) \geq 0$.

90.9 Volume of a Parallelepiped Spanned by Three Vectors

Consider the parallelepiped spanned by three vectors a, b and c, according to Fig. 90.5.

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*

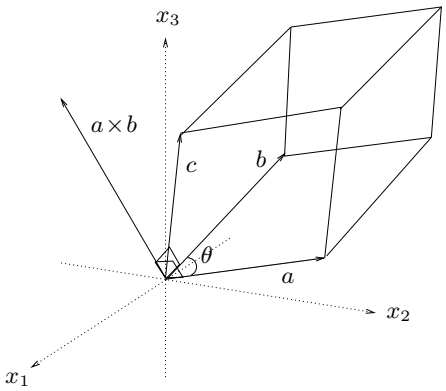


FIGURE 90.5. Parallelepiped spanned by three vectors

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 (90.12) and (90.4), we know that

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

and thus

$$V(a,b,c) = |c \cdot (a \times b)|. \tag{90.15}$$

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,

$$V(a,b,c) = |a \cdot (b \times c)| = |b \cdot (a \times c)| = |c \cdot (a \times b)|. \tag{90.16}$$

EXAMPLE 90.2. 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 $a \cdot (b \times c) = (1,2,3) \cdot (1,-5,7) = 12$.

90.10 The Triple Product $a \cdot b \times c$

The expression $a \cdot (b \times c)$ occurs in the formulas (90.15) and (90.16). This is called the *triple product* of the three vectors a, b and c. We usually write the triple product without the parenthesis following the convention that the vector product $\times$ is performed first. In fact, the alternative interpretation $(a \cdot b) \times c$ does not make sense since $a \cdot b$ is a scalar and the vector product $\times$ requires vector factors!

The following properties of the triple product can be readily verified by direct application of the definition of the scalar and vector products,

$$a \cdot b \times c = c \cdot a \times b = b \cdot c \times a,$$
$$a \cdot b \times c = -a \cdot c \times b = -b \cdot a \times c = -c \cdot b \times a.$$

To remember these formulas, we note that if two of the vectors change place then the sign changes, while if all three vectors are cyclically permuted (for example the order a, b, c is replaced by c, a, b or b, c, a), then the sign is unchanged.

Using the triple product $a \cdot b \times c$, we can express the geometric quantity of the volume $V(a, b, c)$ of the parallelepiped spanned by a, b and c in the concise algebraic form,

$$V(a, b, c) = |a \cdot b \times c|. \tag{90.17}$$

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.

90.11 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 (90.17),

$$\begin{aligned} \pm V(a_1, a_2, a_3) &= a_1 \cdot a_2 \times a_3 \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) \\ &\quad + a_{13}(a_{21}a_{32} - a_{22}a_{31}). \end{aligned} \tag{90.18}$$

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

$$a_{1j_1}a_{2j_2}a_{3j_3} \tag{90.19}$$

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 (90.18) states that the signed volume $\pm V(a_1,a_2,a_3)$ is equal to the determinant of the 3×3 matrix $A=(a_{ij})$.

90.12 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. 90.6. If $\hat{x}=0$, then the line passes through the origin.

Example 90.3. The line through $(1,2,3)$ in the direction $(4,5,6)$ is given by

$$x=(1,2,3)+s(4,5,6)=(1+4s,2+5s,3+6s)\quad s\in\mathbb{R}.$$

The line through $(1,2,3)$ and $(3,1,2)$ has the direction $(3,1,2)-(1,2,3)$ $=(2,-1,-1)$, and is thus given by $x=(1,2,3)+s(2,-1,-1)$.

Note that by choosing other vectors to represent the direction of the line, it may also be represented, for example, as $x=(1,2,3)+\hat{s}(-2,1,1)$ or $x=(1,2,3)+\tilde{s}(6,-3,-3)$. Also, the "point of departure" on the line, corresponding to $s=0$, can be chosen arbitrarily on the line of course. For example, the point $(1,2,3)$ could be replaced by $(-1,3,4)$ which is another point on the line.

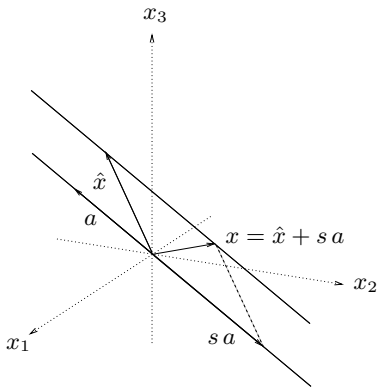


FIGURE 90.6. Line in $\mathbb{R}^3$ of the form $x = \hat{x} + s\,a$.

90.13 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

$$Pb = \hat{x} + \frac{b \cdot a - \hat{x} \cdot a}{|a|^2} a\,. \tag{90.20}$$

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 (89.9) in $\mathbb{R}^2$.

90.14 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

$$x = s_1 a_1 + s_2 a_2, \tag{90.21}$$

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. 90.7. 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

$$n \cdot x = 0. \tag{90.22}$$

We may thus describe the points x in the plane by the representation (90.21) or the equation (90.22). Note that (90.21) is a vector equation corresponding to 3 scalar equations, while (90.22) is a scalar equation. Eliminating the parameters s_1 and s_2 in the system (90.21), we obtain the scalar equation (90.22).

Let $\hat{x}$ be a given point in $\mathbb{R}^3$. The points x in $\mathbb{R}^3$ that can be expressed as

$$x = \hat{x} + s_1 a_1 + s_2 a_2, \tag{90.23}$$

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. 90.8.

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

$$n \cdot x = n \cdot \hat{x}. \tag{90.24}$$

Again, we obtain the scalar equation (90.24) if we eliminate the parameters s_1 and s_2 in the system (90.23).

We summarize:

Theorem 90.2 *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 90.4. Consider the plane $x_1 + 2x_2 + 3x_3 = 4$, that is the plane $(1, 2, 3) \cdot (x_1, x_2, x_3) = 4$ with normal $n = (1, 2, 3)$. To write the points x in this plane on the form $x = \hat{x} + s_1 a_1 + s_2 a_2$, we first choose a point $\hat{x}$ in the plane, for example, $\hat{x} = (2, 1, 0)$ noting that $n \cdot \hat{x} = 4$. We next choose two non-parallel vectors a_1 and a_2 such that $n \cdot a_1 = 0$ and $n \cdot a_2 = 0$, for example $a_1 = (-2, 1, 0)$ and $a_2 = (-3, 0, 1)$. Alternatively, we choose one vector a_1 satisfying $n \cdot a_1 = 0$ and set $a_2 = n \times a_1$, which is a vector orthogonal to both n and a_1. To find a vector a_1 satisfying $a_1 \cdot n = 0$, we may choose an arbitrary non-zero vector m non-parallel to n and set $a_1 = m \times n$, for example $m = (0, 0, 1)$ giving $a_1 = (-2, 1, 0)$.

Conversely, given the plane $x = (2, 1, 0) + s_1(-2, 1, 0) + s_2(-3, 0, 1)$, that is $x = \hat{x} + s_1 a_1 + s_2 a_2$ with $\hat{x} = (2, 1, 0)$, $a_1 = (-2, 1, 0)$ and $a_2 = (-3, 0, 1)$, we obtain the equation $x_1 + 2x_2 + 3x_3 = 4$ simply by computing $n = a_1 \times a_2 = (1, 2, 3)$ and $n \cdot \hat{x} = (1, 2, 3) \cdot (2, 1, 0) = 4$, from which we obtain the following equation for the plane: $n \cdot x = (1, 2, 3) \cdot (x_1, x_2, x_3) = x_1 + 2x_2 + 3x_3 = n \cdot \hat{x} = 4$.

EXAMPLE 90.5. Consider the real-valued function $z = f(x, y) = ax + by + c$ of two real variables x and y, where a, b and c are real numbers.

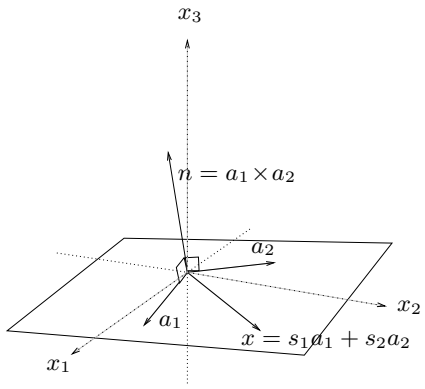


FIGURE 90.7. Plane through the origin spanned by a_1 and a_2, and with normal $n = a_1 \times a_2$.

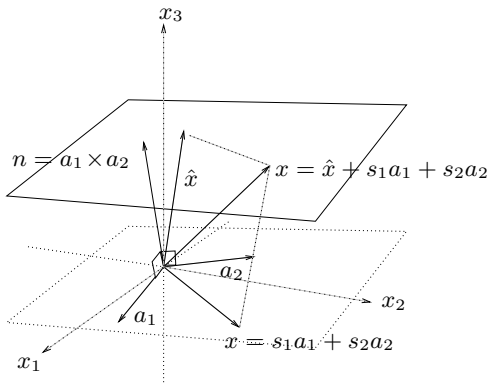


FIGURE 90.8. Plane through $\hat{x}$ with normal n defined by $n \cdot x = d = n \cdot \hat{x}$.

Setting $x_1 = x$, $x_2 = y$ and $x_3 = z$, we can express the graph of $z = f(x, y)$ as the plane $ax_1 + bx_2 - x_3 = -c$ in $\mathbb{R}^3$ with normal $(a, b, -1)$.

90.15 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

$$x = \hat{x} + (d - n \cdot \hat{x})/(n \cdot a)\, a. \tag{90.25}$$

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 90.6. 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.

90.16 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

$$\begin{aligned} n_1 \cdot x &= d_1, \\ n_2 \cdot x &= d_2. \end{aligned} \tag{90.26}$$

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 (90.26) as

$$x = \hat{x} + s\, n_1 \times n_2, \quad s \in \mathbb{R}. \tag{90.27}$$

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 90.3 *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 90.7. 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 $\hat{x} = (0, 1, 1)$.

90.17 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. 90.9,

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

We conclude that $Pb = b - \lambda n$ and the equation $n \cdot Pb = d$ thus gives $n \cdot (b - \lambda n) = d$. So, $\lambda = \frac{b \cdot n - d}{|n|^2}$ and thus

$$Pb = b - \frac{b \cdot n - d}{|n|^2}\, n. \tag{90.28}$$

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

$$Pb = b - \frac{b \cdot n}{|n|^2}\, n. \tag{90.29}$$

If the plane is given in the form $x = \hat{x} + s_1 a_1 + s_2 a_2$ with a_1 and a_2 two given non-parallel vectors in $\mathbb{R}^3$, then we may alternatively compute the projection Pb of a point b onto the plane by seeking real numbers x_1 and x_2 so that $Pb = \hat{x} + x_1 a_1 + x_2 a_2$ and $(b - Pb) \cdot a_1 = (b - Pb) \cdot a_2 = 0$. This gives the system of equations

$$\begin{aligned} x_1 a_1 \cdot a_1 + x_2 a_2 \cdot a_1 &= b \cdot a_1 - \hat{x} \cdot a_1, \\ x_1 a_1 \cdot a_2 + x_2 a_2 \cdot a_2 &= b \cdot a_2 - \hat{x} \cdot a_2 \end{aligned} \tag{90.30}$$

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 90.24.

EXAMPLE 90.8. 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 90.9. 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 $Pb = (2, 2, 3) - \frac{(2,2,3)\cdot(1,-2,1)}{6}(1, -2, 1) = (2, 2, 3) - \frac{1}{6}(1, -2, 1) = \frac{1}{6}(11, 14, 17)$.

90.18 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. 90.9 and Problem 90.22.

EXAMPLE 90.10. The distance from the point $(2, 2, 3)$ to the plane $x_1 + x_2 + x_3 = 1$ is equal to $\frac{|(2,2,3)\cdot(1,1,1)-1|}{\sqrt{3}} = 2\sqrt{3}$.

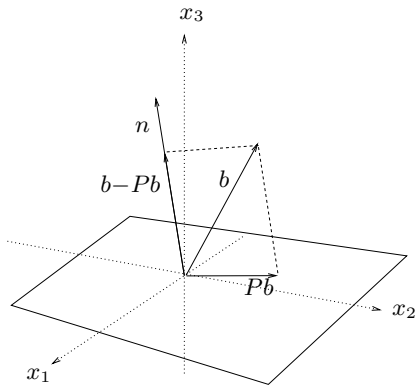


FIGURE 90.9. Projection of a point/vector onto a plane.

90.19 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 89.21, 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. 90.10. We write

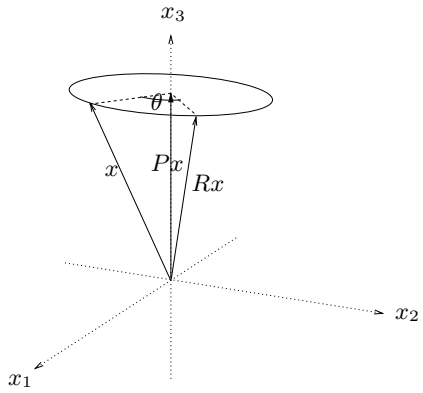


FIGURE 90.10. Rotation around $a = (0, 0, 1)$ a given angle θ.

$Rx - Px$ as $Rx - Px = \alpha(x - Px) + \beta\, a \times (x - Px)$ for real numbers α and β, noting that $Rx - Px$ is orthogonal to a and $a \times (x - Px)$ is orthogonal to both a and $(x - Px)$. Taking the scalar product with $(x - Px)$, we use (ii) to get $\alpha = \cos(\theta)$ and then use (i) to find $\beta = \frac{\sin(\theta)}{|a|}$ with a suitable

orientation. Thus, we may express Rx in terms of the projection Px as

$$Rx = Px + \cos(\theta)(x - Px) + \frac{\sin(\theta)}{|a|}a \times (x - Px). \tag{90.31}$$

90.20 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.

90.21 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):

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3, \end{aligned} \tag{90.32}$$

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 $\mathbb{R}^3$:

$$x_1a_1 + x_2a_2 + x_3a_3 = b, \tag{90.33}$$

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 (90.32) 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

$$V(a_1, a_2, a_3) = |a_1 \cdot a_2 \times a_3| \neq 0, \tag{90.34}$$

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 (90.32). We can do this by mimicking what we did in the case of a 2×2 system: Taking the scalar product of both sides of the vector equation $x_1 a_1 + x_2 a_2 + x_3 a_3 = b$ by successively $a_2 \times a_3$, $a_3 \times a_1$, and $a_2 \times a_3$, we get the following solution formula (recalling that $a_1 \cdot a_2 \times a_3 = a_2 \cdot a_3 \times a_1 = a_3 \cdot a_1 \times a_2$):

$$\begin{aligned} x_1 &= \frac{b \cdot a_2 \times a_3}{a_1 \cdot a_2 \times a_3}, \\ x_2 &= \frac{b \cdot a_3 \times a_1}{a_2 \cdot a_3 \times a_1} = \frac{a_1 \cdot b \times a_3}{a_1 \cdot a_2 \times a_3}, \\ x_3 &= \frac{b \cdot a_1 \times a_2}{a_3 \cdot a_1 \times a_2} = \frac{a_1 \cdot a_2 \times b}{a_1 \cdot a_2 \times a_3}, \end{aligned} \tag{90.35}$$

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 (90.35) shows that the system (90.32) has a unique solution if $a_1 \cdot a_2 \times a_3 \neq 0$.

Note the pattern of the solution formula (90.35), 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 (90.35) is also called *Cramer's rule*. We have proved the following basic result:

Theorem 90.4 *If $a_1 \cdot a_2 \times a_3 \neq 0$, then the system of equations (90.32) or the equivalent vector-equation (90.33) has a unique solution given by Cramer's rule (90.35).*

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 $\{a_1, a_2, a_3\}$ is *linearly independent*, or that the three vectors a_1, a_2 and a_3 are *linearly independent*.

90.22 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 (90.32), 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 (90.32) in the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1, \\ (a_{22}a_{11} - a_{21}a_{12})x_2 + (a_{23}a_{11} - a_{21}a_{13})x_3 &= a_{11}b_2 - a_{21}b_1, \\ (a_{32}a_{11} - a_{31}a_{12})x_2 + (a_{33}a_{11} - a_{31}a_{13})x_3 &= a_{11}b_3 - a_{31}b_1, \end{aligned} \tag{90.36}$$

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

$$\begin{aligned} a_{11}x_1 + \quad a_{12}x_2 + a_{13}x_3 &= b_1, \\ \hat{a}_{22}x_2 + \hat{a}_{23}x_3 &= \hat{b}_2, \\ \hat{a}_{32}x_2 + \hat{a}_{33}x_3 &= \hat{b}_3, \end{aligned} \tag{90.37}$$

with modified coefficients $\hat{a}_{ij}$ and $\hat{b}_i$. We now proceed in the same way considering the 2×2-system in (x_2, x_3), and bring the system to the final triangular form

$$\begin{aligned} a_{11}x_1 + \quad a_{12}x_2 + a_{13}x_3 &= b_1, \\ \hat{a}_{22}x_2 + \hat{a}_{23}x_3 &= \hat{b}_2, \\ \tilde{a}_{33}x_3 &= \tilde{b}_3, \end{aligned} \tag{90.38}$$

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 90.11. We give an example of Gaussian elimination: Consider the system

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

Subtracting the first equation multiplied by 2 from the second and the first equation multiplied by 3 from the third equation, we get the system

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

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

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

from which we find $x_3 = 1$ and then from the second equation $x_2 = 1$ and finally from the first equation $x_1 = 1$.

90.23 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

$$c_{ij} = \sum_{k=1}^{3} a_{ik} b_{kj} \quad i, j = 1, 2, 3. \tag{90.39}$$

Matrix multiplication is *associative* so that $(AB)C = A(BC)$ for matrices A, B and C, see Problem 90.10. The matrix product is however not *commutative* in general, that is there are matrices A and B such that $AB \neq BA$, see Problem 90.11.

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 (90.39) 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 $b = (b_i)$.

90.24 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:

$$x_1 a_1 + x_2 a_2 + x_3 a_3 = b, \tag{90.40}$$

or as a system of 3 scalar equations:

$$\begin{aligned} \hat{a}_1 \cdot x &= b_1 \\ \hat{a}_2 \cdot x &= b_2 \\ \hat{a}_3 \cdot x &= b_3, \end{aligned} \tag{90.41}$$

where the rows $\hat{a}_i$ may be interpreted as normals to planes. We know from the discussion following (90.34) that (90.40) 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 90.3 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 (90.41) 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 (90.41) at one point. Thus, if $\hat{a}_1 \cdot \hat{a}_2 \times \hat{a}_3 \neq 0$ then (90.41) 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 (90.18) gives the more precise result,

Theorem 90.5 *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 $V(a_1, a_2, a_3) = V(\hat{a}_1, \hat{a}_2, \hat{a}_3)$.*

90.25 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 90.21, 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 90.6 *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.*

90.26 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

$$AA^{-1} = I, \tag{90.42}$$

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 (90.42) $AA^{-1}x = Ix = x \neq 0$. 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 90.7 *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, $(AB)^{-1}=B^{-1}A^{-1}$.*

90.27 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 90.4 implies that any given $b\in\mathbb{R}^3$ can be uniquely expressed as a linear combination of $\{a_1,a_2,a_3\}$,

$$b=x_1a_1+x_2a_2+x_3a_3, \tag{90.43}$$

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 $x=A^{-1}b$.

90.28 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 90.8 *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 $x=0$.*

90.29 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 90.12. The matrix

$$Q = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{90.44}$$

is orthogonal and corresponds to rotation of an angle θ around the x_3 axis.

90.30 Linear Transformations Versus Matrices

Let $A = (a_{ij})$ be a 3×3 matrix. The mapping $x \to Ax$, that is the function $y = f(x) = Ax$, is a transformation from $\mathbb{R}^3$ to $\mathbb{R}^3$. This transformation is linear since $A(x+y) = Ax + Ay$ and $A(\lambda x) = \lambda Ax$ for $\lambda \in \mathbb{R}$. Thus, a 3×3 matrix A generates a linear transformation $f : \mathbb{R}^3 \to \mathbb{R}^3$ with $f(x) = Ax$.

Conversely to each linear transformation $f : \mathbb{R}^3 \to \mathbb{R}^3$, we can associate a matrix A with coefficients given by

$$a_{ij} = f_i(e_j)$$

where $f(x) = (f_1(x), f_2(x), f_3(x))$. The linearity of $f(x)$ implies

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

which shows that a linear transformation $f : \mathbb{R}^3 \to \mathbb{R}^3$ can be represented as $f(x) = Ax$ with the matrix $A = (a_{ij})$ with coefficients $a_{ij} = f_i(e_j)$.

EXAMPLE 90.13. The projection $Px = \frac{x\cdot a}{|a|^2}a$ onto a non-zero vector $a \in \mathbb{R}^3$ takes the matrix form

$$Px = \begin{pmatrix} \frac{a_1^2}{|a|^2} & \frac{a_1a_2}{|a|^2} & \frac{a_1a_3}{|a|^2} \\ \frac{a_2a_1}{|a|^2} & \frac{a_2^2}{|a|^2} & \frac{a_2a_3}{|a|^2} \\ \frac{a_3a_1}{|a|^2} & \frac{a_3a_2}{|a|^2} & \frac{a_3^2}{|a|^2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

EXAMPLE 90.14. 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

$$Px = \begin{pmatrix} 1 - \frac{n_1^2}{|n|^2} & -\frac{n_1n_2}{|n|^2} & -\frac{n_1n_3}{|n|^2} \\ -\frac{n_2n_1}{|n|^2} & 1 - \frac{n_2^2}{|n|^2} & -\frac{n_2n_3}{|n|^2} \\ -\frac{n_3n_1}{|n|^2} & -\frac{n_3n_2}{|n|^2} & 1 - \frac{n_3^2}{|n|^2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

EXAMPLE 90.15. 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

$$(2P - I)x = \begin{pmatrix} 2\frac{a_1^2}{|a|^2} - 1 & 2\frac{a_1a_2}{|a|^2} & 2\frac{a_1a_3}{|a|^2} \\ 2\frac{a_2a_1}{|a|^2} & 2\frac{a_2^2}{|a|^2} - 1 & 2\frac{a_2a_3}{|a|^2} \\ 2\frac{a_3a_1}{|a|^2} & 2\frac{a_3a_2}{|a|^2} & 2\frac{a_3^2}{|a|^2} - 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

90.31 The Scalar Product Is Invariant Under Orthogonal Transformations

Let Q be the matrix $\{q_1, q_2, q_2\}$ formed by taking the columns to be the basis vectors q_j. We assume that Q is orthogonal, which is the same as assuming that $\{q_1, q_2, q_2\}$ is an orthogonal basis, that is the q_j are pairwise orthogonal and have length 1. The coordinates $\hat{x}$ of a vector x in the standard basis with respect to the basis $\{q_1, q_2, q_2\}$ are given by $\hat{x} = Q^{-1}x = Q^\top x$. We shall now prove that if $\hat{y} = Q^\top y$, then

$$\hat{x} \cdot \hat{y} = x \cdot y,$$

which states that the scalar product is invariant under orthogonal coordinate changes. We compute

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

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

$$\begin{aligned} (Ax) \cdot y &= \sum_{i=1}^{3}(\sum_{j=1}^{3} a_{ij}x_j)y_i = \sum_{j=1}^{3}(\sum_{i=1}^{3} a_{ij}y_i)x_j \\ &= (A^\top y) \cdot x = x \cdot (A^\top y), \end{aligned} \tag{90.45}$$

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.

90.32 Looking Ahead to Functions $f : \mathbb{R}^3 \to \mathbb{R}^3$

We have met linear transformations $f : \mathbb{R}^3 \to \mathbb{R}^3$ of the form $f(x) = Ax$, where A is a 3×3 matrix. Below we shall meet more general (non-linear) transformations $f : \mathbb{R}^3 \to \mathbb{R}^3$ that assign a vector $f(x) = (f_1(x), f_2(x), f_3(x)) \in \mathbb{R}^3$ to each $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. For example,

$$f(x) = f(x_1, x_2, x_3) = (x_2 x_3, x_1^2 + x_3, x_3^4 + 5)$$

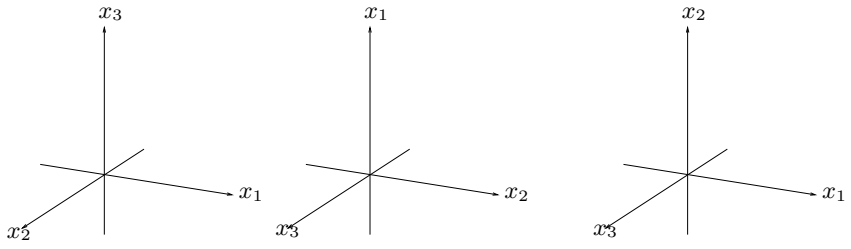
with $f_1(x) = x_2 x_3$, $f_2(x) = x_1^2 + x_3$, $f_3(x) = x_3^4 + 5$. We shall see that we may naturally extend the concepts of Lipschitz continuity and differentiability for functions $f : \mathbb{R} \to \mathbb{R}$ to functions $f : \mathbb{R}^3 \to \mathbb{R}^3$. For example, we say that $f : \mathbb{R}^3 \to \mathbb{R}^3$ 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 90 Problems

90.1. 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.

90.2. Which of the following coordinate systems are righthanded?



90.3. Indicate the direction of $a \times b$ and $b \times a$ in Fig. 90.1 if b points in the direction of the x_1-axis. Consider also the same question in Fig. 90.2.

90.4. Given $a = (1, 2, 3)$ and $b = (1, 3, 1)$, compute $a \times b$.

90.5. Compute the volume of the parallelepiped spanned by the three vectors $(1, 0, 0)$, $(1, 1, 1)$ and $(-1, -1, 1)$.

90.6. What is the area of the triangle formed by the three points: $(1, 1, 0)$, $(2, 3, -1)$ and $(0, 5, 1)$?

90.7. 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.

90.8. 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?

90.9. 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) $\lambda(A + B) = \lambda A + \lambda B$?

90.10. 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.

90.11. 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 $AB \neq BA$.

90.12. Prove Theorem 90.5.

90.13. Write down the three matrices corresponding to rotations around the x_1, x_2 and x_3 axis.

90.14. Find the matrix corresponding to a rotation by the angle θ around a given vector b in $\mathbb{R}^3$.

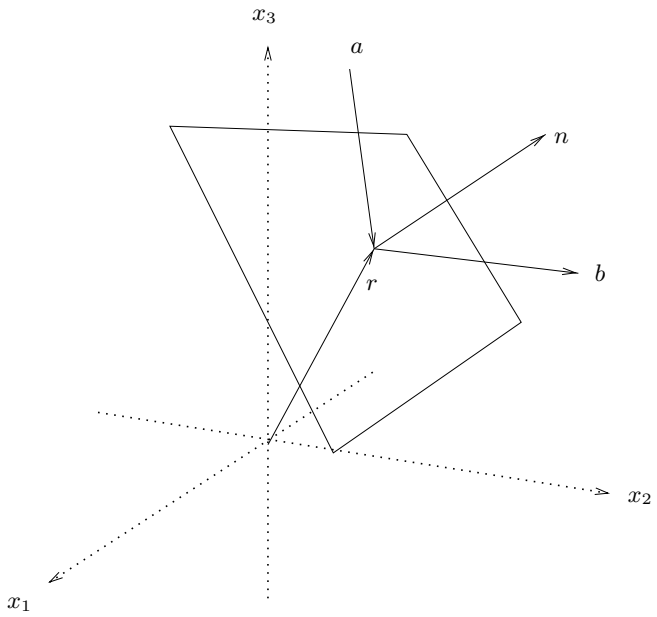
90.15. Give the matrix corresponding to be mirroring a vector through the $x_1 - x_2$-plane.

90.16. 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 $\hat{p}_2$.

90.17. 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?

90.18. 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?

90.19. 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.



90.20. Prove that (90.42) implies that the inverse of A^{-1} exists.

90.21. 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.

90.22. 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 $(b-Pb, Pb-x) = 0$).

90.23. Express (90.31) i matrix form.

90.24. Complete the proof of the claim that (90.30) is uniquely solvable.

90.34 Gösta Mittag-Leffler

The Swedish mentor of Sonya Kovalevskaya was Gösta Mittag-Leffler (1846-1927), famous Swedish mathematician and founder of the prestigous journal Acta Mathematica, see Fig. 90.34. 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.

FIGURE 90.11. 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".

91
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)

91.1 Introduction

In this chapter, we introduce the set of *complex numbers* $\mathbb{C}$. A complex number, typically denoted by z, is an ordered pair $z = (x, y)$ of real numbers x and y, where x represents the *real part* of z and y the *imaginary part* of z. We may thus identify $\mathbb{C}$ with $\mathbb{R}^2$ and we often refer to $\mathbb{C}$ as the *complex plane*. We further identify the set of complex numbers with zero imaginary part with the set of real numbers and write $(x, 0) = x$, viewing the real line $\mathbb{R}$ as the x-axis in the complex plane $\mathbb{C}$. We may thus view $\mathbb{C}$ as an extension of $\mathbb{R}$. Similarly, we identify the set of complex numbers with zero real part with the y-axis, which we also refer to as the set of *purely imaginary* numbers. The complex number $(0, 1)$ is given a special name $i = (0, 1)$, and we refer to i as the *imaginary unit*.

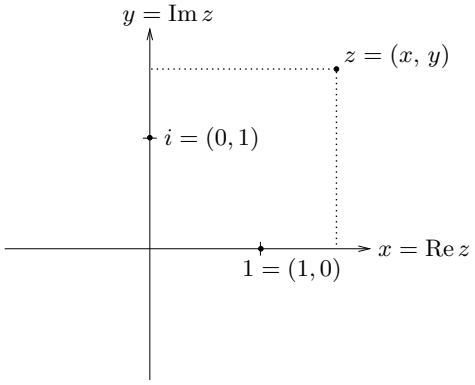


FIGURE 91.1. The complex plane $\mathbb{C} = \mathbb{R}^2$

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.

91.2 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:

$$(a, b) + (c, d) = (a + c, b + d), \tag{91.1}$$

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:

$$(a,b)(c,d)=(ac-bd,ad+bc). \tag{91.2}$$

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

$$z=(x,y)=(x,0)+(0,y)=(x,0)+(0,1)(y,0)=x+iy, \tag{91.3}$$

referring to the identification of complex numbers of the form $(x,0)$ with x, (and similarly $(y,0)$ with y of course) and the notation $i=(0,1)$ introduced above. We refer to x is the *real part of* z and y as the *imaginary part of* z, writing $x=\text{Re } z$ and $y=\text{Im } z$, that is

$$z=\text{Re } z+i\,\text{Im } z=(\text{Re } z,\text{Im } z). \tag{91.4}$$

We note in particular that

$$i^2=i\,i=(0,1)(0,1)=(-1,0)=-(1,0)=-1, \tag{91.5}$$

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 (91.2) 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

$$|z|=(x^2+y^2)^{1/2}, \tag{91.6}$$

that is, $|z|$ is simply the length or norm of the corresponding vector $(x,y)\in\mathbb{R}^2$. We note that if $z=x+iy$, then in particular

$$|x|=|\text{Re } z|\le|z|,\qquad |y|=|\text{Im } z|\le|z|. \tag{91.7}$$

91.3 The Triangle Inequality

If z_1 and z_2 are two complex numbers, then

$$|z_1+z_2|\le|z_1|+|z_2|. \tag{91.8}$$

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

91.4 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.

91.5 Polar Representation of Complex Numbers

Using polar coordinates in $\mathbb{R}^2$, we can express a complex number as follows

$$z = (x, y) = r(\cos(\theta), \sin(\theta)) = r(\cos(\theta) + i\sin(\theta)), \tag{91.9}$$

where $r = |z|$ is the modulus of z and $\theta = \arg z$ is the *argument* of z, and we also used (91.3). We usually assume that $\theta \in [0, 2\pi)$, but by periodicity we may replace θ by $\theta + 2\pi n$ with $n = \pm 1, \pm 2, ...,$. Choosing $\theta \in [0, 2\pi)$, we obtain the *principal argument* of z, which we denote by Arg z.

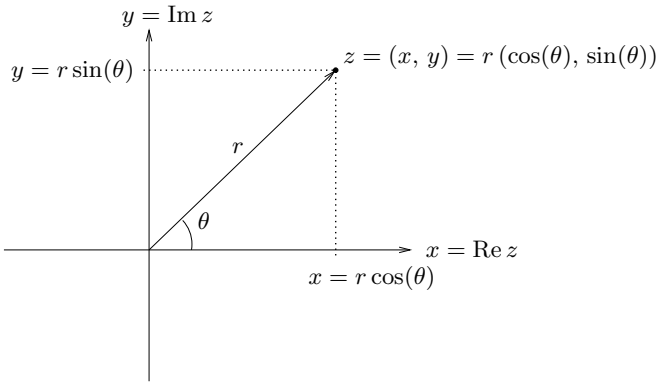


FIGURE 91.2. Polar representation of a complex number

EXAMPLE 91.1. 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 $z = 2(\cos(60°), \sin(60°))$.

91.6 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

$$\arg z\zeta = \text{Arg } z + \text{Arg } \zeta, \quad |z\zeta| = |z||\zeta|. \tag{91.10}$$

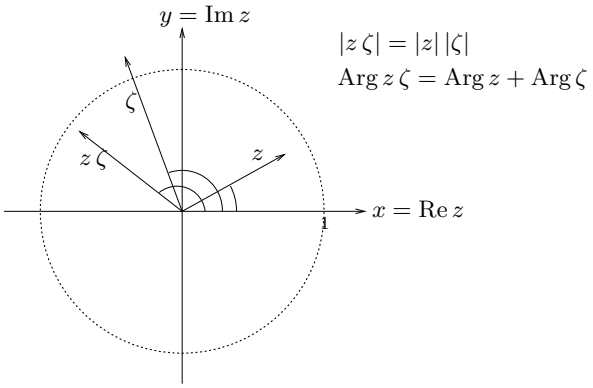


FIGURE 91.3. Geometrical interpretation of multiplication of a complex numbers

EXAMPLE 91.2. Multiplication by i corresponds to rotation counterclockwise $\frac{\pi}{2}$, or 90°.

91.7 Complex Conjugation

If $z = x + iy$ is a complex number with x and y real, we define the complex conjugate $\bar{z}$ of z as

$$\bar{z} = x - iy.$$

We see that z is real if and only if $\bar{z} = z$ and that z is *purely imaginary*, that is Re z = 0, if and only $z = -\bar{z}$.

Identifying $\mathbb{C}$ with $\mathbb{R}^2$, we see that complex conjugation corresponds to reflection in the real axis. We also note the following relations, easily

verified,

$$|z|^2 = z\bar{z}, \quad \text{Re } z = \frac{1}{2}(z + \bar{z}), \quad \text{Im } z = \frac{1}{2i}(z - \bar{z}). \tag{91.11}$$

91.8 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 91.3. 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 ....

91.9 The Fundamental Theorem of Algebra

Consider a polynomial equation $p(z) = 0$, where $p(z) = a_0 + a_1 z + ... + a_n z^n$ is a polynomial in z of degree n with complex coefficients $a_0, ..., a_n$. The Fundamental Theorem of Algebra states that the equation $p(z)$ has at least one complex root z_1 satisfying $p(z_1) = 0$. By the factorization algorithm, it follows that $p(z)$ can be factored into

$$p(z) = (z - z_1)p_1(z),$$

where $p_1(z)$ is a polynomial of degree at most $n-1$. Indeed, the factorization algorithm from the Chapter Combinations of functions (Section 11.4) shows that

$$p(z) = (z - z_1)p_1(z) + c,$$

where c is a constant. Setting $z = z_1$, it follows that $c = 0$. Repeating the argument, we find that $p(z)$ can be factored into

$$p(z) = c(z - z_1)....(z - z_n),$$

where $z_1, .., z_n$ are the (complex valued in general) roots of $p(z) = 0$.

91.10 Roots

Consider the equation in $w \in \mathbb{C}$

$$w^n = z,$$

where $n = 1, 2, ..$ is a natural number and $z \in \mathbb{C}$ is given. Using polar coordinates with $z = |z|(\cos(\theta), \sin(\theta)) \in \mathbb{C}$ and $w = |w|(\cos(\varphi), \sin(\varphi)) \in \mathbb{C}$, the equation $w^n = z$ takes the form

$$|w|^n(\cos(n\varphi), \sin(n\varphi)) = |z|(\cos(\theta), \sin(\theta))$$

from which it follows that

$$|w| = |z|^{\frac{1}{n}}, \qquad \varphi = \frac{\theta}{n} + 2\pi\frac{k}{n},$$

where $k = 0, ..., n-1$. We conclude that the equation $w^n = z$ has n distinct roots on the circle $|w| = |z|^{\frac{1}{n}}$. In particular, the equation $w^2 = -1$ has the two roots $w = \pm i$. The n roots of the equation $w^n = 1$ are called the n *roots of unity*.

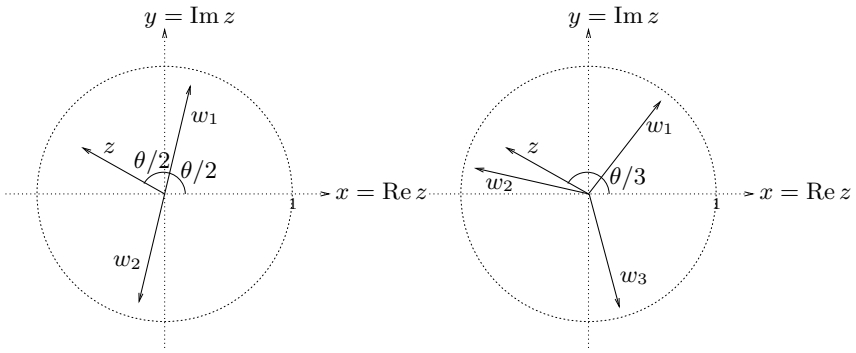


FIGURE 91.4. The "square" and "cubic" roots of z.

91.11 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 91 Problems

91.1. Show that (a) $\frac{1}{i} = -i$, (b) $i^4 = 1$.

91.2. Find (a) Re $\frac{1}{1+i}$, (b) Im $\frac{3+4i}{7-i}$, (c) Im $\frac{z}{\bar{z}}$.

91.3. 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) $\frac{z_1}{z_1+z_2}$.

91.4. 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.

91.5. Represent in polar form (a) $1 + i$, (b) $\frac{1+i}{1-i}$, (c) $\frac{2+3i}{5+4i}$.

91.6. Solve the equations (a) $z^2 = i$, (b) $z^8 = 1$, (c) $z^2 + z + 1 = -i$, (d) $z^4 - 3(1 + 2i)z^2 + 6i = 0$.

91.7. Determine the sets in the complex plane represented by (a) $|\frac{z+i}{z-i}| = 1$, (b) Im $z^2 = 2$, (c) $|\text{Arg } z| \leq \frac{\pi}{4}$.

91.8. Express z/w in polar coordinates in terms of the polar coordinates of z and w.

91.9. 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) $f(z) = z^{\frac{1}{2}}$.

92
Analytic Geometry in $\mathbb{R}^n$

> I also think that the (mathematical) mine has become too deep and sooner or later it will be necessary to abandon it if new ore-bearing veins shall not be discovered. Physics and Chemistry display now treasures much more brilliant and easily exploitable, thus, apparently, everybody has turned completely in this direction, and possibly posts in Geometry in the Academy of Sciences will some day be like chairs in Arabic Language in universities at present. (Lagrange, 1781)

92.1 Introduction and Survey of Basic Objectives

We now generalize the discussion of analytic geometry to $\mathbb{R}^n$, where n is an arbitrary natural number. Following the pattern set above for $\mathbb{R}^2$ and $\mathbb{R}^3$, we define $\mathbb{R}^n$ to be the set of all possible ordered n-tuples of the form $(x_1, x_2, ..., x_n)$ with $x_i \in \mathbb{R}$ for $i = 1, ..., n$. We refer to $\mathbb{R}^n$ as *n-dimensional Euclidean space*.

We all have a direct concrete experience of $\mathbb{R}^3$ as the three-dimensional space of the real World, and we may think of $\mathbb{R}^2$ as an infinite flat surface, but we don't have a similar experience with for example $\mathbb{R}^4$, except possibly from some science fiction novel with space ships travelling in four-dimensional space-time. Actually, Einstein in his theory of relativity used $\mathbb{R}^4$ as the set of space-time coordinates (x_1, x_2, x_3, x_4) with $x_4 = t$ representing time, but of course had the same difficulty as we all have of "seeing"

an object in $\mathbb{R}^4$. In Fig. 42.1, we show a projection into $\mathbb{R}^3$ of a 4-cube in $\mathbb{R}^4$, and we hope the clever reader can "see" the 4-cube.

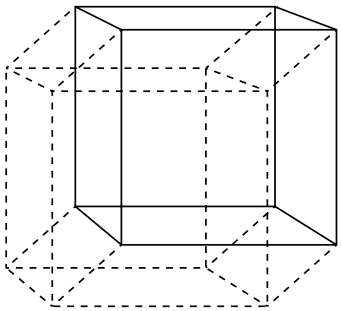


FIGURE 92.1. A cube in $\mathbb{R}^4$

More generally, the need of using $\mathbb{R}^n$ arises as soon as we have n different variables to deal with, which occurs all the time in applications, and $\mathbb{R}^n$ is thus one of the most useful concepts in mathematics. Fortunately, we can work with $\mathbb{R}^n$ purely algebraically without having to draw geometric pictures, that is we can use the tools of analytic geometry in $\mathbb{R}^n$ in pretty much the same way as we have done in $\mathbb{R}^2$ and $\mathbb{R}^3$.

Most of this chapter is one way or the other connected to systems of m linear equations in n unknowns $x_1, ..., x_n$, of the form

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \quad \text{for } i = 1, ..., m, \tag{92.1}$$

that is,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n &= b_2, \\ &..... \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n &= b_m, \end{aligned} \tag{92.2}$$

where the a_{ij} are given (real) coefficients and $(b_1, ..., b_m) \in \mathbb{R}^m$ is a given right-hand side. We will write this system in matrix form as

$$Ax = b, \tag{92.3}$$

that is

$$\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ . \\ . \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ . \\ . \\ b_m \end{pmatrix}, \tag{92.4}$$

where $A = (a_{ij})$ is a $m \times n$ matrix with rows $(a_{i1}, ..., a_{in})$, $i = 1, ..., m$, and columns $(a_{1j}, ..., a_{mj})$, $j = 1, ..., n$, and we view $x = (x_1, ..., x_n) \in \mathbb{R}^n$ and

$b = (b_1, ..., b_m) \in \mathbb{R}^m$ as column vectors. We will also write the system in the form

$$x_1 a_1 + \cdots + x_n a_n = b, \tag{92.5}$$

expressing the given column vector $b \in \mathbb{R}^m$ as a linear combination of the column vectors $a_j = (a_{1j}, a_{2j}, ..., a_{mj})$, $j = 1, 2, ..., n$, with coefficients $(x_1, ..., x_n)$. Notice that we use both (column) vectors in $\mathbb{R}^m$ (such as the columns of the matrix A and the right hand side b) and (column) vectors in $\mathbb{R}^n$ such as the solution vector x.

We shall view $f(x) = Ax$ as a function or transformation $f : \mathbb{R}^n \to \mathbb{R}^m$, and we thus focus on a particular case of our general problem of solving systems of equations of the form $f(x) = b$, where $f : \mathbb{R}^n \to \mathbb{R}^m$ is the *linear transformation* $f(x) = Ax$. We shall denote by $R(A)$ the *range* of $f(x) = Ax$, that is

$$R(A) = \{Ax \in \mathbb{R}^m : x \in \mathbb{R}^n\} = \{\sum_{j=1}^{n} x_j a_j : x_j \in \mathbb{R}\},$$

and by $N(A)$ the *null space* of $f(x) = Ax$ that is

$$N(A) = \{x \in \mathbb{R}^n : Ax = 0\} = \{x \in \mathbb{R}^n : \sum_{j=1}^{n} x_j a_j = 0\}.$$

We are interested in the question of existence and/or uniqueness of solutions $x \in \mathbb{R}^n$ to the problem $Ax = b$ for a given $m \times n$ matrix A and right hand side $b \in \mathbb{R}^m$. Of particular interest is the case $m = n$ with as many equations as unknowns.

Existence of a solution x to $Ax = b$ is of course the same as saying that $b \in R(A)$, which is the same as saying that b is a linear combination of the columns of A. Uniqueness is the same as saying that $N(A) = 0$, because if x and $\hat{x}$ satisfy $Ax = b$ and $A\hat{x} = b$, then by linearity, $A(x - \hat{x}) = 0$, and if $N(A) = 0$ then $x - \hat{x} = 0$ that is $x = \hat{x}$. Further, the non-uniqueness of solutions of $Ax = b$ is described by $N(A)$: If $A\hat{x} = b$ and $Ax = b$, then $x - \hat{x} \in N(A)$.

We may thus formulate the following prime objectives of our study of the linear transformation $f(x) = Ax$ given by the matrix A:

- Determine $R(A)$.
- Determine $N(A)$.
- Solve $Ax = b$ for given b.

We state here the following partial answer given by the *Fundamental Theorem of Linear Algebra*, which we will prove in a couple of different ways below: Let $m = n$ and suppose that $N(A) = 0$. Then $Ax = b$ has a unique

solution for any $b \in \mathbb{R}^m$, that is, $R(A) = \mathbb{R}^m$. In other words, if $m = n$, then uniqueness implies existence.

In our study we will be led to concepts such as: linear combination, linear span, linear space, vector space, subspace, linear independence, basis, determinant, linear transformation and projection, which we have already met in the chapters on analytic geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$ above.

This chapter focusses mostly on theoretical issues while the computational methods such as Gaussian elimination and iterative methods are considered in more detail in Chapter *Solving systems of linear equations* below.

92.2 Body/Soul and Artificial Intelligence

Before plunging into the geometry of $\mathbb{R}^n$, we take a brake and return to the story of Body and Soul which continues into our time with new questions: Is it possible to create computer programs for Artificial Intelligence AI, that is, can we give the computer some more or less advanced capability of acting like an intelligent organism with some ability of "thinking"? It appears that this question does not yet have a clear positive answer, despite many dreams in that direction during the development of the computer. In seeking an answer, Spencer's principle of adaptivity of course plays an important role: an intelligent system must be able to adapt to changes in its environment. Further, the presence of a goal or final cause according to Leibniz, seems to be an important feature of intelligence, to judge if an action of a system is stupid or not. Below we will design adaptive IVP-solvers, which are computer programs for solving systems of differential equations, with features of adaptive feed-back from the computational process towards the goal of error control. These IVP-solvers thus show some kind of rudiments of intelligence, and at any rate are infinitely much more "clever" than traditional non-adaptive IVP-solvers with no feed-back.

92.3 The Vector Space Structure of $\mathbb{R}^n$

We view $\mathbb{R}^n$ as a *vector space* consisting of *vectors* which are ordered n-tuples, $x = (x_1, ..., x_n)$ with components $x_i \in \mathbb{R}$, $i = 1, ..., n$. We write $x = (x_1, ..., x_n)$ for short, and refer to $x \in \mathbb{R}^n$ as a vector with component x_i in position i.

We may *add* two vectors $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ in $\mathbb{R}^n$ by componentwise addition to get a new vector $x + y$ in $\mathbb{R}^n$ defined by

$$x + y = (x_1 + y_1, x_2 + y_2, ..., x_n + y_n). \tag{92.6}$$

Further, we may multiply a vector $x = (x_1, ..., x_n)$ by a real number λ by componentwise multiplication with λ, to get a new vector λx in $\mathbb{R}^n$ defined

by

$$\lambda x = (\lambda x_1, ..., \lambda x_n). \tag{92.7}$$

The operations of adding two vectors in $\mathbb{R}^n$ and multiplying a vector in $\mathbb{R}^n$ with a real number, of course directly generalize the corresponding operations from the cases $n = 2$ and $n = 3$ considered above. The generalization helps us to deal with $\mathbb{R}^n$ using concepts and tools which we have found useful in $\mathbb{R}^2$ and $\mathbb{R}^3$.

We may thus add vectors in $\mathbb{R}^n$ and multiply them by real numbers (scalars), the usual commutative and distributive rules hold for these operations, and $\mathbb{R}^n$ is thus a vector space. We say that $(0, 0, .., 0)$ is the *zero vector* in $\mathbb{R}^n$ and write $0 = (0, 0, .., 0)$.

Linear algebra concerns vectors in vector spaces, also refereed to as *linear spaces*, and linear functions of vectors, that is *linear transformations* of vectors. As we just saw, $\mathbb{R}^n$ is a vector space, but there are also many other types of vector spaces, where the vectors have a different nature. In particular, we will below meet vector spaces consisting of vectors which are functions. In this chapter we focus on $\mathbb{R}^n$, the most basic of all vector spaces. We know that linear transformations in $\mathbb{R}^2$ and $\mathbb{R}^3$ lead to 2×2 and 3×3 matrices, and we shall now generalize to linear transformations from $\mathbb{R}^n$ into $\mathbb{R}^m$ which can be represented by $m \times n$ *matrices*.

We give in this chapter a condensed (and dry) presentation of some basic facts of linear algebra in $\mathbb{R}^n$. Many applications of the theoretical results presented will appear in the rest of the book.

92.4 The Scalar Product and Orthogonality

We define the *scalar product* $x \cdot y = (x, y)$ of two vectors x and y in $\mathbb{R}^n$, by

$$x \cdot y = (x, y) = \sum_{i=1}^{n} x_i y_i. \tag{92.8}$$

This generalizes the scalar product in $\mathbb{R}^2$ and $\mathbb{R}^3$. Note that here we introduce a new notation for the scalar product of two vectors x and y, namely (x, y), as an alternative to the "dot product" $x \cdot y$ used in $\mathbb{R}^2$ and $\mathbb{R}^3$. We should be ready to use both notations.

The scalar product is *bilinear* in the sense that $(x+y, z) = (x, z)+(y, z)$, $(\lambda x, z) = \lambda(x, z)$, $(x, y + z) = (x, y) + (x, z)$ and $(x, \lambda y) = \lambda(x, y)$, and *symmetric* in the sense that $(x, y) = (y, x)$, for all vectors $x, y, z \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$.

We say that two vectors x and y in $\mathbb{R}^n$ are *orthogonal* if $(x, y) = 0$. We define

$$|x| = (\sum_{i=1}^{n} x_i^2)^{1/2} = (x, x)^{1/2} \tag{92.9}$$

to be the Euclidean *length* or *norm* of the vector x. Note that this definition of *length* is a direct generalization of the natural length $|x|$ of a vector x in $\mathbb{R}^n$, $n = 1, 2, 3$.

EXAMPLE 92.1. Let $x = (2, -4, 5, 1, 3)$ and $y = (1, 4, 6, -1, 2)$ be two vectors in $\mathbb{R}^5$. We compute $(x, y) = 2 \times 1 + (-4) \times 4 + 5 \times 6 + 1 \times (-1) + 3 \times 2 = 21$.

92.5 Cauchy's Inequality

Cauchy's inequality states that for $x, y \in \mathbb{R}^n$,

$$|(x, y)| \leq |x|\,|y|.$$

In words: the absolute value of the scalar product of two vectors is bounded by the product of the norms of the vectors. We prove Cauchy's inequality by noting that for all $s \in \mathbb{R}$,

$$0 \leq |x + sy|^2 = (x + sy, x + sy) = |x|^2 + 2s(x, y) + s^2|y|^2,$$

and then assuming that $y \neq 0$, choosing $s = -(x, y)/|y|^2$ (which minimizes the right-hand side), to get

$$0 \leq |x|^2 - 2\frac{(x, y)^2}{|y|^2} + \frac{(x, y)^2}{|y|^2} = |x|^2 - \frac{(x, y)^2}{|y|^2},$$

which proves the desired result.

We recall that for $n = 2, 3$,

$$(x, y) = x \cdot y = \cos(\theta)|x||y|,$$

where θ is the angle between x and y, from which of course Cauchy's inequality follows directly using the fact that $|\cos(\theta)| \leq 1$.

We define the *angle* $\theta \in [0, 2\pi)$ between two non-zero vectors x and y in $\mathbb{R}^n$ by

$$\cos(\theta) = \frac{(x, y)}{|x||y|}, \tag{92.10}$$

which generalizes the corresponding notion for $n = 2, 3$.

EXAMPLE 92.2. The angle between the vectors $x = (1, 2, 3, 4)$ and $y = (4, 3, 2, 1)$ in $\mathbb{R}^4$ is equal to $\arccos\frac{2}{3} \approx 0.8411 \approx 48^\circ$ since $(x, y) = 20$ and $|x| = |y| = \sqrt{30}$.

92.6 The Linear Combinations of a Set of Vectors

We know that two non-parallel vectors a_1 and a_2 in $\mathbb{R}^3$ define a plane in $\mathbb{R}^3$ through the origin consisting of all the linear combinations $\lambda_1 a_1 + \lambda_2 a_2$ with coefficients λ_1 and λ_2 in $\mathbb{R}$. The normal to the plane is given by $a_1 \times a_2$. A plane through the origin is an example of *subspace* of $\mathbb{R}^3$, which is a subset of $\mathbb{R}$ with the property that vector addition and scalar multiplication does not lead outside the set. So, a subset S of $\mathbb{R}^3$ is a subspace if the sum of any two vectors in S belongs to S and scalar multiplication of a vector in S gives a vector in S. Clearly, a plane through the origin is a subspace of $\mathbb{R}^3$. Similarly, a line through the origin defined as the scalar multiples $\lambda_1 a_1$ with coefficients $\lambda_1 \in \mathbb{R}$ and a_1 a given vector in $\mathbb{R}^3$, is a subspace of $\mathbb{R}^3$. The subspaces of $\mathbb{R}^3$ consist of lines and planes through the origin. Notice that a plane or line in $\mathbb{R}^3$ not passing through the origin, is not a subspace.

More generally, we use the concept of a *vector space* to denote a set of vectors for which the operations of vector addition and scalar multiplication does not lead outside the set. Of course, $\mathbb{R}^3$ is a vector space. A subspace of $\mathbb{R}^3$ is a vector space. A plane or line in $\mathbb{R}^3$ through the origin is a vector space. The concept of vector space is fundamental in mathematics and we will meet this term many times below.

We will now generalize to $\mathbb{R}^m$ with $m > 3$ and we will then meet new examples of vector spaces and subspaces of vector spaces. Let a_1, a_2,...,a_n, be n non-zero vectors in $\mathbb{R}^m$. A vector b in $\mathbb{R}^m$ of the form

$$b = \lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_n a_n, \tag{92.11}$$

where the $\lambda_i \in \mathbb{R}$, is said to be a *linear combination* of the set of vectors $\{a_1, ..., a_n\}$ with *coefficients* $\lambda_1, ..., \lambda_n$. If

$$c = \mu_1 a_1 + \mu_2 a_2 + \cdots + \mu_n a_n, \tag{92.12}$$

is another linear combination of $\{a_1, ..., a_n\}$ with coefficients $\mu_j \in \mathbb{R}$, then the vector

$$b + c = (\lambda_1 + \mu_1)a_1 + (\lambda_2 + \mu_2)a_2 + \cdots + (\lambda_+ \mu_n)a_n, \tag{92.13}$$

is again a linear combination of $\{a_1, ..., a_m\}$ now with coefficients $\lambda_j + \mu_j$. Further, for any $\alpha \in \mathbb{R}$ the vector

$$\alpha b = \alpha\lambda_1 a_1 + \alpha\lambda_2 a_2 + \cdots + \alpha\lambda_m a_m \tag{92.14}$$

is also a linear combination of $\{a_1, ..., a_m\}$ with coefficients $\alpha\lambda_j$. This means that if we let $S(a_1, ...a_n)$ denote the set of all linear combinations

$$\lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_n a_n, \tag{92.15}$$

of $\{a_1, ..., a_n\}$, where the coefficients $\lambda_j \in \mathbb{R}$, then $S(a_1, ...a_n)$ is indeed a vector space, since vector addition and multiplication by scalars do not lead

outside the set. The sum of two linear combinations of $\{a_1,...,a_n\}$ is also a linear combination of $\{a_1,...,a_n\}$, and a linear combination of $\{a_1,...,a_n\}$ multiplied by a real number is also a linear combination of $\{a_1,...,a_n\}$.

We refer to the vector space $S(a_1,...a_n)$ of all linear combinations of the form (92.15) of the vectors $\{a_1,...,a_n\}$ in $\mathbb{R}^m$ as the *subspace of* $\mathbb{R}^m$ *spanned by the vectors* $\{a_1,...,a_n\}$, or simply just the *span of* $\{a_1,...,a_n\}$, which we may describe as:

$$S(a_1,...a_n) = \Big\{\sum_{i=1}^{n} \lambda_j a_j :\ \lambda_j \in \mathbb{R}, j = 1,...,n\Big\}.$$

If $m = 2$ and $n = 1$, then the subspace $S(a_1)$ is a line in $\mathbb{R}^2$ through the origin with direction a_1. If $m = 3$ and $n = 2$, then $S(a_1, a_2)$ corresponds to the plane in $\mathbb{R}^3$ through the origin spanned by a_1 and a_2 (assuming a_1 and a_2 are non-parallel), that is, the plane through the origin with normal given by $a_1 \times a_2$.

Note that for any $\mu \in \mathbb{R}$, we have

$$S(a_1, a_2, ..., a_n) = S(a_1, a_2 - \mu a_1, a_3, ..., a_n), \tag{92.16}$$

since we can replace each occurrence of a_2 by the linear combination $(a_2 - \mu a_1) + \mu a_1$ of $a_2 - \mu a_1$ and a_1. More generally, we can add any multiple of one vector to one of the other vectors without changing the span of the vectors! Of course we may also replace any vector a_j with a μa_j where μ is a non-zero real number without changing the span. We shall return to these operations below.

92.7 The Standard Basis

The set of vectors in $\mathbb{R}^n$:

$$\{(1,0,0,\cdots,0,0),(0,1,0,\cdots,0,0),\cdots,(0,0,0,\cdots,0,1)\},$$

commonly denoted by $\{e_1,....,e_n\}$, where $e_i = (0,0,..,0,1,0,...,0)$ with a single coefficient 1 at position i, is called the *standard basis* for $\mathbb{R}^n$. Any vector $x = (x_1,...,x_n) \in \mathbb{R}^n$ can be written as a linear combination of the basis vectors $\{e_1,...,e_n\}$:

$$x = x_1e_1 + x_2e_2 + \cdots + x_ne_n, \tag{92.17}$$

with the coefficients x_j of x appearing as coefficients of the basis vectors e_j. We note that $(e_j, e_k) = e_j \cdot e_k = 0$ for $j \neq k$, that is the standard basis vectors are *pairwise orthogonal*, and of length one since $(e_j, e_j) = |e_j|^2 = 1$. We may thus express the coefficients x_i of a given vector $x = (x_1,...,x_n)$ with respect to the standard basis $\{e_1,....,e_n\}$ as follows:

$$x_i = (e_i, x) = e_i \cdot x. \tag{92.18}$$

92.8 Linear Independence

We recall that to specify a plane in $\mathbb{R}^3$ as the set of linear combinations of two given vectors a_1 and a_2, we assume that a_1 and a_2 are non-parallel. The generalization of this condition to a set $\{a_1, ..., a_m\}$ of m vectors in $\mathbb{R}^n$, is referred to as *linear independence*, which we now proceed to define. Eventually, this will lead us to the concept of *basis* of a vector space, which is one of the most basic(!) concepts of linear algebra.

A set $\{a_1, ..., a_n\}$ of vectors in $\mathbb{R}^m$ is said to be *linearly independent* if none of the vectors a_i can be expressed as a linear combination of the others. Conversely, if some of the vectors a_i can be expressed as a linear combination of the others, for example if

$$a_1 = \lambda_2 a_2 + ... + \lambda_n a_n \tag{92.19}$$

for some numbers λ_2,...,λ_n, we say that the set $\{a_1, a_2, ..., a_n\}$ is *linearly dependent*. As a test of linear independence of $\{a_1, a_2, ..., a_n\}$, we can use: if

$$\lambda_1 a_1 + \lambda_2 a_2 + ... + \lambda_n a_n = 0 \tag{92.20}$$

implies that $\lambda_1 = \lambda_2 = ... = \lambda_n = 0$, then $\{a_1, a_2, ..., a_n\}$ is linearly independent. This is because if (92.20) holds with some of the λ_j different from 0, for example, $\lambda_1 \neq 0$, then we could divide by λ_1 and express a_1 as a linear combination of $\{a_2, ..., a_m\}$:

$$a_1 = -\frac{\lambda_2}{\lambda_1} a_2 + ... + -\frac{\lambda_n}{\lambda_1} a_n. \tag{92.21}$$

The standard basis $\{e_1, ..., e_n\}$ is (of course) a linearly independent set, since if

$$\lambda_1 e_1 + ... + \lambda_n e_n = 0,$$

then $\lambda_i = 0$ for $i = 1, ..., n$, because $0 = (0, 0, ..., 0) = \lambda_1 e_1 + ... + \lambda_n e_n = (\lambda_1, ..., \lambda_n)$.

92.9 Reducing a Set of Vectors to get a Basis

Consider the subspace $S(a_1, ..., a_n)$ spanned by the set of vectors $\{a_1, a_2, ..., a_n\}$. If the set $\{a_1, a_2, ..., a_n\}$ is linearly dependent, say that a_n can be expressed as a linear combination of $\{a_1, ..., a_{n-1}\}$, then $S(a_1, ..., a_n)$ is in fact spanned by $\{a_1, ..., a_{n-1}\}$ and thus $S(a_1, ..., a_n) = S(a_1, ..., a_{n-1})$. This follows simply by replacing all occurrences of a_n by its linear combination of $\{a_1, ..., a_{n-1}\}$. Continuing this way, eliminating linearly dependent vectors, we may express $S(a_1, ..., a_n)$ as the span of $\{a_1, a_2, ..., a_k\}$ (with a suitable enumeration), that is, $S(a_1, ..., a_n) = S(a_1, a_2, ..., a_k\}$, where $k \leq n$, and the set $\{a_1, a_2, ..., a_k\}$ is linearly independent. This means that $\{a_1, a_2, ..., a_k\}$ is

a *basis* for the vector space $S = S(a_1, ..., a_n)$ in the sense that the following two conditions are fulfilled:

- any vector in S can be expressed as a linear combination of $\{a_1, a_2, ..., a_k\}$,
- the set $\{a_1, a_2, ..., a_k\}$ is linearly independent.

Note that by the linear independence the coefficients in the linear combination are uniquely determined: if two linear combinations $\sum_{j=1}^k \lambda_j a_j$ and $\sum_{j=1}^k \mu_j a_j$ are equal, then $\lambda_j = \mu_j$ for $j = 1, .., k$.

Each vector $b \in S$ can thus be expresses as a unique linear combination of the basis vectors $\{a_1, a_2, ..., a_k\}$:

$$b = \sum_{j=1}^k \lambda_j a_j,$$

and we refer to $(\lambda_1, ..., \lambda_k)$ as the *coefficients* of b with respect to the basis $\{a_1, a_2, ..., a_k\}$.

The *dimension* of a vector space S is equal to the number of basis vectors in a basis for S. We prove below that the dimension is uniquely defined so that two sets of basis vectors always have the same number of elements.

EXAMPLE 92.3. Consider the three vectors $a_1 = (1, 2, 3, 4)$, $a_2 = (1, 1, 1, 1)$, and $a_3 = (3, 3, 5, 6)$ in $\mathbb{R}^4$. We see that $a_3 = a_1 + 2a_2$, and thus the set $\{a_1, a_2, a_3\}$ is linearly dependent. The span of $\{a_1, a_2, a_3\}$ thus equals the span of $\{a_1, a_2\}$, since each occurrence of a_3 can be replaced by $a_1 + 2a_2$. The vector a_3 is thus redundant, since it can be replaced by a linear combination of a_1 and a_2. Evidently, $\{a_1, a_2\}$ is linearly independent, since a_1 and a_2 are non-parallel. Thus, $\{a_1, a_2\}$ is a linearly independent set spanning the same subset as $\{a_1, a_2, a_3\}$. We can also express a_2 in terms of a_1 and a_3, or a_1 in terms of a_2 and a_3, and thus any set of two vectors $\{a_1, a_2\}$, $\{a_1, a_3\}$ or $\{a_2, a_3\}$, can serve as a basis for the subspace spanned by $\{a_1, a_2, a_3\}$.

92.10 Using Column Echelon Form to Obtain a Basis

We now present a constructive process for determining a basis for the vector space $S(a_1, ..., a_n)$ spanned by the set of vectors $\{a_1, a_2, ..., a_n\}$, where $a_j = (a_{1j}, ..., a_{mj}) \in \mathbb{R}^m$ for $j = 1, ..., n$ which we view as column vectors. We refer to this process as *reduction to column echelon form*. It is of fundamental importance and we shall return to it below in several different contexts. Assume then first that $a_{11} = 1$ and choose $\mu \in \mathbb{R}$ so that

$\mu a_{11} = a_{12}$, and note that $S(a_1, ..., a_n) = S(a_1, a_2 - \mu a_1, a_3, ..., a_n)$, where now the first component of $a_2 - \mu a_1$ is zero. We here used the fact that we can add one vector multiplied by a scalar to another vector without changing the span of the vectors. Continuing in the same way we obtain $S(a_1, ..., a_n) = S(a_1, \hat{a}_2, \hat{a}_3, ...\hat{a}_n)$ where $\hat{a}_{1j} = 0$ for $j > 1$. In matrix form with the $a_j \in \mathbb{R}^m$ as column vectors, we may express this as follows:

$$S\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ a_{21} & a_{22} & .. & a_{2n} \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix} = S\begin{pmatrix} 1 & 0 & .. & 0 \\ a_{21} & \hat{a}_{22} & .. & \hat{a}_{22} \\ . & . & .. & . \\ a_{m1} & \hat{a}_{m2} & .. & \hat{a}_{mn} \end{pmatrix}$$

We can now repeat the process by cutting out the first row and first column and reduce to a set of $n-1$ vectors in $\mathbb{R}^{m-1}$. Before doing this we take care of the case $a_{11} \neq 1$. If $a_{11} \neq 0$, then we transform to the case $a_{11} = 1$ by replacing a_1 by μa_1 with $\mu = 1/a_{11}$, noting that we can multiply any column with a non-zero real number without changing the span. By renumbering the vectors we may then assume that either $a_{11} \neq 0$, which thus led to the above construction, or $a_{1j} = 0$ for $j = 1, ..., n$, in which case we seek to compute a basis for

$$S\begin{pmatrix} 0 & 0 & .. & 0 \\ a_{21} & a_{22} & .. & a_{22} \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix}$$

with only zeros in the first row. We may then effectively cut out the first row and reduce to a set of n vectors in $\mathbb{R}^{m-1}$.

Repeating now the indicated process, we obtain with $k \leq \min(n, m)$,

$$S\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix} = S\begin{pmatrix} 1 & 0 & 0 & .. & 0 & .. & 0 \\ \hat{a}_{21} & 1 & 0 & .. & 0 & .. & 0 \\ . & . & . & 0 & 0 & .. & 0 \\ . & . & .. & 1 & 0 & .. & 0 \\ . & . & .. & . & . & .. & . \\ \hat{a}_{m1} & \hat{a}_{m2} & .. & \hat{a}_{mk} & 0 & .. & 0 \end{pmatrix},$$

where we refer to the matrix on the right as the *column echelon form* of the matrix to the left, and k is the number of non-zero columns. We see that each non-zero column $\hat{a}_j$, $j = 1, ..., k$, in the echelon form has a coefficient equal to 1 and that all matrix elements to the right and above that coefficient is equal to zero. Further, the ones appear in a staircase form descending to the right on or below the diagonal. The set of non-zero columns $\{\hat{a}_1, ..., \hat{a}_k\}$ is linearly independent, because if

$$\sum_{j=1}^{k} \hat{x}_j \hat{a}_j = 0,$$

then we get successively $\hat{x}_1 = 0$, $\hat{x}_2 = 0$,..., $\hat{x}_k = 0$, and thus $\{\hat{a}_1, ..., \hat{a}_k\}$ forms a basis for $S(a_1, ..., a_n)$. The dimension of $S(a_1, ..., a_n)$ is equal to k. If zero columns appear in the echelon form, then the original set $\{a_1, ..., a_n\}$ is linearly dependent.

We note that, because of the construction, zero columns must appear if $n > m$, and we thus understand that a set of n vectors in $\mathbb{R}^m$ is linearly dependent if $n > m$. We may also understand that if $n < m$, then the set $\{a_1, ..., a_n\}$ cannot span $\mathbb{R}^m$, because if $k < m$, then there are vectors $b \in \mathbb{R}^m$ which cannot be expressed as linear combinations of $\{\hat{a}_1, ..., \hat{a}_k\}$ as we now show: if

$$b = \sum_{j=1}^{k} \hat{x}_j \hat{a}_j,$$

then the coefficients $\hat{x}_1, ..., \hat{x}_k$ are determined by the coefficients $b_1, ..., b_k$, of b occurring in the rows with the coefficient 1. For example, in the case the 1s appear on the diagonal, we first compute $\hat{x}_1 = b_1$, then $\hat{x}_2 = b_1 - \hat{a}_{21}\hat{x}_1$ etc, and thus the remaining coefficients $b_{k+1}, ..., b_m$ of b cannot be arbitrary.

92.11 Using Column Echelon Form to Obtain $R(A)$

By reduction to column echelon form we can construct a basis for $R(A)$ for a given $m \times n$ matrix A with column vectors $a_1, ..., a_n$ because

$$Ax = \sum_{j=1}^{n} x_j a_j$$

and thus $R(A) = S(a_1, ..., a_n)$ expressing that the range $R(A) = \{Ax : x \in \mathbb{R}^n\}$ is equal to the vector space $S(a_1, ..., a_n)$ of all linear combinations of the set of column vectors $\{a_1, ..., a_n\}$. Setting now

$$A = \begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix}, \qquad \hat{A} = \begin{pmatrix} 1 & 0 & 0 & .. & 0 & .. & 0 \\ \hat{a}_{21} & 1 & 0 & .. & 0 & .. & 0 \\ . & . & . & 0 & 0 & .. & 0 \\ . & . & .. & 1 & 0 & .. & 0 \\ . & . & .. & . & . & .. & . \\ \hat{a}_{m1} & \hat{a}_{m2} & .. & \hat{a}_{mk} & 0 & .. & 0 \end{pmatrix}$$

with $\hat{A}$ obtained from A by reduction to column echelon form, we have

$$R(A) = R(\hat{A}) = S(\hat{a}_1, ..., \hat{a}_k),$$

and thus $\{\hat{a}_1, ..., \hat{a}_k\}$ forms a basis for $R(A)$. In particular we can easily check if a given vector $b \in \mathbb{R}^m$ belongs to $R(A)$, by using the echelon form. By reduction to column echelon form we can thus give an answer

to the basic problem of determining $R(A)$ for a given matrix A. Not bad. For example, in the case $m = n$ we have that $R(A) = \mathbb{R}^m$ if and only if $k = n = m$, in which case the echelon form $\hat{A}$ has 1s all along the diagonal.

We give an example showing the sequence of matrices appearing in reduction to column echelon form:

EXAMPLE 92.4. We have

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 7 \\ 1 & 3 & 4 & 5 & 8 \\ 1 & 4 & 5 & 6 & 9 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 2 & 3 & 6 \\ 1 & 2 & 3 & 4 & 7 \\ 1 & 3 & 4 & 5 & 8 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & -1 & -2 & -5 \\ 1 & 3 & -2 & -4 & -10 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & -2 & -5 \\ 1 & 3 & 2 & -4 & -10 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 2 & 0 & 0 \end{pmatrix} = \hat{A}.$$

We conclude that $R(A)$ is spanned by the 3 non-zero columns of $\hat{A}$ and thus in particular that the dimension of $R(A)$ is equal to 3. In this example, A is a 4×5 matrix and $R(A)$ does not span $\mathbb{R}^4$. Solving the system

$$\hat{A}\hat{x} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 2 & 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \\ \hat{x}_5 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}$$

we compute uniquely $\hat{x}_1$, $\hat{x}_2$ and $\hat{x}_3$ from the first three equations, and to have the fourth equation satisfied, we must have $b_4 = \hat{x}_1 + 3\hat{x}_2 + 2\hat{x}_3$ and thus b_4 can not be chosen freely.

92.12 Using Row Echelon Form to Obtain $N(A)$

We take the chance to solve the other basic problem of determining $N(A)$ by *reduction to row echelon form*, which is analogous to reduction to column echelon form working now with the rows instead of the columns. We thus consider a $m \times n$ matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix},$$

and perform the operations of (i) multiplying one row with a real number and (ii) multiplying one row with a real number and subtracting it from another row. We then obtain the *row echelon form* of A (possibly by reordering rows):

$$\hat{A} = \begin{pmatrix} 1 & \hat{a}_{12} & . & . & .. & . & \hat{a}_{1n} \\ 0 & 1 & . & .. & . & .. & \hat{a}_{2n} \\ . & . & . & . & . & .. & . \\ 0 & 0 & .. & 1 & . & .. & \hat{a}_{kn} \\ 0 & 0 & .. & 0 & 0 & .. & 0 \\ . & . & .. & . & . & .. & . \\ 0 & 0 & .. & 0 & 0 & .. & 0 \end{pmatrix}$$

Each non-zero row of the row echelon matrix $\hat{A}$ has one element equal to 1 and all elements to the left and below are equal to zero, and the 1s appear in a staircase form on or to the right of the diagonal from the upper left corner.

We notice that the row operations do not change the null space $N(A) = \{x : Ax = 0\}$, because we may perform the row operations in the system of equations $Ax = 0$, that is

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n &= 0, \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n &= 0, \\ &..... \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n &= 0, \end{aligned}$$

to reduce it to the echelon form system $\hat{A}x = 0$ without changing the vector $x = (x_1, ..., x_n)$. We conclude that

$$N(A) = N(\hat{A})$$

and we may thus determine $N(A)$ by using that we can directly determine $N(\hat{A})$ from the echelon form of A. It is easy to see that the dimension of $N(A) = N(\hat{A})$ is equal to $n - k$, as illustrated in the following example. In the case $n = m$, we have that $N(A) = 0$ if and only if $k = m = n$ in which all diagonal elements of $\hat{A}$ are equal to 1.

We give an example showing the sequence of matrices appearing in reduction to row echelon form:

EXAMPLE 92.5. We have

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 7 \\ 1 & 3 & 4 & 5 & 8 \\ 1 & 4 & 5 & 6 & 9 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \\ 0 & 2 & 3 & 4 & 7 \\ 0 & 3 & 4 & 5 & 8 \end{pmatrix} \rightarrow$$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \\ 0 & 0 & -1 & -2 & -5 \\ 0 & 0 & -2 & -4 & -10 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \\ 0 & 0 & 1 & 2 & 5 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \hat{A}.$$

We now determine $N(A)$ by determining $N(\hat{A}) = N(A)$ by seeking the solutions $x = (x_1, ..., x_5)$ of the system $\hat{A}x = 0$, that is

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \\ 0 & 0 & 1 & 2 & 5 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

We see that we can freely choose x_4 and x_5 and then solve for x_3, x_2 and x_1 to get the solution in the form

$$x = \lambda_1 \begin{pmatrix} 0 \\ 1 \\ -2 \\ 1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 \\ 4 \\ -5 \\ 0 \\ 1 \end{pmatrix}$$

where λ_1 and λ_2 are any real numbers. We have now computed a basis for $N(A)$ and we see in particular that the dimension of $N(A)$ is equal to 2. We recall that the dimension of $R(A)$ is equal to 3 and we note that the sum of the dimensions of $R(A)$ and $N(A)$ happens to be equal to 5 that is the number of columns of A. This is a general fact which we prove in the Fundamental Theorem below.

92.13 Gaussian Elimination

Gaussian elimination to compute solutions to the system

$$\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ . \\ . \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ . \\ . \\ b_m \end{pmatrix}$$

closely couples to reduction to row echelon form. Performing row operations we may reduce to a system of the form

$$\hat{A} = \begin{pmatrix} 1 & \hat{a}_{12} & . & . & .. & . & \hat{a}_{1n} \\ 0 & 1 & . & .. & . & .. & \hat{a}_{2n} \\ . & . & . & . & . & .. & . \\ 0 & 0 & .. & 1 & . & .. & \hat{a}_{kn} \\ 0 & 0 & .. & 0 & 0 & .. & 0 \\ . & . & .. & . & . & .. & . \\ 0 & 0 & .. & 0 & 0 & .. & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ . \\ . \\ x_n \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ . \\ . \\ \hat{b}_m \end{pmatrix}$$

with the same solution vector x. We may assume, by possibly by renumbering the components of x, that the 1s appear on the diagonal. We see

that solvability is equivalent to having $\hat{b}_j = 0$ for $j = k+1, .., m$, and the non-uniqueness is expressed by $N(A)$ as explained above. In the case $m = n$ we have that $N(A) = 0$ if and only if $k = m = n$ in which case all diagonal elements of $\hat{A}$ are equal to 1, and the system $\hat{A}x = \hat{b}$ is uniquely solvable for all $\hat{b} \in \mathbb{R}^m$, and thus $Ax = b$ is uniquely solvable for all $b \in \mathbb{R}^m$. We conclude that if $m = n$, then uniqueness implies existence. We may thus say that by Gaussian elimination or reduction to row echelon form, we may solve our basic problems of existence and uniqueness of solutions to the system $Ax = b$. We shall add more information on these problems in the Fundamental Theorem of Linear Algebra below. For more information on Gaussian elimination, we refer to Chapter *Solving Linear Algebraic Systems* below.

92.14 A Basis for $\mathbb{R}^n$ Contains n Vectors

Let us now prove that if $\{a_1, ..., a_m\}$ is a basis for $\mathbb{R}^n$, then $m = n$, that is any basis for $\mathbb{R}^n$ has exactly n elements, no more no less. We already deduced this fact from the column echelon form above, but we here give a "coordinate-free" proof which applies to more general situations.

We recall that a set $\{a_1, ..., a_m\}$ of vectors in $\mathbb{R}^n$ is a *basis* for $\mathbb{R}^n$ if the following two conditions are fulfilled:

- $\{a_1, ..., a_m\}$ is linearly independent,
- any vector $x \in \mathbb{R}^n$ can be expressed as a linear combination $x = \sum_{j=1}^m \lambda_j a_j$ of $\{a_1, ..., a_m\}$ with coefficients λ_j.

Of course, $\{e_1, ..., e_n\}$ is a basis for $\mathbb{R}^n$ in this sense.

To prove that $m = n$, we consider the set $\{e_1, a_1, a_2, ..., a_m\}$. Since $\{a_1, ..., a_m\}$ is a basis for $\mathbb{R}^n$, that is spans $\mathbb{R}^n$, the vector e_1 can be expressed as a linear combination of $\{a_1, ..., a_m\}$:

$$e_1 = \sum_{j=1}^m \lambda_j a_j,$$

with some $\lambda_j \neq 0$. Suppose $\lambda_1 \neq 0$. Then, dividing by λ_1 expresses a_1 as a linear combination of $\{e_1, a_2, ..., a_m\}$. This means that $\{e_1, a_2, ..., a_m\}$ spans $\mathbb{R}^n$. Consider now the set $\{e_1, e_2, a_2, ..., a_m\}$. The vector e_2 can be expressed as a linear combination of $\{e_1, a_2, ..., a_m\}$ and some of the coefficients of the a_j must be non-zero, since $\{e_1, e_2\}$ are linearly independent. Supposing the coefficient of a_2 is non-zero, we can eliminate a_2 and thus the set $\{e_1, e_2, a_3, ..., a_m\}$ now spans $\mathbb{R}^n$. Continuing this way we get the set $\{e_1, e_2, ..., e_n, a_{n+1}, ..., a_m\}$ if $m > n$ and the set $\{e_1, e_2, ..., e_n\}$ if $m = n$, which both span $\mathbb{R}^n$. We conclude that $m \geq n$, since if e.g. $m = n - 1$,

we would end up with the set $\{e_1, e_2, ..., e_{n-1}\}$ which does not span $\mathbb{R}^n$ contrary to the assumption.

Repeating this argument with the roles of the basis $\{e_1, e_2, ..., e_n\}$ and $\{a_1, a_2, ..., a_m\}$ interchanged, we get the reverse inequality $n \geq m$ and thus $n = m$. Of course, intuitively, there are n independent directions in $\mathbb{R}^n$ and thus a basis of $\mathbb{R}^n$ has n elements, no more no less.

We also note that if $\{a_1, ..., a_m\}$ is a linearly independent set in $\mathbb{R}^n$, then it can be extended to a basis $\{a_1, ..., a_m, a_{m+1}, .., a_n\}$ by adding suitable elements a_{m+1},..,a_n. The extension starts by adding a_{m+1} as any vector which cannot be expressed as a linear combination of the set $\{a_1, ..., a_m\}$. Then $\{a_1, ..., a_m, a_{m+1}\}$ is linearly independent, and if $m + 1 < n$, the process may be continued.

We summarize as follows:

Theorem 92.1 *Any basis of $\mathbb{R}^n$ has n elements. Further, a set of n vectors in $\mathbb{R}^n$ span $\mathbb{R}^n$ if and only if it is linearly independent, that is a set of n vectors in $\mathbb{R}^n$ that spans $\mathbb{R}^n$ or is independent, must be a basis. Also, a set of fewer than n vectors in $\mathbb{R}^n$ cannot span $\mathbb{R}^n$, and a set of more than n vectors in $\mathbb{R}^n$ must be linearly dependent.*

The argument used to prove this result can also be used to prove that the dimension of a vector space S is well defined in the sense that any two bases have the same number of elements.

92.15 Coordinates in Different Bases

There are many different bases in $\mathbb{R}^n$ if $n > 1$ and the coordinates of a vector with respect to one basis are not equal to the coordinates with respect to another basis.

Suppose $\{a_1, ..., a_n\}$ is a basis for $\mathbb{R}^n$ and let us seek the connection between the coordinates of one and the same vector in the standard basis $\{e_1, ..., e_n\}$ and the basis $\{a_1, ..., a_n\}$. Assume then that the coordinates of the basis vectors a_j in the standard basis $\{e_1, ..., e_n\}$ are given by $a_j = (a_{1j}, ...a_{nj})$ for $j = 1, ...n$, that is

$$a_j = \sum_{i=1}^{n} a_{ij} e_i.$$

Denoting the coordinates of a vector x with respect to $\{e_1, ..., e_n\}$ by x_j and the coordinates with respect to $\{a_1, ..., a_n\}$ by $\hat{x}_j$, we have

$$x = \sum_{j=1}^{n} \hat{x}_j a_j = \sum_{j=1}^{n} \hat{x}_j \sum_{i=1}^{n} a_{ij} e_i = \sum_{i=1}^{n} (\sum_{j=1}^{n} a_{ij} \hat{x}_j) e_i, \tag{92.22}$$

that is since also $x = \sum_{i=1}^n x_i e_i$ and the coefficients x_i of x are unique,

$$x_i = \sum_{j=1}^n a_{ij}\hat{x}_j \quad \text{for } i = 1, ...n. \tag{92.23}$$

This relation expresses the connection between the coordinates $\hat{x}_j$ with respect to the basis $\{a_1, ..., a_n\}$, and the coordinates x_i with respect to the standard basis $\{e_1, ..., e_n\}$, in terms of the coordinates a_{ij} of the basis vectors a_j with respect to $\{e_1, ..., e_n\}$. This is a basic connection, which will play a central role in the development to come.

Using the scalar product we can express the coordinates a_{ij} of the basis vector a_j as $a_{ij} = (e_i, a_j)$. To find the connection (92.23) between the coordinates $\hat{x}_j$ with respect to the basis $\{a_1, ..., a_n\}$, and the coordinates x_i with respect to the standard basis $\{e_1, ..., e_n\}$, we may start from the equality $\sum_{j=1}^n x_j e_j = x = \sum_{j=1}^n \hat{x}_j a_j$ and take the scalar product of both sides with e_i, to get

$$x_i = \sum_{j=1}^n \hat{x}_j (e_i, a_j) = \sum_{j=1}^n a_{ij}\hat{x}_j, \tag{92.24}$$

where $a_{ij} = (e_i, a_j)$.

EXAMPLE 92.6. The set $\{a_1, a_2, a_3\}$ with $a_1 = (1, 0, 0)$, $a_2 = (1, 1, 0)$, $a_3 = (1, 1, 1)$ in the standard basis, forms a basis for $\mathbb{R}^3$ since the set $\{a_1, a_2, a_3\}$ is linearly independent. This is because if $\lambda_1 a_1 + \lambda_2 a_2 + \lambda_3 a_3 = 0$, then $\lambda_3 = 0$ and thus also $\lambda_2 = 0$ and thus also $\lambda_1 = 0$. If (x_1, x_2, x_3) are the coordinates with respect to the standard basis and $(\hat{x}_1, \hat{x}_2, \hat{x}_3)$ are the coordinates with respect to $\{a_1, a_2, a_3\}$ of a certain vector, then the connection between the coordinates is given by $(x_1, x_2, x_3) = \hat{x}_1 a_1 + \hat{x}_2 a_2 + \hat{x}_3 a_3 = (\hat{x}_1 + \hat{x}_2 + \hat{x}_3, \hat{x}_2 + \hat{x}_3, \hat{x}_3)$. Solving for the $\hat{x}_j$ in terms of the x_i, we get $(\hat{x}_1, \hat{x}_2, \hat{x}_3) = (x_1 - x_2, x_2 - x_3, x_3)$.

92.16 Linear Functions $f : \mathbb{R}^n \to \mathbb{R}$

A *linear function* $f : \mathbb{R}^n \to \mathbb{R}$ satisfies

$$f(x+y) = f(x) + f(y),\ f(\alpha x) = \alpha f(x) \quad \text{for all } x, y \in \mathbb{R}^n,\ \alpha \in \mathbb{R}. \tag{92.25}$$

We say that $f(x)$ is a *scalar linear function* since $f(x) \in \mathbb{R}$. Expressing $x = x_1 e_1 + ... + x_n e_n$ in the standard basis $\{e_1, ..., e_n\}$, and using the linearity of $f(x)$, we find that

$$f(x) = x_1 f(e_1) + ... + x_n f(e_n), \tag{92.26}$$

and thus $f(x)$ has the form

$$f(x) = f(x_1, ..., x_n) \quad = \quad a_1x_1 + a_2x_2 + ... + a_nx_n, \tag{92.27}$$

where the $a_j = f(e_j)$ are real numbers. We can write $f(x)$ as

$$f(x) = (a, x) = a \cdot x, \tag{92.28}$$

where $a = (a_1, ..., a_n) \in \mathbb{R}^n$, that is $f(x)$ can be expressed as the scalar product of x with the vector $a \in \mathbb{R}^n$ with components a_j given by $a_j = f(e_j)$.

The set of scalar linear functions is the mother of all other functions. We now generalize to systems of scalar linear functions. Linear algebra is the study of systems of linear functions.

EXAMPLE 92.7. $f(x) = 2x_1 + 3x_2 - 7x_3$ defines a linear function $f : \mathbb{R}^3 \to \mathbb{R}$ with coefficients $f(e_1) = a_1 = 2$, $f(e_2) = a_2 = 3$ and $f(e_3) = a_3 = -7$.

92.17 Linear Transformations $f : \mathbb{R}^n \to \mathbb{R}^m$

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *linear* if

$$f(x+y) = f(x)+f(y),\ f(\alpha x) = \alpha f(x) \quad \text{for all } x, y \in \mathbb{R}^n,\ \alpha \in \mathbb{R}. \tag{92.29}$$

We also refer to a linear function $f : \mathbb{R}^n \to \mathbb{R}^m$ as a *linear transformation* of $\mathbb{R}^n$ into $\mathbb{R}^m$.

The image $f(x)$ of $x \in \mathbb{R}^n$ is a vector in $\mathbb{R}^m$ with components which we denote by $f_i(x)$, $i = 1, 2, ..., m$, so that $f(x) = (f_1(x), ..., f_m(x))$. Each *coordinate function* $f_i(x)$ is a linear scalar function $f_i : \mathbb{R}^n \to \mathbb{R}$ if $f : \mathbb{R}^n \to \mathbb{R}^m$ is linear. We can thus represent a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ as

$$\begin{aligned} f_1(x) &= a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n \\ f_2(x) &= a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n \\ &..... \\ f_m(x) &= a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n \end{aligned} \tag{92.30}$$

with the coefficients $a_{ij} = f_i(e_j) = (e_i, f(e_j)) \in \mathbb{R}$.

We can write (92.30) in condensed form as

$$f_i(x) = \sum_{j=1}^{n} a_{ij}x_j \quad \text{for } i = 1, ..., m. \tag{92.31}$$

EXAMPLE 92.8. $f(x) = (2x_1 + 3x_2 - 7x_3, x_1 + x_3)$ defines a linear function $f : \mathbb{R}^3 \to \mathbb{R}^2$ with coefficients $f_1(e_1) = a_{11} = 2$, $f_1(e_2) = a_{12} = 3$ and $f_1(e_3) = a_{13} = -7$, $f_2(e_1)a_{21} = 1$, $f_2(e_2)a_{22} = 0$ and $f_2(e_3) = a_{23} = 1$.

92.18 Matrices

We now return to the notion of a *matrix* and develop a matric calculus. The connection to linear transformations is very important. We define the $m \times n$ *matrix* $A = (a_{ij})$ as the rectangular array

$$\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix} \tag{92.32}$$

with rows $(a_{i1}, ..., a_{in})$, $i = 1, ..., m$, and columns $(a_{1j}, ..., a_{mj})$, $j = 1, ..., n$, where $a_{ij} \in \mathbb{R}$.

We may view each row $(a_{i1}, ..., a_{in})$ as a n-row vector or as a $1 \times n$ matrix, and each column $(a_{1j}, ..., a_{mj})$ as an m-column vector or a $m \times 1$ matrix. We can thus view the $m \times n$ matrix $A = (a_{ij})$ with elements a_{ij}, as consisting of m row vectors $(a_{i1}, ..., a_{in})$, $i = 1, ..., m$ or n column vectors $(a_{1j}, ..., a_{mj})$, $j = 1, ..., n$.

92.19 Matrix Calculus

Let $A = (a_{ij})$ and $B = (b_{ij})$ be two $m \times n$ matrices. We define $C = A + B$ as the $m \times n$ matrix $C = (c_{ij})$ with elements

$$c_{ij} = a_{ij} + b_{ij}, \quad i = 1, ..., n,\ j = 1, ..., m. \tag{92.33}$$

We may thus add two $m \times n$ matrices by adding the corresponding elements.

Similarly, we define for λ a real number the matrix λA with elements (λa_{ij}), corresponding to multiplying all elements of A by the real number λ.

We shall now define matrix multiplication and we start by defining the product Ax of an $m \times n$ matrix $A = (a_{ij})$ with a $n \times 1$ column vector $x = (x_j)$ as the $m \times 1$ column vector $y = Ax$ with elements $y_i = (Ax)_i$ given by

$$(Ax)_i = \sum_{j=1}^{n} a_{ij} x_j, \tag{92.34}$$

or with matrix notation

$$\begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_m \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & & & \\ a_{m1} & a_{m2} & ... & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{pmatrix}.$$

The element $y_i = (Ax)_i$ of the matrix-vector product Ax is thus obtained by taking the scalar product of row i of A with the vector x, as expressed by (92.34).

We can now express a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ as a matrix-vector product

$$f(x) = Ax,$$

where $A = (a_{ij})$ is an $m \times n$ matrix with elements $a_{ij} = f_i(e_j) = (e_i, f(e_j))$, where $f(x) = (f_1(x), ..., f_m(x))$. This is a restatement of (92.31).

We now proceed to define the product of an $m \times p$ matrix $A = (a_{ij})$ with a $p \times n$ matrix $B = (b_{ij})$. We do this by connecting the matrix product to the composition $f \circ g : \mathbb{R}^n \to \mathbb{R}^m$ given by

$$f \circ g(x) = f(g(x)) = f(Bx) = A(Bx), \tag{92.35}$$

where $f : \mathbb{R}^p \to \mathbb{R}^m$ is the linear transformation given by $f(y) = Ay$, where $A = (a_{ij})$ and $a_{ik} = f_i(e_k)$, and $g : \mathbb{R}^n \to \mathbb{R}^p$ is the linear transformation given by $g(x) = Bx$, where $B = (b_{kj})$ and $b_{kj} = g_k(e_j)$. Here e_k denote the standard basis vectors in $\mathbb{R}^p$, and e_j the corresponding basis vectors in $\mathbb{R}^n$. Clearly, $f \circ g : \mathbb{R}^n \to \mathbb{R}^m$ is linear and thus can be represented by an $m \times n$ matrix. Letting $(f \circ g)_i(x))$ denote the components of $(f \circ g)(x)$, we have

$$(f \circ g)_i(e_j) = f_i(g(e_j)) = f_i(\sum_{k=1}^{p} b_{kj}e_k) = \sum_{k=1}^{p} b_{kj}f_i(e_k) = \sum_{k=1}^{p} a_{ik}b_{kj},$$

which shows that $f \circ g(x) = Cx$, where $C = (c_{ij})$ is the $m \times n$ matrix with elements c_{ij} given by the formula

$$c_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj}, \quad i = 1, ..., m,\ j = 1, ..., n. \tag{92.36}$$

We conclude that $A(Bx) = Cx$, and we are thus led to define the matrix product $AB = C$ by (92.36), where thus A is an $m \times p$ matrix and B is a $p \times n$ matrix, and the product AB is a $m \times n$ matrix. We can then write

$$A(Bx) = ABx$$

as a reflection of $f(g(x)) = f \circ g(x)$.

Formally, to get the $m \times n$ format of the product AB, we cancel the p in the $m \times p$ format of A and the $p \times n$ format of B. We see that the formula (92.36) may be expressed as follows: the element c_{ij} in row i and column j of AB is obtained by taking the scalar product of row i of A with column j of B.

We may write the formula for matrix multiplication as follows:

$$(AB)_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj}, \quad \text{for } i = 1, .., n, j = 1, .., m, \tag{92.37}$$

or with matrix notation

$$AB = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1p} \\ a_{21} & a_{22} & ... & a_{2p} \\ ... & & & \\ a_{m1} & a_{m2} & ... & a_{mp} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & ... & b_{1n} \\ b_{21} & b_{22} & ... & b_{2n} \\ ... & & & \\ b_{p1} & b_{p2} & ... & b_{pn} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{k=1}^p a_{1k}b_{k1} & \sum_{k=1}^p a_{1k}b_{k2} & ... & \sum_{k=1}^p a_{1k}b_{kn} \\ \sum_{k=1}^p a_{2k}b_{k1} & \sum_{k=1}^p a_{2k}b_{k2} & ... & \sum_{k=1}^p a_{2k}b_{kn} \\ ... & & & \\ \sum_{k=1}^p a_{mk}b_{k1} & \sum_{k=1}^p a_{mk}b_{k2} & ... & \sum_{k=1}^p a_{mk}b_{kn} \end{pmatrix}.$$

Matrix multiplication is not commutative, that is $AB \neq BA$ in general. In particular, BA is defined only if $n = m$.

As a special case, we have that the product Ax of an $m \times n$ matrix A with an $n \times 1$ matrix x is given by (92.34). We may thus view the matrix-vector product Ax defined by (92.34) as a special case of the matrix product (92.36) with the $n \times 1$ matrix x being a column vector. The vector Ax is obtained taking the scalar product of the rows of A with the column vector x.

We sum up in the following theorem.

Theorem 92.2 *A linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ can be expressed as*

$$f(x) = Ax, \tag{92.38}$$

where $A = (a_{ij})$ is an $m \times n$ matrix with elements $a_{ij} = f_i(e_j) = (e_i, f(e_j))$, where $f(x) = (f_1(x), ..., f_m(x))$. If $g : \mathbb{R}^n \to \mathbb{R}^p$ and $f : \mathbb{R}^p \to \mathbb{R}^m$ are two linear transformations with corresponding matrices A and B, then the matrix of $f \circ g : \mathbb{R}^n \to \mathbb{R}^m$ is given by AB.

92.20 The Transpose of a Linear Transformation

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation defined by $f(x) = Ax$, where $A = (a_{ij})$ is an $m \times n$-matrix. We now define another linear transformation $f^\top : \mathbb{R}^m \to \mathbb{R}^n$, which we refer to as the *transpose* of f, by the relation

$$(x, f^\top(y)) = (f(x), y) \quad \text{for all } x \in \mathbb{R}^n,\ y \in \mathbb{R}^m. \tag{92.39}$$

Using that $f(x) = Ax$, we have

$$(f(x), y) = (Ax, y) = \sum_{i=1}^m \sum_{j=1}^n a_{ij}x_j y_i, \tag{92.40}$$

and thus setting $x = e_j$, we see that

$$(f^\top(y))_j = \sum_{i=1}^m a_{ij}y_i. \tag{92.41}$$

This shows that $f^\top(y) = A^\top y$, where $A^\top$ is the $n \times m$ matrix with elements $(a_{ji}^\top)$ given by $a_{ji}^\top = a_{ij}$. In other words, the columns of $A^\top$ are the rows of A and vice versa. For example, if

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \quad \text{then} \quad A^\top = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}. \tag{92.42}$$

Summing up we have:

Theorem 92.3 *If $A = (a_{ij})$ is a $m \times n$ matrix, then the transpose $A^\top$ is an $n \times m$ matrix with elements $a_{ji}^\top = a_{ij}$, and*

$$(Ax, y) = (x, A^\top y) \quad \textit{for all } x \in \mathbb{R}^n,\ y \in \mathbb{R}^m. \tag{92.43}$$

An $n \times n$ matrix such that $A^\top = A$, that is $a_{ij} = a_{ji}$ for $i, j = 1, ...n$, is said to be a *symmetric* matrix.

92.21 Matrix Norms

In many situations we need to estimate the "size" of a $m \times n$ matrix $A = (a_{ij})$. We may use this information to estimate the "length" of $y = Ax$ in terms of the "length" of x. We observe that

$$\sum_{i=1}^m |y_i| \le \sum_{i=1}^m \sum_{j=1}^n |a_{ij}||x_j| = \sum_{j=1}^n \sum_{i=1}^m |a_{ij}||x_j| \le \max_{j=1,\ldots,n} \sum_{i=1}^m |a_{ij}| \sum_{j=1}^n |x_j|,$$

which shows that if we define $\|x\|_1 = \sum |x_j|$ and $\|y\|_1 = \sum |y_i|$, then

$$\|y\|_1 \le \|A\|_1 \|x\|_1$$

if we define

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^m |a_{ij}|$$

Similarly, we have

$$\max_i |y_i| \le \max_i \sum_{j=1}^n |a_{ij}|x_j| \le \max_i^n \sum_{j=1}^n |a_{ij}| \max_j |x_j|$$

which shows that if we define $\|x\|_\infty = \max_j |x_j|$ and $\|y\|_\infty = \max_i |y_i|$, then

$$\|y\|_\infty \le \|A\|_\infty \|x\|_\infty$$

if we define

$$\|A\|_\infty = \max_{i+1,\ldots,m} \sum_{j=1}^n |a_{ij}|.$$

We may also define the *Euclidean norm* $\|A\|$ by

$$\|A\| = \max_{x\in\mathbb{R}^n} \frac{\|Ax\|}{\|x\|}, \tag{92.44}$$

where we maximize over $x \neq 0$, and $\|\cdot\|$ denotes the Euclidean norm. We thus define $\|A\|$ to be the smallest constant C such that $\|Ax\| \leq C\|x\|$ for all $x \in \mathbb{R}^n$. We shall return in Chapter *The Spectral theorem* below to the problem of giving a formula for $\|A\|$ in terms of the coefficients of A in the case A is symmetric (with in particular $m = n$). By definition, we clearly have

$$\|Ax\| \leq \|A\|\,\|x\|. \tag{92.45}$$

If $A = (\lambda_i)$ is a diagonal $n \times n$ matrix with diagonal elements $a_{ii} = \lambda_i$, then

$$\|A\| = \max_{i=1,\ldots,n} |\lambda_i|. \tag{92.46}$$

92.22 The Lipschitz Constant of a Linear Transformation

Consider a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ given by a $m \times n$ matrix $A = (a_{ij})$, that is

$$f(x) = Ax, \qquad \text{for } x \in \mathbb{R}^n.$$

By linearity we have

$$\|f(x) - f(y)\| = \|Ax - Ay\| = \|A(x - y)\| \leq \|A\|\|x - y\|.$$

We may thus say that the Lipschitz constant of $f : \mathbb{R}^n \to \mathbb{R}^m$ is equal to $\|A\|$. Alternatively, working with the norms $\|\cdot\|_1$ or $\|\cdot\|_\infty$, we may view the Lipschitz constant to be equal to $\|A\|_1$ or $\|A\|_\infty$.

92.23 Volume in $\mathbb{R}^n$: Determinants and Permutations

Let $\{a_1, a_2, ..., a_n\}$ be a set of n vectors in $\mathbb{R}^n$. We shall now generalize the concept of *volume* $V(a_1, ..., a_n)$ spanned by $\{a_1, a_2, ..., a_n\}$, which we have met above in the case $n = 2$ and $n = 3$. In particular, the volume will give us a tool to determine whether the set of vectors $\{a_1, .., a_n\}$ is linearly independent or not. Using the determinant we shall also develop Cramer's solution formula for an $n \times n$ system $Ax = b$, which generalizes the solution formulas for 2×2 and 3×3 which we have already met. The determinant is a quite complicated object, and we try to make the presentation as accessible

as possible. When it comes to computing determinants. we shall return to the column echelon form.

Before actually giving a formula for the volume $V(a_1, ..., a_n)$ in terms of the coordinates $(a_{1j}, ..., a_{nj})$ of the vectors a_j, $j = 1, 2, ..., n$, we note that from our experience in $\mathbb{R}^2$ and $\mathbb{R}^3$, we expect $V(a_1, ..., a_n)$ to be a *multilinear alternating form*, that is

$$\begin{aligned} V(a_1, ..., a_n) &\in \mathbb{R}, \\ V(a_1, ..., a_n) &\quad \text{is linear in each argument } a_j, \\ V(a_1, ..., a_n) &= -V(\hat{a}_1, ..., \hat{a}_n), \end{aligned}$$

where $\hat{a}_1, ..., \hat{a}_n$ is a listing of $a_1, ..., a_n$ with two of the a_j interchanged. For example $\hat{a}_1 = a_2$, $\hat{a}_2 = a_1$ and $\hat{a}_j = a_j$ for $j = 3, ..., n$. We note that if two of the arguments in an alternating form is the same, for example $a_1 = a_2$, then $V(a_1, a_2, a_3, .., a_n) = 0$. This follows at once from the fact that $V(a_1, a_2, a_3, .., a_n) = -V(a_2, a_1, a_3, .., a_n)$. We are familiar with these properties in the case $n = 2, 3$.

We also need a little preliminary work on permutations. A *permutation* of the ordered list $\{1, 2, 3, 4, ..., n\}$ is a reordering of the list. For example $\{2, 1, 3, 4, ..., n\}$ is a permutation corresponding to interchanging the elements 1 and 2. Another permutation is $\{n, n-1, ..., 2, 1\}$ corresponding to reversing the order.

We can also describe a permutation as a one-to-one mapping of the set $\{1, 2, ..., n\}$ onto itself. We may denote the mapping by $\pi : \{1, 2, ..., n\} \to \{1, 2, ..., n\}$, that is $\pi(j)$ is one of the numbers $1, 2, ..., n$ for each $j = 1, 2, ..., n$ and $\pi(i) \neq \pi(j)$ if $i \neq j$. We can then talk about the *product* $\sigma\tau$ of two permutations σ and τ defined as the composition of τ and σ:

$$\sigma\tau(j) = \sigma(\tau(j)), \quad \text{for } j = 1, ..., n, \tag{92.47}$$

which is readily seen to be a permutation. Note that the order may be important: in general the permutation $\sigma\tau$ is different from the permutation $\tau\sigma$. In other words, multiplication of permutations is not commutative. However, multiplication is associative:

$$(\pi\sigma)\tau = \pi(\sigma\tau), \tag{92.48}$$

which directly follows from the definition by composition of functions.

A permutation corresponding to interchanging two elements, is called a *transposition*. More precisely, if π is a transposition then there are two elements p and q of the elements $\{1, 2, ..., n\}$, such that

$$\begin{aligned} \pi(p) &= q \\ \pi(q) &= p \\ \pi(j) &= j \quad \text{for } j \neq p,\ j \neq q. \end{aligned}$$

The permutation π with $\pi(j) = j$ for $j = 1, ..., n$ is called the identity permutation.

We shall use the following basic fact concerning permutations (a proof will be given in the Appendix).

Theorem 92.4 *Every permutation can be written as a product of transpositions. The representation is not unique, but for a given permutation the number of transpositions in such a representation cannot be odd in one case and even in another case; it is odd for all representations or even for all representations.*

We call a permutation *even* if it contains an even number of transposition factors, and *odd* if it contains an odd number of transpositions. The number of even perturbations is equal to the number of odd perturbations, and thus the total number of perturbations, including the identity, is even.

92.24 Definition of the Volume $V(a_1, ...a_n)$

Assuming that $V(a_1, ..., a_n)$ is multilinear and alternating and that $V(e_1, e_2, ..., e_n) = 1$, we get the following relation

$$\begin{aligned} V(a_1, ..., a_n) &= V(\sum_j a_{j1}e_j, \sum_j a_{j2}e_j, ..., \sum_j a_{jn}e_j) \\ &= \sum_\pi \pm a_{\pi(1)\,1}a_{\pi(2)\,2}\cdots a_{\pi(n)\,n}, \end{aligned} \tag{92.49}$$

where we sum over all permutations π of the set $\{1, ..., n\}$, and the sign indicates if the permutation is even (+) or odd (-). Note that we give the identity permutation, which is included among the permutations, the sign +. We recall that $a_j = (a_{1j}, ..., a_{nj})$ for $j = 1, ..., n$.

We now turn around in this game, and simply take (92.49) as a definition of the volume $V(a_1, ..., a_n)$ spanned by the set of vectors $\{a_1, ..., a_n\}$. By this definition it follows that $V(a_1, ..., a_n)$ is indeed a multilinear alternating form on $\mathbb{R}^n$. Further, $V(e_1, ..., e_n) = 1$, since the only non-zero term in the sum (92.49) in this case corresponds to the identity perturbation.

We can transform the definition $V(a_1, ..., a_n)$ to matrix language as follows. Let $A = (a_{ij})$ be the $n \times n$ matrix

$$\begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ . & . & ... & . \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix} \tag{92.50}$$

formed by the column vectors $a_1, ..., a_n$ with coefficients $a_j = (a_{1j}, ...a_{nj})$. We define the *determinant* det A of A, by

$$\det A = V(a_1, ..., a_n) = \sum_{\pi} \pm a_{\pi(1)\,1} a_{\pi(2)\,2} \cdots a_{\pi(n)\,n},$$

where we sum over all permutations π of the set $\{1, ..., n\}$, and the sign indicates if the permutation is even (+) or odd (-).

We note that since the unit vectors e_j in $\mathbb{R}^n$ are mapped by A into the column vectors a_j, that is since $Ae_j = a_j$, we have that A maps the unit n-cube in $\mathbb{R}^n$ onto the parallelepiped in $\mathbb{R}^n$ spanned by a_1,..., a_n. Since the volume of the n-cube is one and the volume of the the parallelepiped spanned by a_1,..., a_n is $V(a_1, ...a_n)$, the *volume scale* of the mapping $x \to Ax$ is equal to $V(a_1, ..., a_n)$.

92.25 The Volume $V(a_1, a_2)$ in $\mathbb{R}^2$

If A is the 2×2-matrix

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

then det $A = V(a_1, a_2)$ is given

$$\det A = V(a_1, a_2) = a_{11}a_{22} - a_{21}a_{12}. \tag{92.51}$$

of course, $a_1 = (a_{11}, a_{21})$ and $a_2 = (a_{12}, a_{22})$ are the column vectors of A.

92.26 The Volume $V(a_1, a_2, a_3)$ in $\mathbb{R}^3$

If A is the 3×3-matrix

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

then det $A = V(a_1, a_2, a_3)$ is given by

$$\begin{aligned} &\det A = V(a_1, a_2, a_3) = a_1 \cdot a_2 \times a_3 \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}). \end{aligned} \tag{92.52}$$

We see that we can express det A as

$$\begin{aligned} \det A &= a_{11} \det A_{11} - a_{12} \det A_{12} + a_{13}A_{13} \\ &= a_{11}V(\hat{a}_2, \hat{a}_3) - a_{12}V(\hat{a}_1, \hat{a}_3) + a_{13}V(\hat{a}_1, \hat{a}_2) \end{aligned} \tag{92.53}$$

where the A_{1j} are 2×2 matrices formed by cutting out the first row and j:th column of A, explicitly given by

$$A_{11} = \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} \quad A_{12} = \begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix} \quad A_{13} = \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}$$

and $\hat{a}_1 = (a_{21}, a_{31})$, $\hat{a}_2 = (a_{22}, a_{32})$, $\hat{a}_3 = (a_{23}, a_{33})$ are the 2-column vectors formed by cutting out the first element of the 3-columns a_j. We say that (92.53) is an *expansion* of the 3×3 matrix A in terms of the elements of the first row of A and the corresponding 2×2 matrices. The expansion formula follows by collecting all the terms with a_{11} as a factor, and all the terms with a_{12} as a factor and all the terms with a_{13} as a factor.

92.27 The Volume $V(a_1, a_2, a_3, a_4)$ in $\mathbb{R}^4$

Using the expansion formula we can compute the determinant $\det A = V(a_1, ..., a_4)$ of a 4×4 matrix $A = (a_{ij})$ with column vectors $a_j = (a_{1j}, ..., a_{4j})$ for $j = 1, 2, 3, 4$. We have

$$\begin{aligned} \det A = V(a_1, a_2, a_3, a_4) &= a_{11} V(\hat{a}_2, \hat{a}_3, \hat{a}_4) - a_{12} V(\hat{a}_1, \hat{a}_3, \hat{a}_4) \\ &+ a_{13} V(\hat{a}_1, \hat{a}_2, \hat{a}_4) - a_{14} V(\hat{a}_1, \hat{a}_2, \hat{a}_3), \end{aligned}$$

where the $\hat{a}_j$, $j = 1, 2, 3, 4$ are the 3-column vectors corresponding to cutting out the first coefficient of the a_j. We have now expressed the determinant of the 4×4 matrix A as a sum of determinants of 3×3 matrices with the first row of A as coefficients.

92.28 The Volume $V(a_1, ..., a_n)$ in $\mathbb{R}^n$

Iterating the row-expansion formula indicated above, we can compute the determinant of an arbitrary $n \times n$ matrix A. As an example we give the expansion formula for a 5×5 matrix $A = (a_{ij})$:

$$\begin{aligned} \det A = V(a_1, a_2, a_3, a_4, a_5) &= a_{11} V(\hat{a}_2, \hat{a}_3, \hat{a}_4, \hat{a}_5) - a_{12} V(\hat{a}_1, \hat{a}_3, \hat{a}_4, \hat{a}_5) \\ &+ a_{13} V(\hat{a}_1, \hat{a}_2, \hat{a}_4, \hat{a}_5) - a_{14} V(\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_5) + a_{15} V(\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4). \end{aligned}$$

Evidently, we can formulate the following a rule of sign for the term with the factor a_{ij}: choose the $+$ if $i + j$ is even and the $-$ if $i + j$ is odd. This rule generalizes to expansions with respect to any row of A.

92.29 The Determinant of a Triangular Matrix

Let $A = (a_{ij})$ be a *upper triangular* $n \times n$ matrix, that is $a_{ij} = 0$ for $i > j$. All elements a_{ij} of A below the diagonal are zero. In this case the only

non-zero term in the expression for $\det A$, is the product of the diagonal elements of A corresponding to the identity perturbation, so that

$$\det A = a_{11}a_{22}\cdots a_{nn}. \tag{92.54}$$

This formula also applies to a *lower triangular* $n \times n$ matrix $A = (a_{ij})$ with $a_{ij} = 0$ for $i < j$.

92.30 Using the Column Echelon Form to Compute det A

We now present a way to compute $\det A = V(a_1, ..., a_n)$, where the a_j are the columns of a $n \times n$ matrix $A = (a_{ij})$, based on reduction to column echelon form. We then use that the volume does not change if we subtract one column multiplied by a real number from another column, to obtain

$$\det A = V(a_1, a_2, ..., a_n) = V(\hat{a}_1, \hat{a}_2, \hat{a}_3, ..., \hat{a}_n)$$

where $\hat{a}_{ij} = 0$ if $j > i$, that is the corresponding matrix $\hat{A}$ is a lower triangular matrix. We then compute $V(\hat{a}_1, \hat{a}_2, \hat{a}_3, ..., \hat{a}_n)$ by multiplying the diagonal elements. As usual, if we meet a zero diagonal term we interchange columns until we meet a nonzero diagonal term, or if all diagonal terms appearing this way are zero, we proceed to modify the next row. At least one of the diagonal terms in the final triangular matrix will then be zero, and thus the determinant will be zero.

Example 92.9. We show the sequence of matrices in a concrete case:

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 4 & 6 \\ 3 & 4 & 6 \end{pmatrix} \quad \to \begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 4 \\ 3 & 1 & 3 \end{pmatrix} \quad \to \begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & 1 & 1 \end{pmatrix}$$

and conclude that $det\, A = 2$.

92.31 The Magic Formula $\det AB = \det A \det B$

Let A and B be two $n \times n$ matrices. We know that AB is the matrix of the composite transformation $f(g(x))$, where $f(y) = Ay$ and $g(x) = Bx$. The volume scale of the mapping $x \to Bx$ is equal to $\det B$ and the volume scale of the mapping $y \to Ay$ is $\det A$, and hence the volume scale of the mapping $x \to ABx$ is equal to $\det A \det B$. This proves that

$$\det AB = \det A \det B,$$

which is one of the corner stones of the calculus of determinants. The proof suggested is a "short proof" avoiding algebraic computations. One can also give a direct algebraic proof using suitable expansion formulas for the determinant.

92.32 Test of Linear Independence

To test the linear independence of a given set of n vectors $\{a_1, a_2, ..., a_n\}$ in $\mathbb{R}^n$, we can use the volume $V(a_1, a_2, ..., a_n)$. More precisely, we shall prove that $\{a_1, a_2, ..., a_n\}$ is linearly independent if and only if $V(a_1, a_2, ..., a_n) \neq 0$. First, we note that if $\{a_1, a_2, ..., a_n\}$ is linearly dependent, for example if $a_1 = \sum_{j=2}^n \lambda_j a_j$ is a linear combination of $\{a_2, ..., a_n\}$, then $V(a_1, a_2, ..., a_n) = \sum_{j=2}^n \lambda_j V(a_j, a_2, ..., a_n) = 0$, since each factor $V(a_j, a_2, ..., a_n)$ has two equal vectors.

Secondly, if $\{a_1, a_2, ..., a_n\}$ is linearly independent, i.e., $\{a_1, a_2, ..., a_n\}$ is a basis for $\mathbb{R}^n$, then we must have $V(a_1, ..., a_n) \neq 0$. We see this as follows. We express each e_j as a linear combination of the set $\{a_1, a_2, ..., a_n\}$, for example $e_1 = \sum \lambda_{1j} a_j$. We have, since V is multilinear and vanishes if two arguments are the same, and $V(a_{\pi(1)}, .., a_{\pi(n)}) = \pm V(a_1, .., a_n)$ for any permutation π, that

$$\begin{aligned} 1 = V(e_1, .., e_n) &= V(\sum_j \lambda_{1j} a_j, e_2, .., e_n) = \sum_j \lambda_{1j} V(a_j, e_2, .., e_n) \\ &= \sum_j \lambda_{1j} V(a_j, \sum_k \lambda_{2k} a_k, e_3, .., e_n) = ... = cV(a_1, .., a_n), \end{aligned} \tag{92.55}$$

for some constant c. We conclude that $V(a_1, ..., a_n) \neq 0$. We summarize as follows:

Theorem 92.5 *A set $\{a_1, a_2, ..., a_n\}$ of n vectors in $\mathbb{R}^n$ is linearly independent if and only if $V(a_1, ..., a_n) \neq 0$.*

We may restate this result in matrix language as follows: The columns of an $n \times n$-matrix A are linearly independent if and only if $\det A \neq 0$. We may thus sum up as follows:

Theorem 92.6 *Let A be a $n \times n$ matrix. Then the following statements are equivalent:*

- *The columns of A are linearly independent.*
- *If $Ax = 0$ then $x = 0$.*
- $\det A \neq 0$.

To test linear independence of the columns of a given matrix A we may thus compute $\det A$ and check if $\det A = 0$. We can also use this test in more quantitative form as follows: If $\det A$ is small then the columns are close to being linearly dependent and uniqueness is at risk!

A matrix A with $\det A = 0$ is called *singular*, while matrices with $\det A \neq 0$ are referred to as *non-singular*. Thus an $n \times n$-matrix is non-singular if and only if its columns are linearly independent. Again we can go to quantitative forms and say that a matrix A is close to singular if its determinant is close to zero. The dependence of the solution on the size of the determinant is clearly expressed in the next chapter.

92.33 Cramer's Solution for Non-Singular Systems

Consider again the $n \times n$ linear system of equations

$$Ax = b \tag{92.56}$$

or

$$\sum_{j=1}^{n} a_j x_j = b \tag{92.57}$$

where $A = (a_{ij})$ is an $n \times n$ matrix with columns $a_j = (a_{1j}, ..., a_{nj})$, $j = 1, ..., n$. Suppose that the columns a_j of A are linearly independent, or equivalently, that $\det A = V(a_1, ..., a_n) \neq 0$. We then know that (92.56) has a unique solution $x \in \mathbb{R}^n$ for any given $b \in \mathbb{R}^n$, we shall now seek a formula for the solution x in terms of b and the columns a_j of A.

Using the basic property of the volume function $V(g_1, ..., g_n)$ of a set $\{g_1, ..., g_n\}$ of n vectors g_i, in particular the property that $V(g_1, ..., g_n) = 0$ if any two of the g_i are equal, we obtain the following solution formula (Cramer's formula):

$$\begin{aligned} x_1 &= \frac{V(b, a_2, ..., a_n)}{V(a_1, a_2, ..., a_n)}, \\ &.... \\ x_n &= \frac{V(a_1, ..., a_{n-1}, b)}{V(a_1, a_2, ..., a_n)}. \end{aligned} \tag{92.58}$$

For example, to obtain the formula for x_1, use that

$$V(b, a_2, ..., a_n) = V(\sum_j a_j x_j, a_2, ..., a_n)$$

$$= \sum_{j=1}^{n} x_j V(a_j, a_2, ..., a_n) = x_1 V(a_1, a_2, .., a_n).$$

We summarize:

Theorem 92.7 *If A is a $n \times n$ non-singular matrix with $\det A \neq 0$, then the system of equations $Ax = b$ has a unique solution x for any $b \in \mathbb{R}^n$. The solution is given by Cramer's formula* (92.58).

A result like this was first derived by Leibniz and then by Gabriel Cramer (1704-1752) (who got a Ph.D. at the age of 18 with a thesis on the theory of sound) in *Introduction l'analyse des lignes courbes algbraique*. Throughout the book Cramer makes essentially no use of the Calculus in either Leibniz' or Newton's form, although he deals with such topics as tangents, maxima and minima, and curvature, and cites Maclaurin and Taylor in footnotes. One conjectures that he never accepted or mastered Calculus.

FIGURE 92.2. Gabriel Cramer: "I am friendly, good-humoured, pleasant in voice and appearance, and possess good memory, judgement and health".

Note that Cramer's solution formula for $Ax = b$ is very computationally demanding, and thus cannot be used for actually computing the solution unless n is small. To solve linear systems of equations other methods are used, like Gaussian elimination and iterative methods, see Chapter *Solving systems of linear equations*.

92.34 The Inverse Matrix

Let A be a non-singular $n \times n$ matrix with $V(a_1, ..., a_n) \neq 0$. Then $Ax = b$ can be solved uniquely for all $b \in \mathbb{R}^n$ according to Cramer's solution formula (92.58). Clearly, x depends linearly on b, and the solution x may be expressed as $A^{-1}b$, where A^{-1} is an $n \times n$ matrix which we refer to as the *inverse* of A. The j:th column of A^{-1} is the solution vector corresponding to choosing $b = e_j$. Cramer's formula thus gives the following formula for

the inverse A^{-1} of A:

$$A^{-1} = V(a_1,..,a_n)^{-1}\begin{pmatrix} V(e_1,a_2,..,a_n) & .. & V(a_1,..,a_{n-1},e_1) \\ . & .. & . \\ . & .. & . \\ V(e_n,a_2,..,a_n) & .. & V(a_1,..,a_{n-1},e_n) \end{pmatrix}. \quad (92.59)$$

The inverse matrix A^{-1} of A satisfies

$$A^{-1}A = AA^{-1} = I,$$

where I is the $n \times n$ identity matrix, with ones on the diagonal and zeros elsewhere.

Evidently, we can express the solution to $Ax = b$ in the form $x = A^{-1}b$ if A is a non-singular $n \times n$ matrix (by multiplying $Ax = b$ by A^{-1} from the left)..

92.35 Projection onto a Subspace

Let V be a subspace of $\mathbb{R}^n$ spanned by the linearly independent set of vectors $\{a_1,...,a_m\}$. In other words, $\{a_1,...,a_m\}$ is a basis for V. The projection Pv of a vector $v \in \mathbb{R}^n$ onto V is defined as the vector $Pv \in V$ satisfying the orthogonality relation

$$(v - Pv, w) = 0 \quad \text{for all vectors } w \in V, \quad (92.60)$$

or equivalently

$$(Pv, a_j) = (v, a_j) \quad \text{for } j = 1,..,m. \quad (92.61)$$

To see the equivalence, we note that (92.60) clearly implies (92.61). Conversely, any $w \in V$ is a linear combination of the form $\sum \mu_j a_j$, and multiplying (92.61) by μ_j and summing over j, we obtain $(Pv, \sum_j \mu_j a_j) = (v, \sum_j \mu_j a_j)$, which is (92.60) with $w = \sum_j \mu_j a_j$ as desired.

Expressing $Pv = \sum_{i=1}^m \lambda_i a_i$ in the basis $\{a_1,...,a_m\}$ for V, the orthogonality relation (92.61) corresponds to the $m \times m$ linear system of equations

$$\sum_{i=1}^m \lambda_i (a_i, a_j) = (v, a_j) \quad \text{for } j = 1, 2, ...m. \quad (92.62)$$

We shall now prove that this system admits a unique solution, which proves that the projection Pv of v onto V exists and is unique. By Theorem 92.6 it is enough to prove uniqueness. We thus assume that

$$\sum_{i=1}^m \lambda_i (a_i, a_j) = 0 \quad \text{for } j = 1, 2, ...m.$$

Multiplying by λ_j and summing we get

$$0 = (\sum_{i=1}^m \lambda_i a_i, \sum_{j=1}^m \lambda_j a_j) = |\sum_{i=1}^m \lambda_i a_i|^2,$$

which proves that $\sum_i \lambda_i a_i = 0$ and thus $\lambda_i = 0$ for $i = 1, ..., m$, since the $\{a_1, ..., a_m\}$ is linearly independent.

We have now proved the following fundamental result:

Theorem 92.8 *Let V be a linear subspace of $\mathbb{R}^n$. Then for all $v \in \mathbb{R}^n$ the projection Pv of v onto V, defined by $Pv \in V$ and $(v - Pv, w) = 0$ for all $w \in V$, exists and is unique.*

We note that $P : \mathbb{R}^n \to V$ is a linear mapping. To see this, let v and $\hat{v}$ be two vectors in $\mathbb{R}^n$, and note that since $(v - Pv, w) = 0$ and $(\hat{v} - P\hat{v}, w) = 0$ for all $w \in V$, we have

$$(v + \hat{v} - (Pv + P\hat{v}), w) = (v - Pv, w) + (\hat{v} - P\hat{v}, w) = 0,$$

which shows that $Pv + P\hat{v} = P(v + \hat{v})$. Similarly, $Pw = \lambda Pv$ if $w = \lambda v$, for any $\lambda \in \mathbb{R}$ and $v \in \mathbb{R}^n$. This proves the linearity of $P : \mathbb{R}^n \to V$.

We further note that $PP = P$. We sum up as follows:

Theorem 92.9 *The projection $P : \mathbb{R}^n \to V$ onto a linear subspace V of $\mathbb{R}^n$ is a linear transformation defined by $(v - Pv, w) = 0$ for all $w \in V$, which satisfies $PP = P$.*

92.36 An Equivalent Characterization of the Projection

We shall now prove that the *projection* Pv of a vector $v \in \mathbb{R}^n$ onto V is the vector $Pv \in V$ with minimum distance to v, that is $|v - Pv| \leq |v - w|$ for all $w \in V$.

We state the equivalence of the two definitions of the projection in the following fundamental theorem:

Theorem 92.10 . *Let $v \in \mathbb{R}^n$ be given. The vector $Pv \in V$ satisfies the orthogonality relation*

$$(v - Pv, w) = 0 \quad \text{for all vectors } w \in V, \tag{92.63}$$

if and only if Pv minimizes the distance to v in the sense that

$$|v - Pv| \leq |v - w| \quad \text{for all } w \in V. \tag{92.64}$$

Further, the element $Pv \in V$ satisfying (92.63) and (92.64) is uniquely determined.

To prove the theorem we note that by the orthogonality (92.60), we have for any $w \in V$,

$$\begin{aligned}|v - Pv|^2 &= (v - Pv, v - Pv)\\ &= (v - Pv, v - w) + (v - Pv, w - Pv) = (v - Pv, v - w),\end{aligned}$$

since $w - Pv \in V$. Using Cauchy-Schwarz inequality, we obtain

$$|v - Pv|^2 \le |v - Pv|\,|v - w|,$$

which shows that $|v - Pv| \le |v - w|$ for all $w \in V$.

Conversely, if $|v - Pv| \le |v - w|$ for all $w \in V$, then for all $\epsilon \in \mathbb{R}$ and $w \in V$

$$\begin{aligned}|v - Pv|^2 &\le |v - Pv + \epsilon w|^2\\ &= |v - Pv|^2 + \epsilon(v - Pv, w) + \epsilon^2|w|^2,\end{aligned}$$

that is for all $\epsilon > 0$

$$(v - Pv, w) + \epsilon|w|^2 \ge 0,$$

which proves that

$$(v - Pv, w) \ge 0 \quad \text{for all } w \in V.$$

Changing w to $-w$ proves the reverse inequality and we conclude that $(v - Pv, w) = 0$ for all $w \in V$.

Finally, to prove uniqueness, assume that $z \in V$ satisfies

$$(v - z, w) = 0 \quad \text{for all vectors } w \in V.$$

Then $(Pv - z, w) = (Pv - v, w) + (v - z, w) = 0 + 0 = 0$ for all $w \in V$, and $Pv - z$ is a vector in V. Choosing $w = Pv - z$ thus shows that $|Pv - z|^2 = 0$, that is $z = Pv$. The proof of the theorem is now complete.

The argument just given is very fundamental and will be used many times below in various forms, so it is worth taking the time to understand it now.

92.37 Orthogonal Decomposition: Pythagoras Theorem

Let V be a subspace of $\mathbb{R}^n$. Let P be the projection onto V. Any vector x can be written

$$x = Px + (x - Px) \tag{92.65}$$

where $Px \in V$, and further $(x - Px) \perp V$ since by the definition of P we have $(x - Px, w) = 0$ for all $w \in V$. We say that $x = Px + (x - Px)$ is an *orthogonal decomposition* of x since $(Px, x - Px) = 0$.

Define the *orthogonal complement* $V^\perp$ to V by $V^\perp = \{y \in \mathbb{R}^n : y \perp V\} = \{y \in \mathbb{R}^n : y \perp x \quad \text{for all } x \in V\}$. It is clear that $V^\perp$ is a linear subspace of $\mathbb{R}^n$. We have that if $x \in V$ and $y \in V^\perp$, then $(x, y) = 0$. Further, any vector $z \in \mathbb{R}^n$ can be written $z = x + y$, with $x = Pz \in V$ and $y = (x - Px) \in V^\perp$. We can summarize by saying that

$$V \oplus V^\perp = \mathbb{R}^n, \tag{92.66}$$

is an *orthogonal decomposition* of $\mathbb{R}^n$ into the two orthogonal subspaces V and $V^\perp$: $x \in V$ and $y \in V^\perp$ implies $(x, y) = 0$ and any $z \in \mathbb{R}^n$ can be written uniquely in the form $z = x+y$. The uniqueness of the decomposition $z = Pz + (z - Pz)$ follows from the uniqueness of Pz.

We note the following generalization of Pythagoras theorem: for any $x \in \mathbb{R}^n$, we have

$$|x|^2 = |Px|^2 + |x - Px|^2. \tag{92.67}$$

This follows by writing $x = Px + (x - Px)$ and using that $Px \perp (x - Px)$:

$$|x|^2 = |Px + (x - Px)|^2 = |Px|^2 + 2(Px, x - Px) + |x - Px|^2.$$

More generally, we have if $z = x + y$ with $x \perp y$ (that is $(x, y) = 0$), that

$$|z|^2 = |x|^2 + |y|^2.$$

92.38 Properties of Projections

Let P be the orthogonal projection onto a linear subspace V in $\mathbb{R}^n$. Then $P : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation that satisfies

$$P^\top = P \quad \text{and} \quad PP = P. \tag{92.68}$$

We have already seen that $PP = P$. To see that $P^\top = P$ we note that

$$(w, P^\top v) = (Pw, v) = (Pw, Pv) = (w, Pv) \quad \text{for all } v, w \in \mathbb{R}^n, \tag{92.69}$$

and thus $P^\top = P$. Conversely, let $P : \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation which satisfies (92.68). Then P is the orthogonal projection onto a subspace V of $\mathbb{R}^n$. To see this, set $V = R(P)$ and note that since $P^\top = P$ and $PP = P$, we have

$$\begin{aligned}(x - Px, Px) &= (x, Px) - (Px, Px) = (x, Px) - (x, P^\top Px)\\ &= (x, Px) - (x, Px) = 0.\end{aligned}$$

This shows that $x = Px + (x - Px)$ is an orthogonal decomposition, and thus P is the orthogonal projection onto $V = R(P)$.

92.39 Orthogonalization: The Gram-Schmidt Procedure

Let $\{a_1, ..., a_m\}$ be a basis for a subspace V of $\mathbb{R}^n$, i.e., $\{a_1, ..., a_m\}$ is linearly independent and V is the set of linear combinations of $\{a_1, ..., a_m\}$. We try to construct another basis $\{\hat{e}_1, ..., \hat{e}_m\}$ for V that is *orthonormal*, i.e. such that the basis vectors $\hat{e}_i$ are mutually orthogonal and have length equal to one or

$$(\hat{e}_i, \hat{e}_j) = 0 \quad \text{for } i \neq j, \quad \text{and } |\hat{e}_i| = 1 \tag{92.70}$$

We choose $\hat{e}_1 = \frac{1}{|a_1|}a_1$ and let V_1 be the subspace spanned by $\hat{e}_1$, or equivalently by a_1. Let P_1 be the projection onto V_1. Define

$$\hat{e}_2 = \frac{1}{|a_2 - P_1 a_2|}(a_2 - P_1 a_2).$$

Then $(\hat{e}_1, \hat{e}_2) = 0$ and $|\hat{e}_2| = 1$. Further, the subspace V_2 spanned by $\{a_1, a_2\}$ is also spanned by $\{\hat{e}_1, \hat{e}_2\}$. We now continue in the same way: Let P_2 be the projection onto V_2 and define

$$\hat{e}_3 = \frac{1}{|a_3 - P_2 a_3|}(a_3 - P_2 a_3)$$

Then the subspace V_3 spanned by $\{a_1, a_2, a_3\}$ is also spanned by the orthonormal set $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$.

Continuing, we obtain an orthonormal basis $\{\hat{e}_1, ..., \hat{e}_m\}$ for the subspace spanned by $\{a_1, ..., a_m\}$ with the property that for $i = 1, ..., m$, the subspace spanned by $\{a_1, ..., a_i\}$ is spanned by $\{\hat{e}_1, ..., \hat{e}_i\}$.

Note that since the basis $\{\hat{e}_1, ..., \hat{e}_m\}$ is orthogonal, the system of equations (92.62) corresponding to computing $P_{i-1}a_i$, is diagonal.

92.40 Orthogonal Matrices

Consider the matrix Q with columns $\hat{e}_1, .., \hat{e}_n$, where $\{\hat{e}_1, .., \hat{e}_n\}$ is an orthonormal basis for $\mathbb{R}^n$. Since the vectors $\hat{e}_j$ are pairwise orthogonal and of length one, $Q^\top Q = I$, where I is the $n \times n$ identity matrix. Conversely, if Q is a matrix such that $Q^\top Q = I$, where I is an identity matrix, then the columns of Q must be orthonormal.

An $n \times n$-matrix Q such that $Q^\top Q = I$, is called an *orthogonal matrix*. An orthogonal $n \times n$-matrix can thus be characterized as follows: Its columns form an *orthonormal basis* for $\mathbb{R}^n$, that is a basis consisting of pairwise orthogonal vectors of length, or norm, one.

We summarize:

Theorem 92.11 *An orthogonal matrix Q satisfies $Q^\top Q = QQ^\top = I$, and $Q^{-1} = Q^\top$.*

92.41 Invariance of the Scalar Product Under Orthogonal Transformations

Let Q by an $n \times n$ orthonormal matrix with the columns formed by the coefficients of basis vectors $\hat{e}_j$ of an orthonormal basis $\{\hat{e}_1, .., \hat{e}_n\}$. We then know that the coordinates x of a vector with respect to the standard basis, and the coordinates $\hat{x}$ with respect to the basis $\{\hat{e}_1, .., \hat{e}_n\}$, are connected by

$$x = Q\hat{x}.$$

We now prove that the scalar product is invariant under the orthonormal change of coordinates $x = Q\hat{x}$. We compute setting $y = Q\hat{y}$,

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

that is the scalar product is the same in the $\{e_1, .., e_n\}$ coordinates as in the $\{\hat{e}_1, .., \hat{e}_n\}$ coordinates. We summarize:

Theorem 92.12 *If Q is an orthogonal $n \times n$ matrix, then $(x, y) = (Qx, Qy)$ for all $x, y \in \mathbb{R}^n$.*

92.42 The QR-Decomposition

We can give the Gram-Schmidt orthogonalization procedure the following matrix interpretation: Let $\{a_1, ..., a_m\}$ be m linearly independent vectors in $\mathbb{R}^n$ and let A be the $n \times m$ matrix with the a_j occurring as columns. Let $\{\hat{e}_1, ..., \hat{e}_m\}$ be the corresponding orthonormal set generated by the Gram-Schmidt procedure, and let Q be the $n \times m$ matrix with the $\hat{e}_j$ as columns. Then

$$A = QR, \tag{92.71}$$

where R is a $m \times m$ upper triangular matrix, which expresses each a_j as a linear combination of $\{\hat{e}_1, ..., \hat{e}_j\}$.

The matrix Q satisfies $Q^\top Q = I$, where I is the $m \times m$ identity matrix, since the $\hat{e}_j$ are pairwise orthogonal and have length 1. We conclude that a $m \times n$ matrix A with linearly independent columns can be factored into $A = QR$, where Q satisfies $Q^\top Q = I$, and R is upper triangular. The columns of the matrix Q are orthonormal, as in the case of an orthonormal matrix, but if $m < n$, then they do not span all of $\mathbb{R}^n$.

92.43 The Fundamental Theorem of Linear Algebra

We return to our basic question of existence and uniqueness of solutions to the system $Ax = b$ with A a given $m \times n$ matrix and $b \in \mathbb{R}^m$ a given vector.

We now allow m to be different from n, remembering that we focussed on the case $m = n$ above. We shall now prove the Fundamental Theorem of Linear Algebra giving an answer of theoretical nature to our basic questions of existence and uniqueness.

We note the following chain of equivalent statements for a $m \times n$-matrix A, where "iff" is shorthand for "if and only if":

$$\begin{aligned} x \in N(A) \quad \text{iff} \quad Ax = 0 \quad \text{iff} \quad x \perp \text{ rows of } A \quad \text{iff} \\ x \perp \text{ columns of } A^\top \quad \text{iff} \\ x \perp R(A^\top) \quad \text{iff} \\ x \in (R(A^\top)^\perp. \end{aligned}$$

Thus $N(A) = (R(A^\top)^\perp$, and since $R(A^\top)^\perp \oplus R(A^\top) = \mathbb{R}^n$, we see that

$$N(A) \oplus R(A^\top) = \mathbb{R}^n. \tag{92.72}$$

As a consequence of this orthogonal splitting, we see that

$$\dim\, N(A) + \dim\, R(A^\top) = n, \tag{92.73}$$

where $\dim\, V$ is the dimension of the linear space V. We recall that the dimension $\dim\, V$ of a linear space V is the number of elements in a basis for V. Similarly, replacing A by $A^\top$ and using that $(A^\top)^\top = A$, we have

$$N(A^\top) \oplus R(A) = \mathbb{R}^m, \tag{92.74}$$

and thus in particular,

$$\dim\, N(A^\top) + \dim\, R(A) = m. \tag{92.75}$$

Next we note that, letting $g_1, ..., g_k$ be a basis in $N(A)^\perp$ so that $Ag_1, ..., Ag_k$ span $R(A)$ and thus $\dim\, R(A) \leq k$, we have

$$\dim\, N(A) + \dim\, R(A) \leq n, \quad \text{and also} \quad \dim\, N(A^\top) + \dim\, R(A^\top) \leq m. \tag{92.76}$$

Adding (92.73) and (92.75), we conclude that equality holds in (92.76). We summarize in:

Theorem 92.13 (The Fundamental Theorem of Linear Algebra) *Let A be a $m \times n$ matrix. Then*

$$N(A) \oplus R(A^\top) = \mathbb{R}^n \quad N(A^\top) \oplus R(A) = \mathbb{R}^m,$$

$$\textit{dim}\ N(A) + \textit{dim}\ R(A^\top) = n,\ \textit{dim}\ N(A^\top) + \textit{dim}\ R(A) = m,$$

$$\textit{dim}\ N(A) + \textit{dim}\ R(A) = n,\ \textit{dim}\ N(A^\top) + \textit{dim}\ R(A^\top) = m,$$

$$\textit{dim}\ R(A) = \textit{dim}\ R(A^\top).$$

In the special case $m = n$, we have that $R(A) = \mathbb{R}^m$ if and only if $N(A) = 0$ (which we proved above using Cramer's rule), stating that uniqueness implies existence.

We call dim $R(A)$ the *column rank* of the matrix A. The column rank of A is equal to the dimension of the space spanned by the columns of A. Similarly the *row rank* of A is equal to the dimension of the space spanned by the rows of A. The equality dim $R(A) =$ dim $R(A^\top)$ in the Fundamental Theorem expresses that the the column ranks of A and $A^\top$ are equal, that is that the column rank of A is equal to the *row rank* of A. We state this result as:

Theorem 92.14 *The number of linearly independent columns of A is equal to the number of linearly independent rows of A.*

EXAMPLE 92.10. Returning to Example 41.5, we note that the column echelon form of $A^\top$ is the transpose of the row echelon form of A, that is

$$R(A^\top) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 1 & 3 & 2 & 0 \\ 1 & 6 & 5 & 0 \end{pmatrix}.$$

We check that the columns vectors $(0, 1, -2, 1, 0)$ and $(0, 4, -5, 0, 1)$ spanning $N(A)$ are orthogonal to $R(A^\top)$, that is orthogonal to the columns of the echelon form of $A^\top$. Of course, this is just a restatement of the fact that these vectors are orthogonal to the rows of the row echelon form $\hat{A}$ of A (as is evident from the proof of the Fundamental Theorem). We see that $N(A) \oplus R(A^\top) = \mathbb{R}^5$ as predicted by the Fundamental Theorem.

92.44 Change of Basis: Coordinates and Matrices

Let $\{s_1, ..., s_n\}$ be a basis for $\mathbb{R}^n$ where the coordinates of the basis vectors in the standard basis $\{e_1, ..., e_n\}$ are given by $s_j = (s_{1j}, ..., s_{nj})$. Recalling (92.23), we have the following connection between the coordinates x_i of a vector x with respect to the standard basis and the coordinates $\hat{x}_j$ of x with respect to the basis $\{s_1, ..., s_n\}$:

$$x_i = \sum_{j=1}^{n} s_{ij}\hat{x}_j \quad \text{for } i = 1, ...n. \tag{92.77}$$

This follows from taking the scalar product of $\sum_{j=1}^n x_j e_j = \sum_{i=1}^n \hat{x}_j s_j$ with e_i and using that $s_{ij} = (e_i, s_j)$.

Introducing the matrix $S = (s_{ij})$, we thus have the following connection between the coordinates $x = (x_1, ..., x_n)$ with respect to $\{e_1, ..., e_n\}$, and the coordinates $\hat{x} = (\hat{x}_1, ...\hat{x}_n)$ with respect to $\{s_1, ..., s_n\}$:

$$x = S\hat{x}, \quad \text{that is } \hat{x} = S^{-1}x. \tag{92.78}$$

Consider now a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ with matrix $A = (a_{ij})$ with respect to the standard basis $\{e_1, ...e_n\}$, that is with $a_{ij} = f_i(e_j) = (e_i, f(e_j))$, where $f(x) = (f_1(x), ..., f_n(x))$ in the standard basis $\{e_1, ...e_n\}$, that is

$$y = f(x) = \sum_i f_i(x)e_i = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_j e_i = Ax.$$

Writing $y = S\hat{y}$ and $x = S\hat{x}$, we have

$$S\hat{y} = AS\hat{x} \quad \text{that is } \hat{y} = S^{-1}AS\hat{x}$$

This shows that the matrix of the linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$, with the matrix A with respect to the standard basis, takes the following form in the basis $\{s_1, ...s_n\}$:

$$S^{-1}AS, \tag{92.79}$$

where the the coefficients s_{ij} of the matrix $S = (s_{ij})$ are the coordinates of the basis vectors s_j with respect to the standard basis.

92.45 Least Squares Methods

Consider the $m \times n$ linear system of equations $Ax = b$, or

$$\sum_j^n a_j x_j = b,$$

where $A = (a_{ij})$ is an $m \times n$ matrix with columns $a_j = (a_{1j}, ..., a_{mj})$, $j = 1, ..., n$. We know that if $b \in R(A)$ then the system can be solved, and if $N(A) = 0$, then the solution is unique. Suppose now that $b \notin R(A)$. Then there is no $x \in \mathbb{R}^n$ such that $Ax = b$, and the system $Ax = b$ has no solution. We can replace the problem by the following *least squares problem*

$$\min_{x \in \mathbb{R}^n} |Ax - b|^2.$$

This problem amounts to seeking the projection Pb of b onto $R(A)$, that is the projection of b onto the space spanned by the columns a_j of A.

By the properties of projections given above, we know that $Pb \in R(A)$ exists and is uniquely determined by the relation

$$(Pb, y) = (b, y) \quad \text{for all } y \in R(A),$$

that is we seek $Pb = A\hat{x}$ for some $\hat{x} \in \mathbb{R}^n$ such that

$$(A\hat{x}, Ax) = (b, Ax) \quad \text{for all } x \in \mathbb{R}^n.$$

This relation can be written

$$(A^\top A\hat{x}, x) = (A^\top b, x) \quad \text{for all } x \in \mathbb{R}^n,$$

which is the same as the matrix equation

$$A^\top A\hat{x} = A^\top b,$$

which we refer to as the *normal equations*.

The matrix $A^\top A$ is an $n \times n$ symmetric matrix. Assume now that the columns a_j of A are linearly independent. Then $A^\top A$ is non-singular, because if $A^\top Ax = 0$, then

$$0 = (A^\top Ax, x) = (Ax, Ax) = |Ax|^2,$$

and thus $Ax = 0$ and therefore $x = 0$, since the columns of A are linearly independent. Thus the equation $A^\top A\hat{x} = A^\top b$ has a unique solution $\hat{x}$ for each right hand side $A^\top b$, given by the formula

$$\hat{x} = (A^\top A)^{-1}A^\top b.$$

In particular, we have the following formula for the projection Pb of b onto $R(A)$,

$$Pb = A(A^\top A)^{-1}A^\top b.$$

We can directly check that $P : \mathbb{R}^m \to \mathbb{R}^m$ defined this way is symmetric and satisfies $P^2 = P$.

If the columns of A are linearly dependent, then $\hat{x}$ is undetermined up to vectors $\hat{x}$ in $N(A)$. It is then natural to single out a unique $\hat{x}$ by requiring that $|\hat{x}|^2$ to be minimal. Using the orthogonal decomposition $\mathbb{R}^n = R(A^\top) \oplus N(A)$, this is equivalent to seeking $\hat{x}$ in $R(A^\top)$, since by Pythagoras theorem this minimizes $|\hat{x}|$. We thus seek $\hat{x}$ so that

- $A\hat{x}$ is equal to the projection Pb of b onto $R(A)$
- $\hat{x} \in R(A^\top)$.

This leads to the following equation for $\hat{x} = A^\top\hat{y}$:

$$(A\hat{x}, AA^\top y) = (b, AA^\top y) \quad \text{for all } y \in \mathbb{R}^m, \tag{92.80}$$

with $\hat{x}$ uniquely determined.

Chapter 92 Problems

92.1. Prove that a plane in $\mathbb{R}^3$ not passing through the origin is not a subspace of $\mathbb{R}^3$.

92.2. (a) What is a vector space? (b) What is a subspace of a vector space?

92.3. Verify (92.17) and (92.18).

92.4. Why must a set of more than n vectors in $\mathbb{R}^n$ be linearly dependent? Why must a set of n linearly independent vectors in $\mathbb{R}^n$ be a basis?

92.5. Verify that $R(A)$ and $N(A)$ are linear subspaces of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, and further that the orthogonal complement $V^\top$ of a subspace V of $\mathbb{R}^n$ is also a subspace of $\mathbb{R}^n$.

92.6. (a) Give an example showing that permutations need not commute. (b) Verify the associative law for permutations.

92.7. Compute the determinants of some $n \times n$ matrices with $n = 2, 3, 4, 5$.

92.8. Fill out the details in the proof of Cauchys inequality.

92.9. Write an algorithm for the Gram-Schmidt orthogonalization procedure, and implement it in Matlab, for example.

92.10. Fill in the details in (92.55)

92.11. Verify that for an orthogonal matrix $QQ^\top = I$. Hint: Multiply $Q^\top Q = I$ from the right with C and from the right with Q, where C is the matrix such that $QC = I$.

92.12. Prove for 2×2 matrices A and B that $\det AB = \det A \det B$.

92.13. How many operations are needed to solve an $n \times n$ system of linear equations using Cramer's formula?

92.14. Prove by reduction to column echelon form that a basis for $\mathbb{R}^n$ contains n elements.

92.15. Implement algorithms for reduction to column and row echelon forms.

92.16. Prove that the solution $\hat{x} \in R(A^\top)$ of (92.80) is uniquely determined.

92.17. Construct the row and column echelon forms of different (small) matrices, and check the validity of the Fundamental Theorem.

93
The Spectral Theorem

There seems to be three possibilities (of a Unified Theory of Physics):

1. There really is a complete unified theory, which we will someday discover if we are smart enough.
2. There is no ultimate theory of the Universe, just an infinite sequence of theories that describe the Universe more and more accurately.
3. There is no theory of the Universe; events cannot be predicted beyond a certain extent but occur in a random and arbitrary manner. (Stephen Hawking, in *A Brief History of Time*)

93.1 Eigenvalues and Eigenvectors

Let $A = (a_{ij})$ be a quadratic $n \times n$ matrix. We investigate the situation in which multiplication by A acts like scalar multiplication. To start with, we assume that the elements a_{ij} are real numbers. If $x = (x_1, ..., x_n) \in \mathbb{R}^n$ is a non-zero vector that satisfies

$$Ax = \lambda x, \tag{93.1}$$

where λ is a real number, then we say that $x \in \mathbb{R}^n$ is an *eigenvector* of A and that λ is a corresponding *eigenvalue* of A. An eigenvector x has the property that Ax is parallel to x (if $\lambda \neq 0$), or $Ax = 0$ (if $\lambda = 0$). This is a special property, as easy to verify with almost any example we might make up.

If x is an eigenvector with corresponding eigenvalue λ then $\bar{x} = \mu x$ for any non-zero real number μ is also an eigenvector corresponding to the eigenvalue λ because

$$\text{if } Ax = \lambda x, \text{ then } A\bar{x} = \mu Ax = \mu\lambda x = \lambda\mu x = \lambda\bar{x}.$$

Thus, we may change the length of an eigenvector without changing the corresponding eigenvalue. For example, we may normalize an eigenvector to have length equal to 1. In essence, the direction of an eigenvector is determined, but not its length.

We shall now study the problem of finding eigenvalues and corresponding eigenvectors of a given a quadratic matrix. We shall see that this is a basic problem of linear algebra arising in many different situations. We shall prove the *Spectral Theorem* stating that if A is a symmetric real $n \times n$ matrix, then there is an orthogonal basis for $\mathbb{R}^n$ consisting of eigenvectors. We shall also briefly discuss the case of non-symmetric matrices.

Rewriting (93.1) as $(A-\lambda I)x = 0$ with $x \in \mathbb{R}^n$ a non-zero eigenvector and I the identity matrix, we see that the matrix $A-\lambda I$ must be singular if λ is an eigenvalue, that is $\det(A - \lambda I) = 0$. Conversely, if $\det(A - \lambda I) = 0$ then $A - \lambda I$ is singular and thus the null-space $N(A - \lambda I)$ is different from the zero vector and thus there is a non-zero vector x such that $(A - \lambda I)x = 0$, that is there is an eigenvector x with corresponding eigenvalue λ. Using the expansion formula for the determinant, we see that $\det(A - \lambda I)$ is a polynomial in λ of degree n with coefficients depending on the coefficients a_{ij} of A. The polynomial equation

$$\det(A - \lambda I) = 0$$

is called the *characteristic equation*. We summarize:

Theorem 93.1 *The number λ is an eigenvalue of the $n \times n$ matrix A if and only if λ is a root of the characteristic equation* $\det(A - \lambda I) = 0$.

Example 93.1.

If $A = (a_{ij})$ is a 2×2 matrix, then the characteristic equation is

$$\det(A - \lambda I) = (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0,$$

which is a second order polynomial equation in λ. For example, if

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

then the characteristic equation is $\det(A - \lambda I) = \lambda^2 - 1 = 0$ with roots $\lambda_1 = 1$ and $\lambda_2 = -1$. The corresponding normalized eigenvectors are $s_1 = \frac{1}{\sqrt{2}}(1, 1)$ and $s_2 = \frac{1}{\sqrt{2}}(1, -1)$ since

$$(A - \lambda_1 I)\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

and similarly $(A-\lambda_2)s_2 = 0$. We observe that $(s_1, s_2) = s_1 \cdot s_2 = 0$, that is eigenvectors corresponding to different eigenvalues are orthogonal.

93.2 Basis of Eigenvectors

Suppose $\{s_1, ..., s_n\}$ is a basis for $\mathbb{R}^n$ consisting of eigenvectors of the $n \times n$ matrix $A = (a_{ij})$ with corresponding eigenvalues λ_1,...,λ_n so

$$As_i = \lambda_i s_i \quad \text{for } i = 1, ..., n. \tag{93.2}$$

Let S be the matrix with the columns equal to the eigenvectors s_j expressed in the standard basis. We can then write (93.2) in matrix form as follows,

$$AS = SD, \tag{93.3}$$

where D is the diagonal matrix with the eigenvalues λ_j on the diagonal. We thus have

$$A = SDS^{-1} \qquad \text{or} \qquad D = S^{-1}AS, \tag{93.4}$$

where D is a diagonal matrix. We say that S transforms A into a diagonal matrix D with the eigenvalues on the diagonal.

Conversely, if we can express a matrix A in the form $A = SDS^{-1}$ with S non-singular and D diagonal then $AS = SD$, which says that the columns of S are eigenvectors with corresponding eigenvalues on the diagonal of D.

Viewing the $n \times n$ matrix A as defining a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ by $f(x) = Ax$, we can express the action of $f(x)$ in a basis of eigenvectors $\{s_1, ..., s_n\}$ by the diagonal matrix D since $f(s_i) = \lambda_i s_i$. Thus, the linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ is expressed by the matrix A in the standard basis and by the diagonal matrix D matrix in a basis of eigenvectors. The coupling is given by

$$D = S^{-1}AS.$$

Of course, the action of a diagonal matrix is very easy to describe and to understand and this is the motivation for considering eigenvalues and eigenvectors.

We now formulate the following basic question in two equivalent forms:

- Given a $n \times n$ matrix A, is there a basis of eigenvectors of A?
- Given a $n \times n$ matrix A, is there a non-singular matrix S such that $S^{-1}AS$ is diagonal?

As we have seen, the columns of the matrix S are the eigenvectors of A and the diagonal elements are the eigenvalues.

We shall now give the following partial answer: if A is an $n \times n$ symmetric matrix, then there is an orthogonal basis for $\mathbb{R}^n$ consisting of eigenvectors.

This is the celebrated *Spectral Theorem for symmetric matrices*. Notice the assumption that A is symmetric and that in this case the basis of eigenvectors may be chosen to be orthogonal.

EXAMPLE 93.2. Recalling Example 93.1, we see that $s_1 = \frac{1}{\sqrt{2}}(1,1)$ and $s_2 = \frac{1}{\sqrt{2}}(1,-1)$ form an orthogonal basis. By the orthogonality of S, $S^{-1} = S^\top$, and

$$S^{-1}AS = \frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$$
$$= \frac{1}{2}\begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix} = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}.$$

93.3 An Easy Spectral Theorem for Symmetric Matrices

The following version of the Spectral Theorem for symmetric matrices is easy to prove:

Theorem 93.2 *Let A be a symmetric $n \times n$ matrix. Suppose A has n distinct eigenvalues λ_1,..., λ_n and corresponding normalized eigenvectors s_1,....,s_n with $\|s_j\| = 1$, $j = 1, ..., n$. Then, $\{s_1, ..., s_n\}$ is an orthonormal basis of eigenvectors. Letting $Q = (q_{ij})$ be the orthogonal matrix with the columns $(q_{1j}, ..., q_{nj})$ being the coordinates of the eigenvectors s_j with respect to the standard basis, then $D = Q^{-1}AQ$ is a diagonal matrix with the eigenvalues λ_j on the diagonal and $A = QDQ^{-1}$, where $Q^{-1} = Q^\top$.*

To prove this result, it suffices to prove that eigenvectors corresponding to different eigenvalues are orthogonal. This follows from the assumption that there are n distinct eigenvalues λ_1,..., λ_n with corresponding normalized eigenvectors s_1,....,s_n. If we prove that these eigenvectors are pairwise orthogonal, then they form a basis for $\mathbb{R}^n$ and the proof is complete. Thus, assume that s_i and s_j are eigenvectors corresponding to different eigenvalues λ_i and λ_j. Since A is symmetric and $(Ax, y) = (x, Ay)$ for all $x, y \in \mathbb{R}^n$, we have

$$\lambda_i(s_i, s_j) = (\lambda_i s_i, s_j) = (As_i, s_j) = (s_i, As_j)$$
$$= (s_i, \lambda_j s_j) = \lambda_j(s_i, s_j),$$

which implies that $(s_i, s_j) = 0$ since $\lambda_i \neq \lambda_j$. We state this observation as a theorem because of its basic importance.

Theorem 93.3 *If A is a symmetric $n \times n$ matrix, and s_i and s_j are eigenvectors of A corresponding to the eigenvalues λ_i and λ_j with $\lambda_i \neq \lambda_j$, then*

$(s_i, s_j) = 0$. *In other words, eigenvectors corresponding to different eigenvalues are orthogonal.*

Note that above we prove the Spectral Theorem for a symmetric $n \times n$ matrix A in the case the characteristic equation $\det(A - \lambda I) = 0$ has n different roots. It thus remains to consider the case of multiple roots where there are less than n different roots. We will consider this below. The reader in hurry may skip that proof.

93.4 Applying the Spectral Theorem to an IVP

We show a typical application of the Spectral Theorem. Consider the initial value problem: find $u : [0, 1] \to \mathbb{R}^n$ such that

$$\dot{u} = Au, \quad \text{for } 0 < t \leq 1, \quad u(0) = u_0,$$

where $A = (a_{ij})$ is a symmetric $n \times n$ matrix with real coefficients a_{ij} independent of t. Systems of this form arise in many applications and the behavior of such a system may be very complicated.

Suppose now that $\{g_1, ..., g_n\}$ is an orthonormal basis of eigenvectors of A and let Q be the matrix with columns comprised of the coordinates of the eigenvectors g_j with respect to the standard basis. Then $A = QDQ^{-1}$, where D is the diagonal matrix with the eigenvalues λ_j on the diagonal. We introduce the new variable $v = Q^{-1}u$, that is we set $u = Qv$, where $v : [0, 1] \to \mathbb{R}^n$. Then, the equation $\dot{u} = Au$ takes the form $Q\dot{v} = AQv$, that is $\dot{v} = Q^{-1}AQv = Dv$, where we use the fact that Q is independent of time. Summing up, we get the following diagonal system in the new variable v,

$$\dot{v} = Dv \quad \text{for } 0 < t \leq 1, \quad v(0) = v_0 = Q^{-1}u_0.$$

The solution of this decoupled system is given by

$$v(t) = \begin{pmatrix} \exp(\lambda_1 t) & 0 & 0 & .... & 0 \\ 0 & \exp(\lambda_2 t) & 0 & .... & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & .... & \exp(\lambda_n) \end{pmatrix} v_0 = \exp(Dt)v_0,$$

where $\exp(Dt)$ is a diagonal matrix with the elements $\exp(\lambda_j t)$ on the diagonal. The dynamics of this system is easy to grasp: each component $v_j(t)$ of $v(t)$ evolves according to $v_j(t) = \exp(\lambda_j t)v_{0j}$.

Transforming back, we get the following solution formula in the original variable $u(t)$,

$$u(t) = Q\exp(Dt)Q^{-1}u_0. \tag{93.5}$$

With A as in Example 93.1, we get the solution formula

$$u(t) = \frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}e^t & 0\\ 0 & e^{-t}\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}u_{01}\\ u_{02}\end{pmatrix}$$
$$= \frac{1}{2}\begin{pmatrix}(e^t + e^{-t})u_{01} + (e^t - e^{-t})u_{02}\\ (e^t - e^{-t})u_{01} + (e^t + e^{-t})u_{02}\end{pmatrix}.$$

93.5 The General Spectral Theorem for Symmetric Matrices

Above, we saw that eigenvalues of a matrix A are roots of the characteristic equation $\det(A - \lambda I) = 0$. In principle, we can find the eigenvalues and eigenvectors of given matrix by first solving the characteristic equation to find all the eigenvalues, and then for each eigenvalue λ find corresponding eigenvectors by solving the linear system of equations $(A - \lambda I)x = 0$.

We shall now present an alternative way of constructing/finding the eigenvectors and eigenvalues of a symmetric matrix A that also proves the Spectral Theorem for a symmetric $n \times n$ matrix A in the general case with possibly multiple roots. In the proof, we construct an orthonormal basis of eigenvectors $\{s_1, ..., s_n\}$ of A by constructing the eigenvectors one by one starting with s_1.

Constructing the First Eigenvector s_1

To construct the first eigenvector s_1, we consider the minimization problem: find $\bar{x} \in \mathbb{R}^n$ such that

$$F(\bar{x}) = \min_{x\in\mathbb{R}^n} F(x), \tag{93.6}$$

where

$$F(x) = \frac{(Ax, x)}{(x, x)} = \frac{(f(x), x)}{(x, x)} \tag{93.7}$$

is the so-called *Rayleigh quotient*. We note that the function $F(x)$ is *homogenous of degree zero*, that is for any $\lambda \in \mathbb{R}$, $\lambda \neq 0$, we have

$$F(x) = F(\lambda x),$$

because we can simply divide out the factor λ. In particular, for any $x \neq 0$,

$$F(x) = F(\frac{x}{\|x\|}), \tag{93.8}$$

and thus we may restrict the x in (93.6) to have length one, that is we may consider the equivalent minimization problem: find $\bar{x}$ with $\|\bar{x}\| = 1$ such that

$$F(\bar{x}) = \min_{x\in\mathbb{R}^n \, \|x\|=1} F(x) \tag{93.9}$$

Since $F(x)$ is Lipschitz continuous on the closed and bounded subset $\{x \in \mathbb{R}^n : \|x\| = 1\}$ of $\mathbb{R}^n$, we know by the Chapter Minimization, that the problem (93.9) has a solution $\bar{x}$, and thus also the problem (93.6) has a solution $\bar{x}$. We set $s_1 = \bar{x}$, and check that g_1 is indeed an eigenvector of A, that is an eigenvector of $f : \mathbb{R}^n \to \mathbb{R}^n$.

Since $\bar{x}$ solves the minimization problem (93.6), we have $\nabla F(\bar{x}) = 0$, where ∇F is the gradient of F. Computing $\nabla F(x)$ using the symmetry of $F(x)$ or the matrix A, we find that

$$\nabla F(x) = \frac{(x,x)2Ax - (Ax,x)2x}{(x,x)^2}, \tag{93.10}$$

so that with $x = \bar{x}$ satisfying $(\bar{x},\bar{x}) = 1$,

$$\nabla F(\bar{x}) = 2(A\bar{x} - (A\bar{x},\bar{x})\bar{x}) = 0,$$

that is

$$A\bar{x} = \lambda_1\bar{x}, \tag{93.11}$$

where

$$\lambda_1 = (A\bar{x},\bar{x}) = \frac{(A\bar{x},\bar{x})}{(\bar{x},\bar{x})} = \min_{x\in\mathbb{R}^n} F(x). \tag{93.12}$$

Setting $s_1 = \bar{x}$, we thus have

$$As_1 = \lambda_1 s_1, \quad \lambda_1 = (As_1, s_1), \quad \|s_1\| = 1.$$

We have now constructed the first normalized eigenvector s_1 with corresponding eigenvalue λ_1. We now let V_1 be the orthogonal complement of the space spanned by s_1, consisting of all the vectors $x \in \mathbb{R}^n$ such that $(x, s_1) = 0$. The dimension of V_1 is $n - 1$.

Invariance of A

Note that V_1 is *invariant* with respect to A in the sense that if $x \in V_1$ then $Ax \in V_1$. This follows because if $(x, g_1) = 0$ then $(Ax, s_1) = (x, As_1) = (x, \lambda_1 s_1) = \lambda_1(x, s_1) = 0$. This means that we can now restrict attention to the action of A on V_1, having handled the action of A on the space spanned by the first eigenvector s_1.

Constructing the Second Eigenvector s_2

Consider the minimization problem to find $\bar{x} \in V_1$ such that

$$F(\bar{x}) = \min_{x\in V_1} F(x). \tag{93.13}$$

By the same argument, this problem has a solution which we denote s_2 and which satisfies $As_2 = \lambda_2 s_2$, where $\lambda_2 = \frac{(As_2,s_2)}{(s_2,s_2)}$, and $\|s_2\| = 1$. Because in

(93.13) we minimize over a smaller set than in (93.6), $\lambda_2 \geq \lambda_1$. Note that it may happen that $\lambda_2 = \lambda_1$, although V_1 is a subset of $\mathbb{R}^n$. In that case, we say that $\lambda_1 = \lambda_2$ is a *multiple eigenvalue*.

Continuing the Process

Let V_2 be the orthogonal subspace to the space spanned by s_1 and s_2 Again A is invariant on V_2 and the space spanned by $\{s_1, s_2\}$. Continuing this way, we obtain a orthonormal basis $\{s_1, ..., s_n\}$ of eigenvectors of A with corresponding real eigenvalues λ_i.

We have now proved the famous

Theorem 93.4 (Spectral Theorem): *If $f : \mathbb{R} \to \mathbb{R}$ is a linear symmetric transformation with corresponding symmetric $n \times n$ matrix A in the standard basis, then there is an orthogonal basis $(g_1, ..., g_n)$ of $\mathbb{R}^n$ consisting of eigenvectors g_i of f with corresponding real eigenvalues λ_j satisfying $f(g_j) = Ag_j = \lambda_j g_j$, for $j = 1, ..., n$. We have $D = Q^{-1}AQ$ and $A = QDQ^{-1}$, where Q is the orthogonal matrix with the coefficients of the eigenvectors g_j in the standard basis forming the columns, and D is the diagonal matrix with the eigenvalues λ_j on the diagonal.*

93.6 The Norm of a Symmetric Matrix

We recall that we have defined the *Euclidean norm* $\|A\|$ of a $n \times n$ matrix A by

$$\|A\| = \max_{x \in \mathbb{R}^n} \frac{\|Ax\|}{\|x\|}, \tag{93.14}$$

where we maximize over $x \neq 0$. By the definition, we have

$$\|Ax\| \leq \|A\| \, \|x\|, \tag{93.15}$$

and we may thus view $\|A\|$ to be the smallest constant C such that $\|Ax\| \leq C\|x\|$ for all $x \in \mathbb{R}^n$.

We shall now prove that if A is symmetric, then we can directly relate $\|A\|$ to the eigenvalues $\lambda_1, ..., \lambda_n$ of A:

$$\|A\| = \max_{i=1,...,n} |\lambda_i|. \tag{93.16}$$

We do this as follows. Using the Spectral theorem, we can write A as $A = Q^\top \Lambda Q$ with Q orthogonal and Λ a diagonal matrix with the eigenvalues λ_i on the diagonal. We recall that (cf. (92.46))

$$\|\Lambda\| = \max_{i=1,...,n} |\lambda_i| = |\lambda_j| \tag{93.17}$$

and thus for all $x \in \mathbb{R}^n$,

$$\|Ax\| = \|Q^\top \Lambda Q x\| = \|\Lambda Q x\| \leq \|\Lambda\|\|Qx\| = \|\Lambda\|\|x\| = \max_{i=1,\ldots,n} |\lambda_i| \|x\|,$$

which proves that $\|A\| \leq \max_{i=1,\ldots,n} |\lambda_i|$. Choosing x to be equal to the eigenvector corresponding to the eigenvalue λ_j of maximal modulus proves that indeed $\|A\| = \max_{i=1,\ldots,n} |\lambda_i| = |\lambda_j|$. We have proved the following result, which is a corner stone of numerical linear algebra.

Theorem 93.5 *If A is a symmetric $n \times n$ matrix, then $\|A\| = \max |\lambda_i|$, where $\lambda_1, ..., \lambda_n$ are the eigenvalues of A.*

93.7 Extension to Non-Symmetric Real Matrices

Up until now, we have mainly focussed on the case of *real scalars*, that is we assume that the components of vectors are real numbers. We know that we can also let the components of vectors be *complex numbers*, and we may then allow eigenvalues to be complex numbers. The fundamental theorem of algebra states that a polynomial equation of degree n with complex coefficients, has n complex roots, and thus the characteristic equation $\det(A - \lambda I) = 0$ has n complex roots λ_1,....,λ_n, and thus a $n \times n$ matrix A has n complex eigenvalues λ_1,....,λ_n, if roots are counted with multiplicity. We have in this chapter focussed on *symmetric* matrices A with real coefficients and we have proved that a symmetric matrix with real coefficients has n real eigenvalues, counted with multiplicity. For symmetric matrices we can thus limit ourselves to real roots of the characteristic equation.

Chapter 93 Problems

93.1. Verify (93.10).

93.2. Compute the eigenvalues and eigenvectors of an arbitrary symmetric 2×2 matrix A. Solve the corresponding initial-value problem $\dot{u}(t) = Au(t)$ for $t > 0$, $u(0) = u^0$.

94
Solving Linear Algebraic Systems

All thought is a kind of computation. (Hobbes)

94.1 Introduction

We are interested in solving a system of linear equations

$$Ax = b,$$

where A is a given $n \times n$ matrix and $b \in \mathbb{R}^n$ is a given n-vector and we seek the solution vector $x \in \mathbb{R}^n$. We recall that if A is non-singular with non-zero determinant, then the solution $x \in \mathbb{R}^n$ is theoretically given by Cramer's formula. However if n is large, the computational work in using Cramer's formula is prohibitively large, so we need to find a more efficient means of computing the solution.

We shall consider two types of methods for solving the system $Ax = b$: (i) *direct methods* based on *Gaussian elimination* that theoretically produce a solution after a finite number of arithmetic operations, and (ii) *iterative methods* that produce a generally infinite sequence of increasingly accurate approximations.

94.2 Direct Methods

We begin by noting that some linear systems are easier to solve than others. For example if $A = (a_{ij})$ is *diagonal*, which means that $a_{ij} = 0$ if $i \neq j$,

then the system is solved in n operations: $x_i = b_i/a_{ii}$, $i = 1, ..., n$. Further, if the matrix is *upper triangular*, which means that $a_{ij} = 0$ if $i > j$, or *lower triangular*, which means that $a_{ij} = 0$ if $i < j$, then the system can be solved by *backward substitution* or *forward substitution* respectively; see Fig. 94.1 for an illustration of these different types. For example if A is

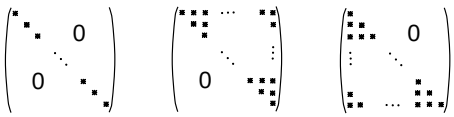


FIGURE 94.1. The pattern of entries in diagonal, upper, and lower triangular matrices. A "$*$" denotes a possibly nonzero entry.

upper triangular, the "pseudo-code" shown in Fig. 94.2 solves the system $Ax = b$ for the vector $x = (x_i)$ given the vector $b = (b_i)$ (assuming that $a_{kk} \neq 0$): In all three cases, the systems have a unique solution as long as

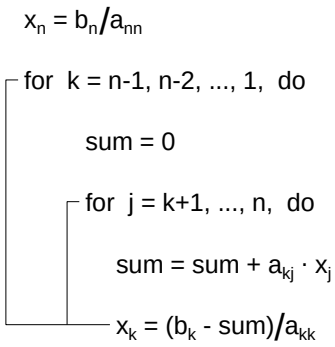


FIGURE 94.2. An algorithm for solving an upper triangular system by back substitution.

the diagonal entries of A are nonzero.

Direct methods are based on Gaussian elimination, which in turn is based on the observation that the solution of a linear system is not changed under the following *elementary row operations*:

- interchanging two equations
- adding a multiple of one equation to another
- multiplying an equation by a nonzero constant.

The idea behind Gaussian elimination is to transform using these operations a given system into an upper triangular system, which is solved by

back substitution. For example, to solve the system

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

we first subtract 2 times the first equation from the third to get the equivalent system,

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_2 + 2x_3 &= 1 \\ -x_2 + x_3 &= -1. \end{aligned}$$

We define the *multiplier* to be the factor 2. Next, we subtract -1 times the second row from the third to get

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_2 + 2x_3 &= 1 \\ 3x_3 &= 0. \end{aligned}$$

In this case, the multiplier is -1. The system is now upper triangular and using back substitution, we obtain $x_3 = 0$, $x_2 = 1$, and $x_1 = 0$. Gaussian elimination can be coded in a straightforward way using matrix notation.

Matrix Factorization

There is another way to view Gaussian elimination that is useful for the purposes of programming and handling special cases. Namely, Gaussian elimination is equivalent to computing a *factorization* of the coefficient matrix, $A = LU$, where L is a lower triangular and U an upper triangular $n \times n$ matrix. Given such a factorization of A, solving the system $Ax = b$ is straightforward. We first set $y = Ux$, then solve $Ly = b$ by forward substitution and finally solve $Ux = y$ by backward substitution.

To see that Gaussian elimination gives an LU factorization of A, consider the example above. We performed row operations that brought the system into upper triangular form. If we view these operations as row operations on the matrix A, we get the sequence

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 1 & 3 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & -1 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix},$$

which is an upper triangular matrix. This is the "U" in the LU decomposition.

The matrix L is determined by the observation that the row operations can be performed by multiplying A on the left by a sequence of special

matrices called *Gauss transformations*. These are lower triangular matrices that have at most one nonzero entry in the off-diagonal positions and 1s down the diagonal. We show a Gauss transformation in Fig. 94.3. Multiplying A on the left by the matrix in Fig. 94.3 has the effect of adding

$$\begin{pmatrix}
1 & 0 & & & & \cdots & & & 0 \\
0 & 1 & 0 & & & & & & 0 \\
 & \ddots & 1 & \ddots & & 0 & & & \\
\vdots & & & & & & & & \vdots \\
 & 0 & 0 & 0 & \ddots & 0 & & & \\
 & 0 & \alpha_{ij} & 0 & \ddots & 1 & \ddots & & \\
 & 0 & 0 & 0 & & & \ddots & & \\
 & & & & & & 0 & 1 & 0 \\
0 & & \cdots & & & & & 0 & 1
\end{pmatrix}$$

FIGURE 94.3. A Gauss transformation.

α_{ij} times row j of A to row i of A. Note that the inverse of this matrix is obtained changing α_{ij} to $-\alpha_{ij}$; we will use this below.

To perform the first row operation on A above, we multiply A on the left by

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix},$$

to get

$$L_1 A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & -1 & -1 \end{pmatrix}.$$

The effect of pre-multiplication by L_1 is to add $-2\times$ row 1 of A to row 3. Note that L_1 is lower triangular and has ones on the diagonal.

Next we multiply $L_1 A$ on the left by

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix},$$

and get

$$L_2 L_1 A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix} = U.$$

L_2 is also lower triangular with ones on the diagonal. It follows that $A = L_1^{-1}L_2^{-1}U$ or $A = LU$, where

$$L = L_1^{-1}L_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix}.$$

It is easy to see that L is also lower triangular with 1's on the diagonal with the multipliers (with sign change) occurring at the corresponding positions. We thus get the factorization

$$A = LU = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix}.$$

Note that the entries in L below the diagonal are exactly the multipliers used to perform Gaussian elimination on A.

A general linear system can be solved in exactly the same fashion by Gaussian elimination using a sequence of Gauss transformations to obtain a factorization $A = LU$.

An LU factorization can be performed *in situ* using the storage space allotted to the matrix A. The fragment of code shown in Fig. 94.4 computes the LU factorization of A, storing U in the upper triangular part of A and storing the entries in L below the diagonal in the part of A below the diagonal. We illustrate the storage of L and U in Fig. 94.5.

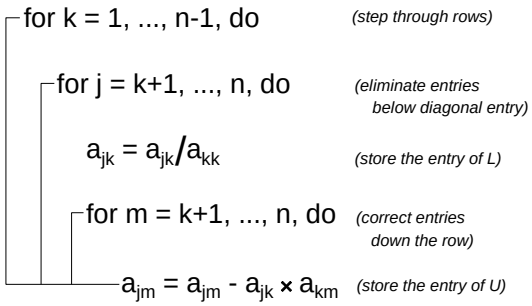


FIGURE 94.4. An algorithm to compute the LU factorization of A that stores the entries of L and U in the storage space of A.

Measuring the Cost

The cost of solving a linear system using a direct method is measured in terms of computer time. In practice, the amount of computer time is

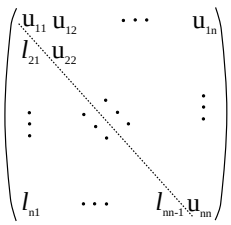


FIGURE 94.5. The matrix output from the algorithm in Fig. 94.4. L and U are stored in the space allotted to A.

proportional to the number of arithmetic and storage operations the computer uses to compute the solution. It is traditional (on a sequential computer) to simplify the cost calculation by equating storing a value, addition, and subtraction and equating multiplication and division when counting operations. Moreover, since multiplication (i.e. multiplications and divisions) generally cost much more than addition on older computers, it is also common to simply count the number of multiplications (=multiplications+divisions).

By this measure, the cost of computing the LU decomposition of an $n \times n$ matrix is $n^3 - n/3 = O(n^3/3)$. We introduce some new notation here, the big "O". The actual count is $n^3/3 - n/3$, however when n is large, the lower order term $-n/3$ becomes less significant. In fact,

$$\lim_{n\to\infty} \frac{n^3/3 - n/3}{n^3/3} = 1, \tag{94.1}$$

and this is the definition of the big "O". (Sometimes the big "O" notation means that the limit of the ratio of the two relevant quantities is any constant). With this notation, the operations count of the LU decomposition is just $O(n^3)$.

The cost of the forward and backward substitutions is much smaller:

Pivoting

During Gaussian elimination, it sometimes happens that the coefficient of a variable in the "diagonal position" becomes zero as a result of previous eliminations. When this happens of course, it is not possible to use that equation to eliminate the corresponding entries in the same column lying below the diagonal position. If the matrix is invertible, it is possible to find a non-zero coefficient in the same column and below the diagonal position, and by switching the two rows, the Gaussian elimination can proceed. This is called *zero pivoting*, or just *pivoting*.

Adding pivoting to the LU decomposition algorithm is straightforward. Before beginning the elimination using the current diagonal entry, we check

to see if that entry is non-zero. If it is zero, we search the entries below in the same column for the first non-zero value, then interchange the row corresponding to that non-zero entry with the row corresponding to the current diagonal entry which is zero. Because the row interchanges involve rows in the "un-factored" part of A, the form of L and U are not affected. We illustrate this in Fig. 94.6.

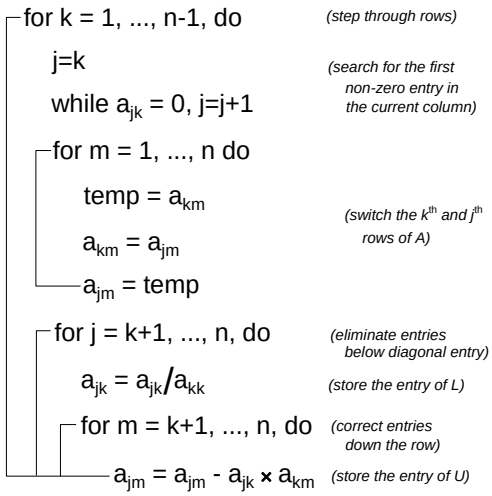


FIGURE 94.6. An algorithm to compute the LU factorization of A that used pivoting to avoid zero-valued diagonal entries.

To obtain the correct solution of the linear system $Ax = b$, we have to mirror all pivots performed on A in the data b. This is easy to do with the following trick. We define the vector of integers $p = (1\ 2\ \ldots\ n)^\top$. This vector is passed to the LU factorization routine and whenever two rows of A are interchanged, we interchange the corresponding entries in p. After getting the altered p vector back, we pass it to the forward/backward routine. Here, we address the vector b indirectly using the vector p, i.e., we use the vector with entries $(b_{p_i})_{i=1}^n$, which has the effect of interchanging the rows in b in the correct fashion.

There are additional reasons to pivot in practice. As we have noted, the computation of the LU decomposition can be sensitive to errors originating from the finite precision of the computer if the matrix A is close to being non-invertible. We discuss this further below. We mention here however that a special kind of pivoting, called *partial pivoting* can be used to reduce this sensitivity. The strategy behind partial pivoting is to search the entries in the same column and below the current diagonal entry for the largest in absolute value. The row corresponding to the largest entry in magnitude is interchanged with the row corresponding to the current entry at the diago-

nal. The use of partial pivoting generally gives more accurate results than factorization without partial pivoting. One reason is that partial pivoting insures that the multipliers in the elimination process are kept as small as possible and consequently the errors in each entry are magnified by as little as possible during the course of the Gaussian elimination. We illustrate this with an example. Suppose that we solve

$$\begin{aligned} .000100x_1 + 1.00x_2 &= 1.00 \\ 1.00x_1 + 1.00x_2 &= 2.00 \end{aligned}$$

on a computer that holds three digits. Without pivoting, we get

$$\begin{aligned} .000100x_1 + 1.00x_2 &= 1.00 \\ -10000x_2 &= -10000 \end{aligned}$$

which implies that $x_2 = 1$ and $x_1 = 0$. Note the large multiplier that is required for the elimination. Since the true answer is $x_1 = 1.0001$ and $x_2 = .9999$, the computed result has an error of 100% in x_1. If we switch the two rows before eliminating, which corresponds exactly to the partial pivoting strategy, we get

$$\begin{aligned} 1.00x_1 + 1.00x_2 &= 2.00 \\ 1.00x_2 &= 1.00 \end{aligned}$$

which gives $x_1 = x_2 = 1.00$ as a result.

94.3 Direct Methods for Special Systems

It is often the case that the matrices arising from the Galerkin finite element method applied to a differential equation have special properties that can be useful during the solution of the associated algebraic equations. For example, the stiffness matrix for the Galerkin finite element approximation of the two-point boundary value problem with no convection is symmetric, positive-definite, and tridiagonal. In this section, we examine a couple of different classes of problems that occur frequently.

Symmetric, Positive-Definite Systems

As we mentioned, symmetric, positive-definite matrices are often encountered when discretizing differential equations (especially if the spatial part of the differential equation is of the type called elliptic). If A is symmetric and positive-definite, then it can be factored as $A = BB^\top$ where B is a lower triangular matrix with positive diagonal entries. This factorization can be computed from the LU decomposition of A, but there is a *compact*

method of factoring A that requires only $O(n^3/6)$ multiplications called *Cholesky's method.*:

$$
\begin{aligned}
&b_{11} = \sqrt{a_{11}} \\
&b_{i1} = \frac{a_{i1}}{b_{11}}, \quad 2 \le i \le n, \\
&\begin{cases} b_{jj} = \left(a_{jj} - \sum_{k=1}^{j-1} b_{jk}^2\right)^{1/2}, \\ b_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} b_{ik} b_{jk}\right) / b_{jj}, \end{cases} \qquad 2 \le j \le n,\ j+1 \le i \le n
\end{aligned}
$$

This is called a compact method because it is derived by assuming that the factorization exists and then computing the coefficients of B directly from the equations obtained by matching coefficients in $BB^\top = A$. For example, if we compute the coefficient in the first row and column of $BB^\top$ we get b_{11}^2, which therefore must equal a_{11}. It is possible to do this because A is positive-definite and symmetric, which implies among other things that the diagonal entries of A remain positive throughout the factorization process and pivoting is not required when computing an LU decomposition.

Alternatively, the square roots in this formula can be avoided by computing a factorization $A = CDC^\top$ where C is a lower triangular matrix with ones on the diagonal and D is a diagonal matrix with positive diagonal coefficients.

Banded Systems

Banded systems are matrices with non-zero coefficients only in some number of diagonals centered around the main diagonal. In other words, $a_{ij} = 0$ for $j \le i - d_l$ and $j \ge i + d_u$, $1 \le i, j \le n$, where d_l is the *lower bandwidth*, d_u is the *upper bandwidth*, and $d = d_u + d_l - 1$ is called the *bandwidth*. We illustrate this in Fig. 94.7. The stiffness matrix computed for the two-point boundary value problem with no convection is an example of a tridiagonal matrix, which is a matrix with lower bandwidth 2, upper bandwidth 2, and bandwidth 3.

When performing the Gaussian elimination used to compute the LU decomposition, we see that the entries of A that are already zero do not have to be reduced further. If there are only relatively few diagonals with non-zero entries, then the potential saving is great. Moreover, there is no need to store the zero-valued entries of A. It is straightforward to adapt the LU factorization and forward/backward substitution routines to a banded pattern, once a storage scheme has been devised. For example, we can store

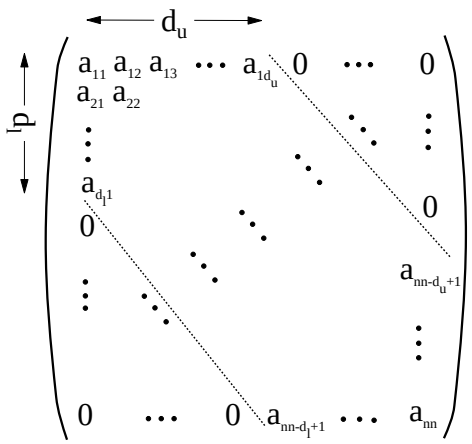


FIGURE 94.7. The notation for a banded matrix.

a tridiagonal matrix as a $3 \times n$ matrix:

$$\begin{pmatrix} a_{21} & a_{31} & 0 & & \cdots & & 0 \\ a_{12} & a_{22} & a_{32} & 0 & \cdots & & 0 \\ 0 & a_{13} & a_{23} & a_{33} & 0 & \cdots & \vdots \\ & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & 0 & a_{1n-1} & a_{2n-1} & a_{3n-1} \\ 0 & & \cdots & & 0 & a_{1n} & a_{2n} \end{pmatrix}.$$

The routine displayed in Fig. 94.8 computes the LU factorization, while the routine in Fig. 94.9 performs the forward/backward substitution.

```
for k = 2, ..., n, do
    a_1k = a_1k / a_2k-1
    a_2k = a_2k - a_1k × a_3k-1
```

FIGURE 94.8. A routine for computing the LU factorization of a tridiagonal system.

The cost of this routine grows linearly with the dimension, rather than at a cubic rate as in the full case. Moreover, we use only the equivalent of six vectors of dimension n for storage. A more efficient version, derived as a compact method, uses even less.

This algorithm assumes that no pivoting is required to factor A. Pivoting during the factorization of a banded matrix raises the difficulty that the

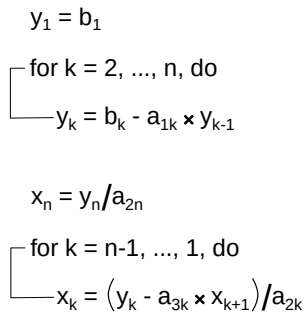


FIGURE 94.9. Using forward and backward substitution to solve a tridiagonal system given the LU factorization.

bandwidth becomes larger. This is easy to see in a tridiagonal matrix, in which case we have to store an extra vector to hold the extra elements above the diagonal that result if two adjacent rows are switched.

As for a tridiagonal matrix, it is straightforward to program special LU factorization and forward/backward substitution routines for a matrix with bandwidth d. The operations count is $O(nd^2/2)$ and the storage requirement is a matrix of dimension $d \times n$ if no pivoting is required. If d is much less than n, the savings in a special approach are considerable.

While it is true that if A is banded, then L and U are also banded, it is also true that in general L and U have non-zero entries in positions where A is zero. This is called *fill-in*. In particular, the stiffness matrix for a boundary value problem in several variables is banded and moreover most of the sub-diagonals in the band have zero coefficients. However, L and U do not have this property and we may as well treat A as if all the diagonals in the band have non-zero entries.

Banded matrices are one example of the class of *sparse* matrices. Recall that a sparse matrix is a matrix with mostly zero entries. As for banded matrices, it is possible to take advantage of sparsity to reduce the cost of factoring A in terms of time and storage. However, it is more difficult to do this than for banded matrices if the sparsity pattern puts non-zero entries at any location in the matrix. One approach to this problem is based on rearranging the equations and variables, or equivalently rearranging the rows and columns to form a banded system.

94.4 Iterative Methods

Instead of solving $Ax = b$ directly, we now consider iterative solution methods based on computing a sequence of approximations $x^{(k)}$, $k = 1, 2, ...$, such that

$$\lim_{k\to\infty} x^{(k)} = x \quad \text{or} \quad \lim_{k\to\infty} \|x^{(k)} - x\| = 0,$$

for some norm $\|\cdot\|$.

Note that the finite precision of a computer has a different effect on an iterative method than it has on a direct method. A theoretically convergent sequence can not reach its limit in general on a computer using a finite number of digits. In fact, at the point at which the change from one iterate to the next occurs outside the range of digits held by the computer, the sequence simply stops changing. Practically speaking, there is no point computing iterations past this point, even if the limit has not been reached. On the other hand, it is often sufficient to have less accuracy than the limit of machine precision, and thus it is important to be able to estimate the accuracy of the current iterate.

Minimization Algorithms

We first construct iterative methods for a linear system $Ax = b$ where A is symmetric and positive-definite. In this case, the solution x can be characterized equivalently as the solution of the quadratic minimization problem: find $x \in \mathbb{R}^n$ such that

$$F(x) \leq F(y) \quad \text{for all } y \in R^n, \tag{94.2}$$

where

$$F(y) = \frac{1}{2}(Ay, y) - (b, y),$$

with $(\cdot, \cdot)$ denoting the usual Euclidean scalar product.

We construct an iterative method for the solution of the minimization problem (94.2) based on the following simple idea: given an approximation $x^{(k)}$, compute a new approximation $x^{(k+1)}$ such that $F(x^{(k+1)}) < F(x^{(k)})$. On one hand, since F is a quadratic function, there must be a "downhill" direction from the current position, unless we are at the minimum. On the other hand, we hope that computing the iterates so that their function values are strictly decreasing, will force the sequence to converge to the minimum point x. Such an iterative method is called a *minimization method*.

Writing $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$, where $d^{(k)}$ is a *search direction* and α_k is a *step length*, by direct computation we get

$$F(x^{(k+1)}) = F(x^{(k)}) + \alpha_k \big(Ax^{(k)} - b, d^{(k)}\big) + \frac{\alpha_k^2}{2}\big(Ad^{(k)}, d^{(k)}\big),$$

where we used the symmetry of A to write $(Ax^{(k)}, d^{(k)}) = (x^{(k)}, Ad^{(k)})$. If the step length is so small that the second order term in α_k can be neglected, then the direction $d^{(k)}$ in which F decreases most rapidly, or the direction of *steepest descent*, is

$$d^{(k)} = -(Ax^{(k)} - b) = -r^{(k)},$$

which is the opposite direction to the residual error $r^{(k)} = Ax^{(k)} - b$. This suggests using an iterative method of the form

$$x^{(k+1)} = x^{(k)} - \alpha_k r^{(k)}. \tag{94.3}$$

A minimization method with this choice of search direction is called a *steepest descent method*. The direction of steepest descent is perpendicular to the *level curve* of F through $x^{(k)}$, which is the curve in the graph of F generated by the points where F has the same value as at $x^{(k)}$. We illustrate this in Fig. 94.10.

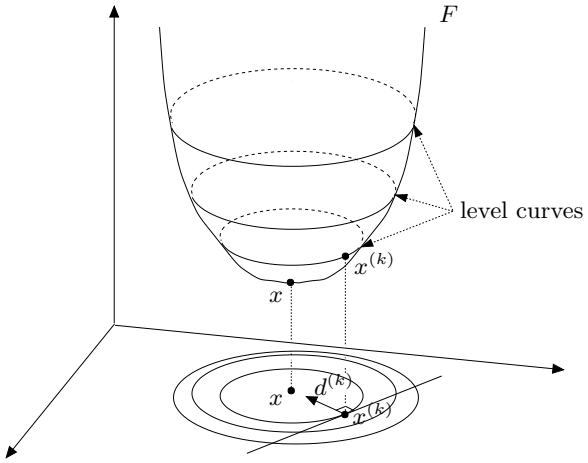


FIGURE 94.10. The direction of steepest descent of F at a point is perpendicular to the level curve of F through the point.

It remains to choose the step lengths α_k. Staying with the underlying principle, we choose α_k to give the minimum value of F in the direction of $d^{(k)}$ starting from $x^{(k)}$. Differentiating $F(x^{(k)} + \alpha_k r^{(k)})$ with respect to α_k and setting the derivative zero gives

$$\alpha_k = -\frac{(r^{(k)}, d^{(k)})}{(d^{(k)}, Ad^{(k)})}. \tag{94.4}$$

As a simple illustration, we consider the case

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \quad 0 < \lambda_1 < \lambda_2, \tag{94.5}$$

and $b = 0$, corresponding to the minimization problem

$$\min_{y \in \mathbb{R}^n} \frac{1}{2}(\lambda_1 y_1^2 + \lambda_2 y_2^2),$$

with solution $x = 0$.

Applying (94.3) to this problem, we iterate according to

$$x^{(k+1)} = x^{(k)} - \alpha_k A x^{(k)},$$

using for simplicity a constant step length with $\alpha_k = \alpha$ instead of (94.4). In Fig. 94.11, we plot the iterations computed with $\lambda_1 = 1$, $\lambda_2 = 9$, and $x^{(0)} = (9, 1)^\top$. The convergence in this case is quite slow. The reason is that if $\lambda_2 \gg \lambda_1$, then the search direction $-(\lambda_1 x_1^{(k)}, \lambda_2 x_2^{(k)})^\top$ and the direction $-(x_1^{(k)}, x_2^{(k)})^\top$ to the solution at the origin, are very different. As a result the iterates swing back and forth across the long, narrow "valley".

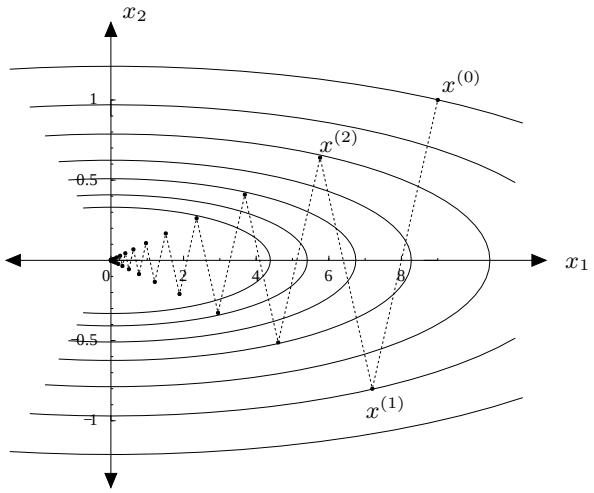


FIGURE 94.11. A sequence generated by the steepest descent method for (94.5) plotted together with some level curves of F.

It turns out that the rate at which the steepest descent method converges in general depends on the *condition number* $\kappa(A) = \lambda_n/\lambda_1$ of A, where $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ are the eigenvalues of A (counted with multiplicity). In other words, the condition number of a symmetric positive definite matrix is the ratio of the largest eigenvalue to the smallest eigenvalue.

The general definition of the condition number of a matrix A in terms of a norm $\|\cdot\|$ is $\kappa(A) = \|A\|\|A^{-1}\|$. In the $\|\cdot\|_2$ norm, the two definitions are equivalent for symmetric matrices. Using any definition, a matrix is said to be *ill-conditioned* if the $\log(\kappa(A))$ is of the order of the number of digits used in the computer. As we said, we can expect to have difficulty solving an ill-conditioned system; which in terms of direct methods means large errors due to rounding errors and in terms of iterative methods means slow convergence.

We now analyze the steepest descent method for $Ax = b$ in the case of a constant step length α, where we iterate according to

$$x^{(k+1)} = x^{(k+1)} - \alpha(Ax^{(k)} - b).$$

Since the exact solution x satisfies $x = x - \alpha(Ax - b)$, we get the following equation for the error $e^{(k)} = x - x^{(k)}$:

$$e^{(k+1)} = (I - \alpha A)e^{(k)}.$$

The iterative method converges if the error tend to zero. Taking norms, we get

$$\|e^{(k+1)}\| \leq \mu \|e^{(k)}\| \tag{94.6}$$

where we use the spectral estimate (93.16) to write

$$\mu = \|I - \alpha A\| = \max_j |1 - \alpha\lambda_j|,$$

since the eigenvalues of the matrix $I - \alpha A$ are $1 - \alpha\lambda_j$, $j = 1, ..., n$. Iterating this estimate we get

$$\|e^{(k+1)}\| \leq \mu^k \|e^{(0)}\|, \tag{94.7}$$

where $e^{(0)}$ is the initial error.

To understand when (94.6), or (94.7), guarantees convergence, consider the scalar sequence $\{\lambda^k\}$ for $k \geq 0$. If $|\lambda| < 1$, then $\lambda^k \to 0$; if $\lambda = 1$, then the sequence is always 1; if $\lambda = -1$, the sequence alternates between 1 and -1 and does not converge; and if $|\lambda| > 1$, then the sequence diverges. Therefore if we want the iteration to converge for any initial value, then we must choose α so that $\mu < 1$. Since the λ_j are positive by assumption, $1 - \alpha\lambda_j < 1$ automatically, and we can guarantee that $1 - \alpha\lambda_j > -1$ if α satisfies $\alpha < 2/\lambda_n$. Choosing $\alpha = 1/\lambda_n$, which is not so far from optimal, we get

$$\mu = 1 - 1/\kappa(A).$$

If $\kappa(A)$ is large, then the convergence can be slow because then the reduction factor $1 - 1/\kappa(A)$ is close to one. More precisely, the number of steps required to lower the error by a given amount is proportional to the condition number.

When an iteration converges in this fashion, i.e. the error decreases (more or less) by a given factor in each iteration, then we say that the iteration converges *linearly*. We define the *rate of convergence* to be $-\log(\mu)$. The motivation is that the number of iterations are required to reduce the error by a factor of 10^{-m} is approximately $-m\log(\mu)$. Note that a faster rate of convergence means a smaller value of μ.

This is an *a priori* estimate of the error reduction per iteration, since we estimate the error before the computation. Such an analysis must account for the slowest possible rate of convergence because it holds for all initial vectors.

Consider the system $Ax = 0$ with

$$A = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}, \tag{94.8}$$

where $0 < \lambda_1 < \lambda_2 < \lambda_3$. For an initial guess $x^{(0)}=(x_1^0, x_2^0, x_3^0)^\top$, the steepest descent method with $\alpha = 1/\lambda_3$ gives the sequence

$$x^{(k)} = \left(\left(1 - \frac{\lambda_1}{\lambda_3}\right)^k x_1^0, \left(1 - \frac{\lambda_2}{\lambda_3}\right)^k x_2^0, 0 \right), \quad k = 1, 2, ...,$$

and,

$$\|e^{(k)}\| = \sqrt{\left(1 - \frac{\lambda_1}{\lambda_3}\right)^{2k} \left(x_1^0\right)^2 + \left(1 - \frac{\lambda_2}{\lambda_3}\right)^{2k} \left(x_2^0\right)^2}, \quad k = 1, 2, ...$$

Thus for a general initial guess, the size of the error is given by the root mean square average of the corresponding iterate and the rate that the errors decrease is the root mean square average of the rates of decrease of the components. Therefore, depending on the initial vector, initially the iterates will generally converge more quickly than the rate of decrease of the first, i.e. slowest, component. In other words, more quickly than the rate predicted by (94.6), which bounds the rate of decrease of the errors by the rate of decrease in the slowest component. However, as the iteration proceeds, the second component eventually becomes much smaller than the first component (as long as $x_1^0 \neq 0$) and we can neglect that term in the expression for the error, i.e.

$$\|e^{(k)}\| \approx \left(1 - \frac{\lambda_1}{\lambda_3}\right)^k |x_1^0| \quad \text{for } k \text{ sufficiently large.} \tag{94.9}$$

In other words, the rate of convergence of the error for almost all initial vectors eventually becomes dominated by the rate of convergence of the slowest component. It is straightforward to show that the number of iterations that we have to wait for this approximation to be valid is determined by the relative sizes of the first and second components of $x^{(0)}$.

This simple error analysis does not apply to the unmodified steepest descent method with varying α_k. However, it is generally true that the rate of convergence depends on the condition number of A, with a larger condition number meaning slower convergence. If we again consider the 2×2 example (94.5) with $\lambda_1 = 1$ and $\lambda_2 = 9$, then the estimate (94.6) for the simplified method suggests that the error should decrease by a factor of $1 - \lambda_1/\lambda_2 \approx .89$ in each iteration. The sequence generated by $x^{(0)} = (9, 1)^\top$ decreases by exactly .8 in each iteration. The simplified analysis over-predicts the rate of convergence for this particular sequence,

though not by a lot. By way of comparison, if we choose $x^{(0)} = (1,1)^\top$, we find that the ratio of successive iterations alternates between $\approx .126$ and $\approx .628$, because α_k oscillates in value, and the sequence converges much more quickly than predicted. On the other hand, there are initial guesses leading to sequences that converge at the predicted rate.

The stiffness matrix A of a linear second order two-point boundary value problem with no convection is symmetric and positive-definite, and its condition number $\kappa(A) \propto h^{-2}$. Therefore the convergence of the steepest descent method is very slow if the number of mesh points is large.

A General Framework for Iterative Methods

We now briefly discuss iterative methods for a general, linear system $Ax = b$, following the classical presentation of iterative methods in Isaacson and Keller ([**?**]). Recall that some matrices, like diagonal and triangular matrices, are relatively easy and cheap to invert, and Gaussian elimination can be viewed as a method of factoring A into such matrices. One way to view an iterative method is an attempt to approximate A^{-1} by the inverse of a part of A that is easier to invert. This is called an approximate inverse of A, and we use this to produce an approximate solution to the linear system. Since we don't invert the matrix A, we try to improve the approximate solution by repeating the partial inversion over and over. With this viewpoint, we start by *splitting* A into two parts:

$$A = N - P,$$

where the part N is chosen so that the system $Ny = c$ for some given c is relatively inexpensive to solve. Noting that the true solution x satisfies $Nx = Px + b$, we compute $x^{(k+1)}$ from $x^{(k)}$ by solving

$$Nx^{(k+1)} = Px^{(k)} + b \quad \text{for } k = 1, 2, ..., \tag{94.10}$$

where $x^{(0)}$ is an initial guess. For example, we may choose N to be the diagonal of A:

$$N_{ij} = \begin{cases} a_{ij}, & i = j, \\ 0, & i \neq j, \end{cases}$$

or triangular:

$$N_{ij} = \begin{cases} a_{ij}, & i \geq j, \\ 0, & i < j. \end{cases}$$

In both cases, solving the system $Nx^{(k+1)} = Px^{(k)} + b$ is cheap compared to doing a complete Gaussian elimination on A. so we could afford to do it many times.

As an example, suppose that

$$A = \begin{pmatrix} 4 & 1 & 0 \\ 2 & 5 & 1 \\ -1 & 2 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \tag{94.11}$$

and we choose

$$N = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 4 \end{pmatrix} \quad \text{and} \quad P = \begin{pmatrix} 0 & -1 & 0 \\ -2 & 0 & -1 \\ 1 & -2 & 0 \end{pmatrix},$$

in which case the equation $Nx^{(k+1)} = Px^{(k)} + b$ reads

$$\begin{aligned} 4x_1^{k+1} &= -x_2^k + 1 \\ 5x_2^{k+1} &= -2x_1^k - x_3^k \\ 4x_3^{k+1} &= x_1^k - 2x_2^k + 3. \end{aligned}$$

Being a diagonal system it is easily solved, and choosing an initial guess and computing, we get

$$x^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \; x^{(1)} = \begin{pmatrix} 0 \\ -.6 \\ .5 \end{pmatrix}, \; x^{(2)} = \begin{pmatrix} .4 \\ -.1 \\ 1.05 \end{pmatrix}, \; x^{(3)} = \begin{pmatrix} .275 \\ -.37 \\ .9 \end{pmatrix},$$
$$x^{(4)} = \begin{pmatrix} .3425 \\ -.29 \\ 1.00375 \end{pmatrix}, \; \cdots \; x^{(15)} = \begin{pmatrix} .333330098 \\ -.333330695 \\ .999992952 \end{pmatrix}, \; \cdots$$

The iteration appears to converge to the true solution $(1/3, -1/3, 1)^\top$.

In general, we could choose $N = N_k$ and $P = P_k$ to vary with each iteration.

To analyze the convergence of (94.10), we subtract (94.10) from the equation $Nx = Px + b$ satisfied by the true solution to get an equation for the error $e^{(k)} = x - x^{(k)}$:

$$e^{(k+1)} = Me^{(k)},$$

where $M = N^{-1}P$ is the *iteration matrix*. Iterating on k gives

$$e^{(k+1)} = M^{k+1}e^{(0)}. \tag{94.12}$$

Rephrasing the question of convergence, we are interested in whether $e^{(k)} \to 0$ as $k \to \infty$. By analogy to the scalar case discussed above, if M is "small", then the errors $e^{(k)}$ should tend to zero. Note that the issue of convergence is independent of the data b.

If $e^{(0)}$ happens to be an eigenvector of M, then it follows from (94.12)

$$\|e^{(k+1)}\| = |\lambda|^{k+1}\|e^{(0)}\|,$$

and we conclude that if the method converges then we must have $|\lambda| < 1$ (or $\lambda = 1$). Conversely, one can show that if $|\lambda| < 1$ for all eigenvalues of M, then the method (94.10) indeed does converge:

Theorem 94.1 *An iterative method converges for all initial vectors if and only if every eigenvalue of the associated iteration matrix is less than one in magnitude.*

This theorem is often expressed using the *spectral radius* $\rho(M)$ of M, which is the maximum of the magnitudes of the eigenvalues of A. An iterative method converges for all initial vectors if and only if $\rho(M) < 1$. In general, the asymptotic limit of the ratio of successive errors computed in $\| \ \|_\infty$ is close to $\rho(M)$ as the number of iterations goes to infinity. We define the *rate of convergence* to be $R_M = -\log(\rho(M))$. The number of iterations required to reduce the error by a factor of 10^m is approximately m/R_M.

Practically speaking, "asymptotic" means that the ratio can vary as the iteration proceeds, especially in the beginning. In previous examples, we saw that this kind of a priori error result can underestimate the rate of convergence even in the special case when the matrix is symmetric and positive-definite (and therefore has an orthonormal basis of eigenvectors) and the iterative method uses the steepest descent direction. The general case now considered is more complicated, because interactions may occur in direction as well as magnitude, and a spectral radius estimate may overestimate the rate of convergence initially. As an example, consider the non-symmetric (even non-normal) matrix

$$A = \begin{pmatrix} 2 & -100 \\ 0 & 4 \end{pmatrix} \tag{94.13}$$

choosing

$$N = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix} \text{ and } P = \begin{pmatrix} 8 & 100 \\ 0 & 6 \end{pmatrix} \text{ gives } M = \begin{pmatrix} .9 & 10 \\ 0 & .8 \end{pmatrix}.$$

In this case, $\rho(M) = .9$ and we expect the iteration to converge. Indeed it does converge, but the errors become quite large before they start to approach zero. We plot the iterations starting from $x^{(0)} = (1, 1)^\top$ in Fig. 94.12.

The goal is obviously to choose an iterative method so that the spectral radius of the iteration matrix is small. Unfortunately, computing $\rho(M)$ in the general case is much more expensive than solving the original linear system and is impractical in general. We recall that $|\lambda| \leq \|A\|$ holds for any norm and any eigenvalue λ of A. The following theorem indicates a practical way to check for convergence.

Theorem 94.2 *Assume that* $\|N^{-1}P\| \leq \mu$ *for some constant* $\mu < 1$ *and matrix norm* $\| \cdot \|$. *Then the iteration converges and* $\|e^{(k)}\| \leq \mu^k \|e^{(0)}\|$ *for* $k \geq 0$.

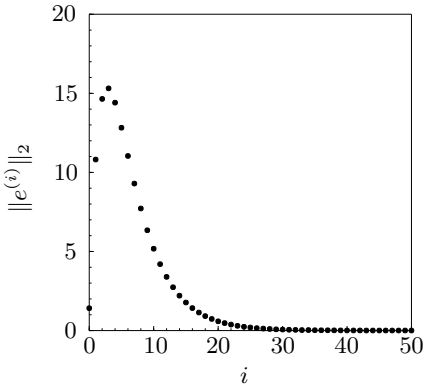


FIGURE 94.12. The results of an iterative method computed using a non-normal matrix.

This theorem is also an a priori convergence result and suffers from the same deficiency as the analysis of the simplified steepest descent method presented above. In fact, choosing an easily computable matrix norm, like $\| \ \|_\infty$, generally leads to an even more inaccurate estimate of the convergence rate than would be obtained by using the spectral radius. In the worst case, it is entirely possible that $\rho(M) < 1 < \|M\|$ for the chosen norm, and hence the iterative method converges even though the theorem does not apply. The amount of "slack" in the bound in Theorem 94.2 depends on how much larger $\|A\|_\infty$ is than $\rho(A)$.

For the 3×3 example (94.11), we compute $\|N^{-1}P\|_\infty = 3/4 = \lambda$ and therefore we know the sequence converges. The theorem predicts that the error will get reduced by a factor of $3/4$ every iteration. If we examine the error of each iterate along with the ratios of successive errors after the first iteration:

i	$\|e^{(i)}\|_\infty$	$\|e^{(i)}\|_\infty / \|e^{(i-1)}\|_\infty$
0	1.333	
1	.5	.375
2	.233	.467
3	.1	.429
4	.0433	.433
5	.0194	.447
6	.00821	.424
7	.00383	.466
8	.00159	.414
9	.000772	.487

we find that after the first few iterations, the errors get reduced by a factor in the range of .4–.5 each iteration and not the factor $3/4$ predicted above. The ratio of $e^{(40)}/e^{(39)}$ is approximately .469. If we compute the eigenvalues of M, we find that $\rho(M) \approx .476$ which is close to the ratio of successive

errors. To decrease the initial error by a factor of 10^{-4} using the predicted decrease of .75 per iteration, we would compute 33 iterations, while only 13 iterations are actually needed.

We get different methods, and different rates of convergence, by choosing different N and P. The method used in the example above is called the *Jacobi* method. In general, this consists of choosing N to be the "diagonal part" of A and P to be the negative of the "off-diagonal" part of A. This gives the set of equations

$$x_i^{k+1} = -\frac{1}{a_{ii}}\left(\sum_{j\neq i} a_{ij}x_j^k - b_i\right), \quad i = 1, ..., n.$$

To derive a more sophisticated method, we write out these equations in Fig. 94.13. The idea behind the *Gauss-Seidel* method is to use the new

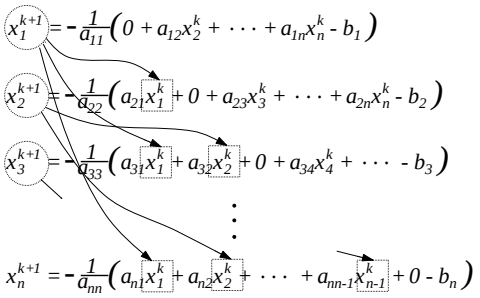


FIGURE 94.13. The Gauss-Seidel method substitutes new values of the iteration as they become available.

values of the approximation in these equations as they become known. The substitutions are drawn in Fig. 94.13. Presumably, the new values are more accurate than the old values, hence we might guess that this iteration will converge more quickly. The equations can be written

$$x_i^{k+1} = \frac{1}{a_{ii}}\left(-\sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}x_j^k + b_i\right).$$

If we decompose A into the sum of its lower triangular L, diagonal D, and upper triangular U parts, $A = L+D+U$, then the equations can be written $Dx^{(k+1)} = -Lx^{(k+1)} - Ux^{(k)} + b$ or

$$(D + L)x^{(k+1)} = -Ux^{(k)} + b.$$

Therefore, $N = D + L$ and $P = -U$. The iteration matrix is $M_{\mathrm{GS}} = N^{-1}P = -(D + L)^{-1}U$.

A diagonally dominant matrix often occurs when a parabolic problem is discretized. We have already seen the other case, if A is symmetric and positive-definite then the Gauss-Seidel method converges. This is quite hard to prove, see Isaacson and Keller ([**?**]) for details.

94.5 Estimating the Error of the Solution

The issue of estimating the error of the numerical solution of a linear system $Ax = b$ arises both in Gaussian elimination, because of the cumulative effects of round-off errors, and when using iterative methods, where we need a stopping criterion. Therefore it is important to be able to estimate the error in some norm with a fair degree of accuracy.

We discussed this problem in the context of iterative methods in the last section when we analyzed the convergence of iterative methods and Theorem 94.2 gives an *a priori* estimate for the convergence rate. It is an a priori estimate because the error is bounded before the computation begins. Unfortunately, as we saw, the estimate may not be very accurate on a particular computation, and it also requires the size of the initial error. In this section, we describe a technique of *a posteriori* error estimation that uses the approximation after it is computed to give an estimate of the error of that particular approximation.

We assume that x_c is a numerical solution of the system $Ax = b$ with exact solution x, and we want to estimate the error $\|x - x_c\|$ in some norm $\|\cdot\|$. We should point out that we are actually comparing the approximate solution $\tilde{x}_c$ of $\tilde{A}\tilde{x} = \tilde{b}$ to the true solution $\tilde{x}$, where $\tilde{A}$ and $\tilde{b}$ are the finite precision computer representations of the true A and b respectively. The best we can hope to do is compute $\tilde{x}$ accurately. To construct a complete picture, it would be necessary to examine the effects of small errors in A and b on the solution x. To simplify things, we ignore this part of the analysis and drop the $\tilde{}$. In a typical use of an iterative method, this turns out to be reasonable. It is apparently less reasonable in the analysis of a direct method, since the errors arising in direct methods are due to the finite precision. However, the initial error caused by storing A and b on a computer with a finite number of digits occurs only once, while the errors in the arithmetic operations involved in Gaussian elimination occur many times, so even in that case it is not an unreasonable simplification.

We start by considering the *residual error*

$$r = Ax_c - b,$$

which measures how well x_c solves the exact equation. Of course, the residual error of the exact solution x is zero but the residual error of x_c is not zero unless $x_c = x$ by some miracle. We now seek to estimate the unknown error $e = x - x_c$ in terms of the computable residual error r.

By subtracting $Ax - b = 0$ from $Ax_c - b = r$, we get an equation relating the error to the residual error:

$$Ae = -r. \tag{94.14}$$

This is an equation of the same from as the original equation and by solving it numerically by the same method used to compute x_c, we get an approximation of the error e. This simple idea will be used in a more sophisticated form below in the context of a posteriori error estimates for Galerkin methods.

We now illustrate this technique on the linear system arising in the Galerkin finite element discretization of a two-point boundary value problem with no convection. We generate a problem with a known solution so that we can compute the error and test the accuracy of the error estimate. We choose the true solution vector x with components $x_i = \sin(\pi ih)$, where $h = 1/(M+1)$, corresponding to the function $\sin(\pi x)$ and then compute the data by $b = Ax$, where A is the stiffness matrix. We use the Jacobi method, suitably modified to take advantage of the fact that A is tridiagonal, to solve the linear system. We use $\| \| = \| \|_2$ to measure the error.

We compute the Jacobi iteration until the residual error becomes smaller than a given *residual tolerance* RESTOL. In other words, we compute the residual $r^{(k)} = Ax^{(k)} - b$ after each iteration and stop the process when $\|r^{(k)}\| \leq$ RESTOL. We present computations using the stiffness matrix generated by a uniform discretization with $M = 50$ elements yielding a finite element approximation with an error of .0056 in the l_2 norm. We choose the value of RESTOL so that the error in the computation of the coefficients of the finite element approximation is about 1% of the error of the approximation itself. This is reasonable since computing the coefficients of the approximation more accurately would not significantly increase the overall accuracy of the approximation. After the computation of $x^{(k)}$ is complete, we use the Jacobi method to approximate the solution of (94.14) and compute the estimate of the error.

Using the initial vector $x^{(0)}$ with all entries equal to one, we compute 6063 Jacobi iterations to achieve $\|r\| <$ RESTOL $= .0005$. The actual error of $x^{(6063)}$, computed using the exact solution, is approximately .0000506233. We solve (94.14) using the Jacobi method for 6063 iterations, reporting the value of the error estimate every 400 iterations:

Iter.	Error Est.	Iter.	Error Est.	Iter.	Error Est.
1	0.00049862	2001	0.000060676	4001	0.000050849
401	0.00026027	2401	0.000055328	4401	0.000050729
801	0.00014873	2801	0.000052825	4801	0.000050673
1201	0.000096531	3201	0.000051653	5201	0.000050646
1601	0.000072106	3601	0.000051105	5601	0.000050634

We see that the error estimate is quite accurate after 6001 iterations and sufficiently accurate for most purposes after 2000 iterations. In general,

we do not require as much accuracy in the error estimate as we do in the solution of the system, so the estimation of the accuracy of the approximate solution is cheaper than the computation of the solution.

Since we estimate the error of the computed solution of the linear system, we can stop the Jacobi iteration once the error in the coefficients of the finite element approximation is sufficiently small so that we are sure the accuracy of the approximation will not be affected. This is a reasonable strategy given an estimate of the error. If we do not estimate the error, then the best strategy to guarantee that the approximation accuracy is not affected by the solution error is to compute the Jacobi iteration until the residual error is on the order of roughly 10^{-p}, where p is the number of digits that the computer uses. Certainly, there is not much point to computing further Jacobi iterations after this. If we assume that the computations are made in single precision, then $p \approx 8$. It takes a total of 11672 Jacobi iterations to achieve this level of residual error using the same initial guess as above. In fact, estimating the error and computing the coefficients of the approximation to a reasonable level of accuracy costs significantly less than this crude approach.

This approach can also be used to estimate the error of a solution computed by a direct method, provided the effects of finite precision are included. The added difficulty is that in general the residual error of a solution of a linear system computed with a direct method is small, even if the solution is inaccurate. Therefore, care has to be taken when computing the residual error because the possibility that subtractive cancellation makes the calculation of the residual error itself inaccurate. *Subtractive cancellation* is the name for the fact that the difference of two numbers that agree to the first i places has i leading zeroes. If only the first p digits of the numbers are accurate then their difference can have at most $p - i$ accurate significant digits. This can have severe consequences on the accuracy of the residual error if Ax_c and b agree to most of the digits used by the computer. One way to avoid this trouble is to compute the approximation in single precision and the residual in double precision (which means compute the product Ax_c in double precision, then subtract b). The actual solution of (94.14) is relatively cheap since the factorization of A has already been performed and only forward/backward substitution needs to be done.

94.6 The Conjugate Gradient Method

We learned above that solving an $n \times n$ linear system of equations $Ax = b$ with A symmetric positive definite using the gradient method, requires a number of iterations, which is proportional to the condition number $\kappa(A) = \lambda_n/\lambda_1$, where $\lambda_1 \leq ... \leq \lambda_n$ are the eigenvalues of A. Thus the number of

iteration will be large, maybe prohibitively so, if the condition number $\kappa(A)$ is large.

We shall now present a variant of the gradient method, referred as the *conjugate gradient method*, where the number of iterations scales instead like $\sqrt{\kappa(A)}$, which may be much smaller than $\kappa(A)$ if $\kappa(A)$ is large.

In the conjugate gradient method each new search direction is chosen to be orthogonal, with respect to the scalar product induced by the positive definite symmetric matrix A, which prevents choosing inefficient search directions as in the usual gradient method.

The conjugate gradient method may be formulated as follows: for $k = 1, 2, ...$ compute an approximate solution $x^k \in \mathbb{R}^n$ as the solution of the minimization problem

$$\min_{y \in K_k(A)} F(y) = \min_{y \in K_k(A)} \frac{1}{2}(Ay, y) - (b, y)$$

where $K_k(A)$ is the *Krylov space* spanned by the vectors $\{b, Ab, ..., A^{k-1}b\}$.

This is the same as defining x^k to be the projection of x onto $K_k(A)$ with respect to the scalar product $< y, z >$ on $\mathbb{R}^n \times \mathbb{R}^n$ defined by $< y, z >= (Ay, z)$, because we have using the symmetry of A and that $Ax = b$:

$$\frac{1}{2}(Ay, y) - (b, y) = \frac{1}{2} < y - x, y - x > - \frac{1}{2} < x, x > .$$

In particular, the conjugate gradient method has the following minimization property

$$\|x - x^k\|_A = \min_{y \in K_k(A)} \|x - y\|_A \leq \|p_k(A)x\|_A$$

where $p_k(x)$ is a polynomial of degree k with $p(0) = 1$, and $\|\cdot\|_A$ is the norm associated with the scalar product $< \cdot, \cdot >$, that is, $\|y\|_A^2 =< y, y >$. This follows by using that since $b = Ax$, we have that $K_k(A)$ is spanned by the vectors $\{Ax, A^2x, ..., A^kx\}$. In particular, we conclude that for all polynomials $p_k(x)$ of degree k such that $p_k(0) = 1$, we have

$$\|x - x^k\|_A \leq \max_{\lambda \in \Lambda} |p_k(\lambda)| \|x\|_A \tag{94.15}$$

where Λ is the set of eigenvalues of A. By choosing the polynomial $p_k(x)$ properly, e.g as a so-called *Chebyshev polynomial* $q_k(x)$ with the property that $q_k(x)$ is small on the interval $[\lambda_1, \lambda_n]$ containing the eigenvalues of A, one can prove that the number of iterations scales like $\sqrt{\kappa(A)}$ if n is large.

If n is not large, we have in particular from (94.15) that we get the exact solution after at most n iterations, since we may choose the polynomial $p_k(x)$ to be zero at the n eigenvalues of A.

We have now defined the conjugate gradient method through it structural properties: projection onto a Krylov space with respect to a certain

scalar product, and we now address the problem of actually computing the sequence x^k step by step. This is done as follows: For $k = 0, 1, 2, ...,$

$$x^{k+1} = x^k + \alpha_k d^k, \quad \alpha_k = -\frac{(r^k, d^k)}{< d^k, d^k >}, \tag{94.16}$$

$$d^{k+1} = -r^{k+1} + \beta_k d^k, \quad \beta_k = \frac{< r^{k+1}, d^k >}{< d^k, d^k >}, \tag{94.17}$$

where $r^k = Ax^k - b$ is the residual of the approximation x^k, and we choose $x^0 = 0$ and $d_0 = b$. Here, (94.17) signifies that the new search direction d^{k+1} gets new directional information from the new residual r^{k+1} and is chosen to be orthogonal (with respect to the scalar product $< \cdot, \cdot >$) to the old search direction d^k. Further, (94.16), expresses that x^{k+1} is chosen so as to to minimize $F(x^{(k)} + \alpha d^k)$ in α, corresponding to projection onto $K_{k+1}(A)$. We prove these properties in a sequence of problems below.

Note that if we choose the initial approximation x^0 different from zero, then we may reduce to the above case by considering instead the problem $Ay = b - Ax^0$ in y, where $y = x - x^0$.

94.7 GMRES

The conjugate gradient method for solving an $n \times n$ system $Ax = b$ builds on the matrix A being symmetric and positive definite. If A is non-symmetric or non-positive definite, but yet non-singular, then we may apply the conjugate gradient method to the least squares problem $A^\top Ax = A^\top b$, but the since the condition number of $A^\top A$ typically is the square of the condition number of A, the required number of iterations may be too large for efficiency.

Instead we may try the *Generalized Minimum Residual* method referred to as *GMRES*, which generates a sequence of approximations x^k of the solution x of $Ax = b$, satisfying for any polynomial $p_k(x)$ of degree at most k with $p_k(0) = 1$

$$\|Ax^k - b\| = \min_{y \in K_k(A)} \|Ay - b\| \leq \|p_k(A)b\|, \tag{94.18}$$

that is x^k is the element in the Krylov space $K_k(A)$ which minimizes the Euclidean norm of the residual $Ay - b$ with $y \in K_k(A)$. Assuming that the matrix A is *diagonalizable*, there exist a nonsingular matrix V so that $A = VDV^{-1}$, where D is a diagonal matrix with the eigenvalues of A on the diagonal. We then have that

$$\|Ax^k - b\| \leq \kappa(V) \max_{\lambda \in \Lambda} |p_k(\lambda)| \|b\|, \tag{94.19}$$

where Λ is the set of eigenvalues of A.

In the actual implementation of GMRES we use the *Arnoldi iteration*, a variant of the Gram-Schmidt orthogonalization, that constructs a sequence of matrices Q_k whose orthogonal column vectors span the successive Krylov spaces $K_k(A)$, and we write $x^k = Q_k c$ to get the following least squares problem:

$$\min_{c\in\mathbb{R}^k} \|AQ_n c - b\|. \tag{94.20}$$

The Arnoldi iteration is based on the identity $AQ_k = Q_{k+1}H_k$, where H_k is an *upper Hessenberg matrix* so that $h_{ij} = 0$ for all $i > j + 1$. Using this identity and multiplying from the left by Q_{k+1}^T gives us another equivalent least squares problem:

$$\min_{c\in\mathbb{R}^k} \|H_k c - Q_{k+1}^T b\|. \tag{94.21}$$

Recalling the construction of the Krylov spaces $K_k(A)$, in particular that $K_1(A)$ is spanned by b, we find that $Q_{k+1}^T b = \|b\|e_1$, where $e_1 = (1, 0, 0, ...)$, and we obtain the final form of the least squares problem to be solved in the GMRES iteration:

$$\min_{c\in\mathbb{R}^k} \|H_k c - \|b\|e_1\|. \tag{94.22}$$

This problem is now easy to solve due to the simpel structure of the Hessenberg matrix H_k.

In Figure 94.14 we compare the performance of the conjugate gradient method and GMRES for system with a tridiagonal 200×200 matrix with 1 on the diagonal, and random off-diagonal entries that take values in $(-0.5, 0.5)$ and the right hand side a random vector with values in $[-1, 1]$. The system matrix in this case is not symmetric, but it is strictly diagonally dominant and thus may be viewed asa perturbation of the identity matrix and should be easy to solve iteratively. We see that both the conjugate gradient method and GMRES converge quite rapidly, with GMRES winning in number of iterations.

In GMRES we need to store the basis vectors for the increasing Krylov space, which may be prohibitive for large systems requiring many iterations. To avoid this problem, we may restart GMRES when we have reached a maximal number of stored basis vector, by using as initial approximation x^0 the last approximation before restart. The trade-off is of course that a retarted GMRES may require more iterations for the same accuracy than GMRES without restart.

We now consider the more challenging problem of solving a 200×200 *stiffness matrix* system, that is a system with a tridiagonal matrix with 2 on the diagonal, and -1 on the off-diagonal (which is not strictly diagonally dominant). We will meet this type of system matrix in Chapter FEM for Two-Point Boundary Value Problems below, and we will see that it has a condition number proportional to the square of the number of unknowns. We thus expect the conjugate gradient method to require about

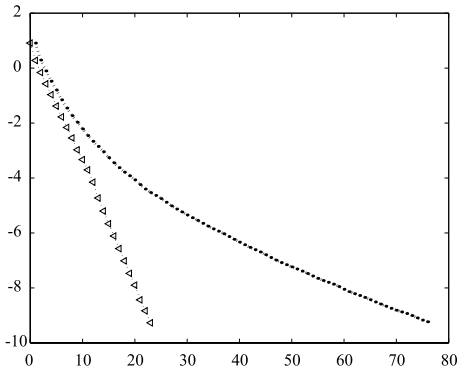


FIGURE 94.14. Log-plot of the residual versus the number of iterations for diagonal dominant random matrix, using the conjugate gradient method ($'\cdot'$) and GMRES ('triangles').

the same number of iterations as the number of unknowns. In Figure 94.15 we compare again the performance of the conjugate gradient method with the GMRES method, now restarted after 100 iterations. We find that the conjugate gradient method as expected converges quite slowly (and non monotonically), until immediate convergence at iteration 200 as predicted by theory. The GMRES iteration on the other hand has a monotone but still quite slow convergence in particular after each restart when the Krylov subspace is small.

In Figure 94.16 we compare different restart conditions for GMRES, and we find that there is a trade-off between the convergence rate and the memory consumption: few restarts give a faster convergence, but require more memory to store more basis vectors for the Krylov space. On the other hand we save memory by using more restarts, but then the convergence rate deteriorates.

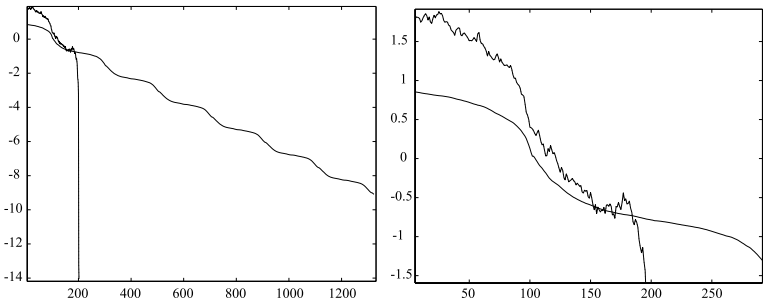


FIGURE 94.15. Log-plot of the residual versus the number of iterations for stiffness matrix, using the conjugate gradient method and GMRES, restarted after 100 iterations.

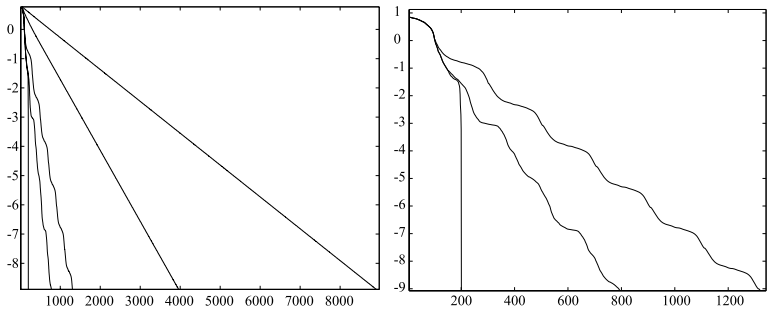


FIGURE 94.16. Log-plot of the residual versus the number of iterations for stiffness matrix using GMRES and restarted GMRES, restarted after 20,50,100,150 iterations (left), and a close-up on the cases of no restart and restart after 100 and 150 iterations (right).

Chapter 94 Problems

94.1. Using a similar format, write down algorithms to solve a diagonal system and then a lower triangular system using forward substitution. Determine the number of arithmetic operations needed to compute the solution.

94.2. Prove that multiplying a square matrix A on the left by the matrix in Fig. 94.3 has the effect of adding α_{ij} times row j of A to row i of A. Prove that the inverse of the matrix in Fig. 94.3 is obtained changing α_{ij} to $-\alpha_{ij}$

94.3. Show that the product of two Gauss transformations is a lower triangular matrix with ones on the diagonal and the inverse of a Gauss transformation is a Gauss transformation.

94.4. Solve the system

$$\begin{aligned} x_1 - x_2 - 3x_3 &= 3 \\ -x_1 + 2x_2 + 4x_3 &= -5 \\ x_1 + x_2 &= -2 \end{aligned}$$

by computing an LU factorization of the coefficient matrix and using forward/backward substitution.

94.5. On some computers, dividing two numbers is up to ten times more expensive than computing the reciprocal of the denominator and multiplying the result with the numerator. Alter this code to avoid divisions. Note: the reciprocal of the diagonal element a_{kk} has to be computed just once.

94.6. Write some pseudo-code that uses the matrix generated by the code in Fig. 94.4 to solve the linear system $Ax = b$ using forward/backward substitution. Hint: the only missing entries of L are the 1s on the diagonal.

94.7. Show that the cost of a backward substitution using an upper triangular matrix of dimension $n \times n$ is $O(n^2/2)$.

94.8. Determine the cost of multiplying a $n \times n$ matrix with another.

94.9. One way to compute the inverse of a matrix is based on viewing the equation $AA^{-1} = I$ as a set of linear equations for the columns of A^{-1}. If $a^{(j)}$ denotes the j^{th} column of A^{-1}, then it satisfies the linear system

$$Aa^{(j)} = e_j$$

where e_j is the standard basis vector of $\mathbb{R}^n$ with a one in the j^{th} position. Use this idea to write a pseudo-code for computing the inverse of a matrix using LU factorization and forward/backward substitution. Note that it suffices to compute the LU factorization only once. Show that the cost of computing the inverse in this fashion is $O(4n^3/3)$.

94.10. Solve the system

$$\begin{aligned} x_1 + x_2 + x_3 &= 2 \\ x_1 + x_2 + 3x_3 &= 5 \\ -x_1 - 2x_3 &= -1. \end{aligned}$$

This requires pivoting.

94.11. Alter the LU decomposition and forward/backward routines to solve a linear system with pivoting.

94.12. Modify the code in Problem 94.11 to use partial pivoting.

94.13. Count the cost of Cholesky's method.

94.14. Compute the Cholesky factorization of

$$\begin{pmatrix} 4 & 2 & 1 \\ 2 & 3 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

94.15. Show that the operations count for solving a tridiagonal system using the solver described in Fig. 94.9 is $O(5n)$.

94.16. Find an algorithm to solve a tridiagonal system that stores only four vectors of dimension n.

94.17. A factorization of a tridiagonal solver can be derived as a compact method. Assume that A can be factored as

$$A = \begin{pmatrix} \alpha_1 & 0 & & \cdots & & 0 \\ \beta_2 & \alpha_2 & 0 & & & \vdots \\ 0 & \beta_3 & \alpha_3 & & & \\ \vdots & & & \ddots & 0 & \\ 0 & \cdots & 0 & \beta_n & \alpha_n & \end{pmatrix} \begin{pmatrix} 1 & \gamma_1 & 0 & \cdots & & 0 \\ 0 & 1 & \gamma_2 & 0 & & \\ \vdots & & \ddots & \ddots & & \\ & & & 1 & \gamma_{n-1} & \\ 0 & \cdots & & 0 & 1 & \end{pmatrix}$$

Multiply out the factors and equate the coefficients to get equations for α, β, and γ. Derive some code based on these formulas.

94.18. Write some code to solve the tridiagonal system resulting from the Galerkin finite element discretization of a two-point boundary value problem. Using 50 elements, compare the time it takes to solve the system with this tridiagonal solver to the time using a full LU decomposition routine.

94.19. Show that the operations count of a banded solver for a $n \times n$ matrix with bandwidth d is $O(nd^2/2)$.

94.20. Write code to solve a linear system with bandwidth five centered around the main diagonal. What is the operations count for your code?

94.21. Prove that the solution of (94.2) is also the solution of $Ax = b$.

94.22. Prove that the direction of steepest descent for a function F at a point is perpendicular to the level curve of F through the same point.

94.23. Prove (94.4).

94.24. Prove that the level curves of F in the case of (94.5) are ellipses with major and minor axes proportional to $1/\sqrt{\lambda_1}$ and $1/\sqrt{\lambda_2}$, respectively.

94.25. Compute the iteration corresponding to $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 3$, and $x^{(0)} = (1, 1, 1)^\top$ for the system $Ax = 0$ with A defined in (94.8). Make a plot of the ratios of successive errors versus the iteration number. Do the ratios converge to the ratio predicted by the error analysis?

94.26. Prove that the estimate (94.9) generalizes to any symmetric positive-definite matrix A, diagonal or not. Hint: use the fact that there is a set of eigenvectors of A that form an orthonormal basis for $\mathbb{R}^n$ and write the initial vector in terms of this basis. Compute a formula for the iterates and then the error.

94.27. (a) Compute the steepest descent iterations for (94.5) corresponding to $x^{(0)} = (9, 1)^\top$ and $x^{(0)} = (1, 1)^\top$, and compare the rates of convergence. Try to make a plot like Fig. 94.11 for each. Try to explain the different rates of convergence.

(b) Find an initial guess which produces a sequence that decreases at the rate predicted by the simplified error analysis.

94.28. Prove that the method of steepest descent corresponds to choosing

$$N = N_k = \frac{1}{\alpha_k} I, \text{ and } P = P_k = \frac{1}{\alpha_k} I - A,$$

with suitable α_k in the general iterative solution algorithm.

94.29. Compute the eigenvalues and eigenvectors of the matrix A in (94.13) and show that A is not normal.

94.30. Prove that the matrix $\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ is normal.

94.31. Prove Theorem 94.2.

94.32. Compute 10 Jacobi iterations using the A and b in (94.11) and the initial guess $x^{(0)} = (-1,\ 1,\ -1)^\top$. Compute the errors and the ratios of successive errors and compare to the results above.

94.33. Repeat Problem 94.32 using

$$A = \begin{pmatrix} 4 & 1 & 100 \\ 2 & 5 & 1 \\ -1 & 2 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}.$$

Does Theorem 94.2 apply to this matrix?

94.34. Show that for the Jacobi iteration, $N = D$ and $P = -(L + U)$ and the iteration matrix is $M_{\mathrm{J}} = -D^{-1}(L + U)$

94.35. (a) Solve (94.11) using the Gauss-Seidel method and compare the convergence with that of the Jacobi method. Also compare $\rho(M)$ for the two methods. (b) Do the same for the system in Problem 94.33.

94.36. (Isaacson and Keller ([**?**])) Analyze the convergence of the Jacobi and Gauss-Seidel methods for the matrix

$$A = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

in terms of the parameter ρ.

In general it is difficult to compare the convergence of the Jacobi method with that of the Gauss-Seidel method. There are matrices for which the Jacobi method converges and the Gauss-Seidel method fails and vice versa. There are two special classes of matrices for which convergence can be established without further computation. A matrix A is *diagonally dominant* if

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, ..., n.$$

If A is diagonally dominant then the Jacobi method converges.

94.37. Prove this claim.

94.38. Derive an algorithm that uses the Jacobi method to solve a tridiagonal system. Use as few operations and as little storage as possible.

94.39. Devise an algorithm to estimate the error of the solution of a linear system using single and double precision as suggested. Repeat the example using a tridiagonal solver and your algorithm to estimate the error.

94.40. Show that the sequences $\{x^k\}$ and $\{d^k\}$ generated by the conjugate gradient method (94.16)-(94.17), with $x^1 = 0$ and $d^1 = b$, satisfies for $k = 1, 2, ...,$ (a) $x^k \in K_k(A) = \{b, ..., A^{k-1}b\}$, (b) d^{k+1} is orthogonal to $K_k(A)$, (c) x^k is the projection of x onto $K_k(A)$ with respect to the scalar product $< y, z >= (Ay, z)$.

94.41. The Cbebyshev polynomial $q_k(x)$ if degree k is defined for $-1 \leq x \leq 1$ by the formula $q_k(x) = \cos(k \arccos(x))$. Show that $q_k'(0) \approx k^2$. Deduce from this result that the number of iterations in the conjugate gradient method scales like $\sqrt{\kappa A)}$.

94.42. Compare the GMRES-algorithm for $Ax = b$ with the conjugate gradient method fro the normal equations $A^\top A = A^\top b$.

94.43. The formula $AQ_k = Q_{k+1}H_k$, with H_k an upper Hessenberg matrix ($h_{ij} = 0$ for all $i > j + 1$), defines a recurrence relation for the column vector q_{k+1} of Q_{k+1} in terms of itself and the previous Krylov vectors. (a) Derive this recurrence relation. (b) Implement an algorithm that computes Q_{k+1} and H_k, given a matrix A (this is the *Arnoldi iteration*).

94.44. Prove that $Q_{k+1}^T b = \|b\|e_1$.

94.45. Implement the GMRES-method.

Part VI

Tool Bags

95
1D Calculus

95.1 Integers

We start with the set of integers $\mathbb{Z} = \{\cdots, -3, -2, -1, 0, 1, 2, 3, \cdots\}$ together with the usual operations of addition, subtraction and multiplication. We define the set of rational numbers $\mathbb{Q}$ as the set of pairs (p, q) with p and $q \neq 0$ integers, and we write $(p, q) = \frac{p}{q}$ along with the arithmetic operations of addition

$$\frac{p}{q} + \frac{r}{s} = \frac{ps + qr}{qs},$$

multiplication

$$\frac{p}{q} \times \frac{r}{s} = \frac{pr}{qs},$$

and division

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

assuming $r \neq 0$. With the operation of division, we can solve the equation $ax = b$ to get $x = b/a$ for $a, b \in \mathbb{Q}$ with $a \neq 0$.

Rational numbers have periodic decimal expansions. There is no rational number x such that $x^2 = 2$.

95.2 Real numbers. Sequences and Limits

Definitions: A real number is specified by an *infinite decimal expansion* of the form

$$\pm p_m \cdots p_0.q_1 q_2 q_3 \cdots$$

with a never ending list of decimals q_1, q_2, ..., where each of the p_i and q_j are one of the 10 digits $0, 1, ..., 9$. The set of (all possible) real numbers is denoted by $\mathbb{R}$.

A sequence $\{x_i\}_{i=1}^{\infty}$ of real numbers converges to a real number x if for any $\epsilon > 0$ there is a natural number N such that $|x_i - x| < \epsilon$ for $i \geq N$ and we then write $x = \lim_{i\to\infty} x_i$.

A sequence $\{x_i\}_{i=1}^{\infty}$ of real numbers is a Cauchy sequence if for all $\epsilon > 0$ there is a natural number N such that

$$|x_i - x_j| \leq \epsilon \quad \text{for} \quad i, j \geq N.$$

Basic properties: A convergent sequence of real numbers is a Cauchy sequence. A Cauchy sequence of real numbers converges to a unique real number. We have $\lim_{i\to\infty} x_i = x$, where $\{x_i\}_{i=1}^{\infty}$ is the sequence of truncated decimal expansions of x.

95.3 Polynomials and Rational Functions

A polynomial function $f : \mathbb{R} \to \mathbb{R}$ of degree n has the form $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ with coefficients $a_i \in \mathbb{R}$. A rational function $h(x)$ has the form $h(x) = f(x)/g(x)$, where $f(x)$ and $g(x)$ are polynomials.

95.4 Lipschitz Continuity

Definition: A function $f : I \to \mathbb{R}$, where I is an interval of real numbers, is Lipschitz continuous on I with Lipschitz constant $L_f \geq 0$ if

$$|f(x_1) - f(x_2)| \leq L_f |x_1 - x_2| \quad \text{for all } x_1, x_2 \in I.$$

Basic facts: Polynomial functions are Lipschitz continuous on bounded intervals. Sums, products and composition of Lipschitz continuous functions are Lipschitz continuous. Quotients of Lipschitz continuous functions are Lipschitz continuous on intervals where the denominator is bounded away from zero. A Lipschitz continuous function $f : I \to \mathbb{R}$, where I is an interval of real numbers, satisfies:

$$f(\lim_{i\to\infty} x_i) = \lim_{i\to\infty} f(x_i),$$

for any convergent sequence $\{x_i\}$ in I with $\lim_{i\to\infty} x_i \in I$.

95.5 Derivatives

Definition:The function $f : (a, b) \to \mathbb{R}$ is differentiable at $\bar{x} \in (a, b)$ with derivative $f'(\bar{x}) = \frac{df}{dx}(\bar{x})$ if there are real numbers $f'(\bar{x})$ and $K_f(\bar{x})$ such that for $x \in (a, b)$ close to $\bar{x}$,

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

If the constant $K_f(\bar{x})$ can be chosen independently of $\bar{x} \in (a, b)$, then $f : (a, b) \to \mathbb{R}$ is said to be uniformly differentiable on (a, b).

Derivative of x^α with $\alpha \neq 0$: The derivative of $f(x) = x^\alpha$ is $f'(x) = \alpha x^{\alpha-1}$ for $\alpha \neq 0$, and $x \neq 0$ for $\alpha < 1$.

Bounded derivative implies Lipschitz continuity: If $f(x)$ is uniformly differentiable on the interval $I = (a, b)$ and there is a constant L such that

$$|f'(x)| \leq L, \quad \text{for } x \in I,$$

then $f(x)$ is Lipschitz continuous on I with Lipschitz constant L.

95.6 Differentiation Rules

Linear Combination rule:

$$(f + g)'(x) = f'(x) + g'(x),$$
$$(cf)'(x) = cf'(x),$$

where c is a constant.

Product rule:

$$(fg)'(x) = f(x)g'(x) + f'(x)g(x).$$

Chain rule:

$$(f \circ g)'(x) = f'(g(x))g'(x), \quad \text{or}$$
$$\frac{dh}{dx} = \frac{df}{dy}\frac{dy}{dx},$$

where $h(x) = f(y)$ and $y = g(x)$, that is $h(x) = f(g(x)) = (f \circ g)(x)$.

Quotient rule:

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2},$$

provided $g(x) \neq 0$.

The derivative of an inverse function:

$$\frac{d}{dy}f^{-1}(y) = \frac{1}{\frac{d}{dx}f(x)}\,.$$

where $y = f(x)$ and $x = f^{-1}(y)$.

95.7 Solving $f(x) = 0$ with $f : \mathbb{R} \to \mathbb{R}$

Bisection: If $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous on $[a, b]$ and $f(a)f(b) < 0$, then the Bisection algorithm converges to a root $\bar{x} \in [a, b]$ of $f(x) = 0$.

Fixed Point Iteration: A Lipschitz continuous function $g : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant $L < 1$ is said to be a contraction mapping. A contraction mapping $g : \mathbb{R} \to \mathbb{R}$ has a unique fixed point $\bar{x} \in \mathbb{R}$ satisfying $\bar{x} = g(\bar{x})$ and any sequence $\{x_i\}_{i=1}^{\infty}$ generated by Fixed Point Iteration $x_i = g(x_{i-1})$ converges to $\bar{x}$.

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 $f(\bar{x}) = 0$ (consequence of Bisection above).

Newton's method: Newton's method $x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$ for computing a root $\bar{x}$ of $f : \mathbb{R} \to \mathbb{R}$ converges quadratically if $f'(x)$ is bounded away from zero for x close to $\bar{x}$ and the initial approximation is sufficiently close to the root $\bar{x}$.

95.8 Fundamental Theorem of Calculus

The Fundamental Theorem of Calclulus: If $f : [a, b]$ is Lipschitz continuous, then there is a unique uniformly differentiable function $u : [a, b] \to \mathbb{R}$, that solves the initial value problem

$$\begin{cases} u'(x) = f(x) & \text{for } x \in (a, b], \\ u(a) = u_a, \end{cases}$$

where $u_a \in \mathbb{R}$ is given. The function $u : [a, b] \to \mathbb{R}$ can be expressed as

$$u(\bar{x}) = u_a + \int_a^{\bar{x}} f(x)\,dx \quad \text{for } \bar{x} \in [a, b],$$

where

$$\int_0^{\bar{x}} f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{j} f(x_{i-1}^n)h_n,$$

with $\bar{x} = x_j^n$, $x_i^n = a + ih_n$, $h_n = 2^{-n}(b-a)$. More precisely, if the Lipschitz constant of f is L_f then for $n = 1, 2, ...,$

$$\Big| \int_a^{\bar{x}} f(x)\,dx - \sum_{i=1}^{j} f(x_{i-1}^n)h_n \Big| \leq \frac{1}{2}(\bar{x} - a)L_f h_n.$$

Furthermore, if $|f(x)| \leq M_f$ for $x \in [a, b]$, then $u : [a, b] \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant M_f and $K_u \leq \frac{1}{2}L_f$, where K_u is the constant of uniform differentiability of u.

95.9 1D Integration Rules

Additivity:

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

Linearity: If α and β are real numbers then,

$$\int_a^b (\alpha f(x) + \beta g(x))\,dx = \alpha \int_a^b f(x)\,dx + \beta \int_a^b g(x)\,dx.$$

Monotonicity: If $f(x) \geq g(x)$ for $a \leq x \leq b$, then

$$\int_a^b f(x)\,dx \geq \int_a^b g(x)\,dx.$$

Differentiation and integration are inverse operations:

$$\frac{d}{dx}\int_a^x f(y)\,dy = f(x).$$

Change of variables: Setting $y = g(x)$, we have with formally $dy = g'(x)\,dx$,

$$\int_a^b f(g(x))g'(x)\,dx = \int_{g(a)}^{g(b)} f(y)\,dy.$$

Integration by parts:

$$\int_a^b u'(x)v(x)\,dx = u(b)v(b) - u(a)v(a) - \int_a^b u(x)v'(x)\,dx.$$

The Mean Value theorem: If $u(x)$ is uniformly differentiable on $[a, b]$ with Lipschitz continuous derivative $u'(x)$, then there is a (at least one) $\bar{x} \in [a, b]$, such that

$$u(b) - u(a) = u'(\bar{x})(b - a).$$

Taylor's theorem:

$$u(x) = u(\bar{x}) + u'(\bar{x})(x-\bar{x}) + \cdots + \frac{u^{(n)}(\bar{x})}{n!}(x-\bar{x})^n + \int_{\bar{x}}^{x} \frac{(x-y)^n}{n!}\, u^{(n+1)}(y)\, dy.$$

95.10 The Logarithm

Definition:

$$\log(x) = \int_1^x \frac{1}{y}\, dy \quad \text{for } x > 0.$$

Basic properties:

$$\begin{aligned} \frac{d}{dx}\log(x) &= \frac{1}{x} \quad \text{for } x > 0, \\ \log(ab) &= \log(a) + \log(b) \quad \text{for } a, b > 0, \\ \log(a^r) &= r\log(a), \quad \text{for } r \in \mathbb{R}, a > 0. \end{aligned}$$

95.11 The Exponential

Definition: $\exp(x) = e^x$ is the unique solution of the differential equation $u'(x) = u(x)$ for $x \in \mathbb{R}$ and $u(0) = 1$.

Basic properties:

$$\begin{aligned} \frac{d}{dx}\exp(x) &= \exp(x), \\ \exp(a+b) &= \exp(a)\exp(b) \quad \text{or } e^{a+b} = e^a e^b, \\ \exp(x) &= \lim_{j\to\infty} (1 + \frac{x}{j})^j. \end{aligned}$$

The inverse of the exponential is the logarithm:

$$y = \exp(x) \quad \text{if and only if } x = \log(y).$$

The function a^x with $a > 0$:

$$a^x = \exp(x\log(a)), \quad \frac{d}{dx}a^x = \log(a)a^x.$$

95.12 The Trigonometric Functions

Definition of $\sin(x)$ **and** $\cos(x)$: The initial value problem $u''(x)+u(x) = 0$ for $x > 0$ with $u_0 = 0$ and $u_1 = 1$, has a unique solution, which is denoted by $\sin(x)$. The initial value problem $u''(x)+u(x) = 0$ for $x > 0$ with $u_0 = 1$ and $u_1 = 0$, has a unique solution, which is denoted by $\cos(x)$. The functions $\sin(x)$ and $\cos(x)$ extend to $x < 0$ as solutions of $u''(x) + u(x) = 0$ and are periodic with period 2π, and $\sin(\pi) = 0$, $\cos(\frac{\pi}{2}) = 0$.

Properties:

$$\frac{d}{dx}\sin(x) = \cos(x),$$
$$\frac{d}{dx}\cos(x) = -\sin(x), \cos(-x) = \cos(x),$$
$$\sin(-x) = -\sin(x),$$
$$\cos(\pi - x) = -\cos(x),$$
$$\sin(\pi - x) = \sin(x),$$
$$\cos(x) = \sin(\frac{\pi}{2} - x),$$
$$\sin(x) = \cos(\frac{\pi}{2} - x),$$
$$\sin(\frac{\pi}{2} + x) = \cos(x),$$
$$\cos(\frac{\pi}{2} + x) = -\sin(x).$$

Definition of $\tan(x)$ **and** $\cot(x)$:

$$\tan(x) = \frac{\sin(x)}{\cos(x)}, \quad \cot(x) = \frac{\cos(x)}{\sin(x)}.$$

Derivatives of $\tan(x)$ **and** $\cot(x)$:

$$\frac{d}{dx}\tan(x) = \frac{1}{\cos^2(x)}, \quad \frac{d}{dx}\cot(x) = -\frac{1}{\sin^2(x)}.$$

Trigonometric formulas:

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

$$\cos(x) - \cos(y) = -2\sin(\frac{x+y}{2})\sin(\frac{x-y}{2}),$$
$$\tan(x+y) = \frac{\tan(x)+\tan(y)}{1-\tan(x)\tan(y)},$$
$$\tan(x-y) = \frac{\tan(x)-\tan(y)}{1+\tan(x)\tan(y)},$$
$$\sin(x) + \sin(y) = 2\sin(\frac{x+y}{2})\cos(\frac{x-y}{2}),$$
$$\sin(x) - \sin(y) = 2\cos(\frac{x+y}{2})\sin(\frac{x-y}{2}),$$
$$\cos(x) + \cos(y) = 2\cos(\frac{x+y}{2})\cos(\frac{x-y}{2}).$$

Inverses of trigonometric functions: The inverse of $f(x) = \sin(x)$ with $D(f) = [\frac{-\pi}{2}, \frac{\pi}{2}]$ is $f^{-1}(y) = \arcsin(y)$ with $D(\arcsin) = [-1, 1]$. The inverse of $f(x) = \tan(x)$ with $D(f) = (\frac{-\pi}{2}, \frac{\pi}{2})$ is $f^{-1}(y) = \arctan(y)$ with $D(\arctan) = \mathbb{R}$. The inverse of $y = f(x) = \cos(x)$ with $D(f) = [0, \pi]$ is $f^{-1}(y) = \arccos(y)$ with $D(\arccos) = [-1, 1]$. The inverse of $f(x) = \cot(x)$ with $D(f) = (0, \pi)$ is $f^{-1}(y) = \operatorname{arccot}(y)$ with $D(\operatorname{arccot}) = \mathbb{R}$. We have

$$\frac{d}{dy}\arcsin(y) = \frac{1}{\sqrt{1-y^2}}$$
$$\frac{d}{dy}\arctan(y) = \frac{1}{1+y^2}$$
$$\frac{d}{dy}\arccos(y) = -\frac{1}{\sqrt{1-y^2}}$$
$$\frac{d}{dy}\operatorname{arccot}(y) = -\frac{1}{1+y^2},$$
$$\arctan(u) + \arctan(v) = \arctan(\frac{u+v}{1-uv}).$$

Definition of sinh(x) and cosh(x):

$$\sinh(x) = \frac{e^x - e^{-x}}{2} \quad \text{and} \quad \cosh(x) = \frac{e^x + e^{-x}}{2} \quad \text{for} x \in \mathbb{R}.$$

Derivatives of of sinh(x) and cosh(x):

$$D\sinh(x) = \cosh(x) \quad \text{and} \quad D\cosh(x) = \sinh(x).$$

Inverses of of sinh(x) and cosh(x): the inverse of $y = f(x) = \sinh(x)$ with $D(f) = \mathbb{R}$ is $f^{-1}(y) = \operatorname{arcsinh}(y)$ with $D(\operatorname{arcsinh}) = \mathbb{R}$. The inverse of $y = f(x) = \cosh(x)$ with $D([0, \infty))$, is $f^{-1}(y) = \operatorname{arccosh}(y)$ with $D(\operatorname{arccosh}) = [1, \infty)$. We have

$$\frac{d}{dy}\operatorname{arcsinh}(y) = \frac{1}{\sqrt{y^2+1}}, \quad \frac{d}{dy}\operatorname{arccosh}(y) = \frac{1}{\sqrt{y^2-1}}.$$

95.13 List of Primitive Functions

$$\int_{x_0}^{x} \frac{1}{s-c}\,ds = \log|x-c| - \log|x_0-c|, \quad c \neq 0,$$

$$\int_{x_0}^{x} \frac{s-a}{(s-a)^2+b^2}\,dx = \frac{1}{2}\log((x-a)^2+b^2) - \frac{1}{2}\log((x_0-a)^2+b^2),$$

$$\int_{x_0}^{x} \frac{1}{(s-a)^2+b^2}\,ds = [\frac{1}{b}\arctan(\frac{x-a}{b})] - [\frac{1}{b}\arctan(\frac{x_0-a}{b})], \quad b \neq 0,$$

$$\int_{0}^{x} y\cos(y)\,dy = x\sin(x) + \cos(x) + 1,$$

$$\int_{0}^{x} \sin(\sqrt{y})\,dy = -2\sqrt{x}\cos(\sqrt{x}) + 2\sin(\sqrt{x}),$$

$$\int_{1}^{x} y^2\log(y)\,dy = \frac{x^3}{3}\log(x) - \frac{x^3}{9} + \frac{1}{9}$$

$$\int_{0}^{x} \frac{1}{\sqrt{1-y^2}}\,dy = \arcsin(x) \quad \text{for } x \in (-1,1)$$

$$\int_{0}^{x} \frac{1}{\sqrt{1-y^2}} = \frac{\pi}{2} - \arccos(x)\,dy \quad \text{for } x \in (-1,1)$$

$$\int_{0}^{x} \frac{1}{1+y^2}\,dy = \arctan(x) \quad \text{for } x \in \mathbb{R}$$

$$\int_{0}^{x} \frac{1}{1+y^2}\,dy = \frac{\pi}{2} - \operatorname{arccot}(x) \quad \text{for } x \in \mathbb{R}.$$

95.14 Series

Definition of convergence: A series $\sum_{i=1}^{\infty} a_i$ converges if and only if the sequence $\{s_n\}_{n=1}^{\infty}$ of partial sums $s_n = \sum_{i=1}^{n} a_i$ converges.

Geometric series: $\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}$ if $|a| < 1$.

Basic facts: A positive series $\sum_{i=1}^{\infty} a_i$ converges if and only if the sequence of partial sums is bounded above.

The series $\sum_{i=1}^{\infty} i^{-\alpha}$ converges if and only if $\alpha > 1$.

An absolutely convergent series is convergent.

An alternating series with the property that the modulus of its terms tends monotonically to zero, converges. Example: $\sum_{i=1}^{\infty}(-i)^{-1}$ converges.

95.15 The Differential Equation $\dot{u} + \lambda(x)u(x) = f(x)$

The solution to the initial-value problem $\dot{u} + \lambda(x)u(x) = f(x)$ for $x > 0$, $u(0) = u^0$, is given by

$$u(x) = \exp(-\Lambda(x))u^0 + \exp(-\Lambda(x)) \int_0^x \exp(\Lambda(y))f(y)\,dy,$$

where $\Lambda(x)$ is a primitive function of $\lambda(x)$ satisfying $\Lambda(0) = 0$.

95.16 Separable Scalar Initial Value Problems

The solution of the separable scalar initial value problem

$$u'(x) = \frac{h(x)}{g(u(x))} \quad \text{for } 0 < x \leq 1,\ u(0) = u_0,$$

where $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ are given functions, satisfies for $0 \leq x \leq 1$ the algebraic equation

$$G(u(x)) = H(x) + C,$$

where $G(v)$ and $H(x)$ are primitive functions of $g(v)$, and $C = G(u_0) - H(0)$.

96
MultiD Calculus

Timeo hominem unius libri. (St. Thomas of Aquino)

96.1 Introduction

We here collect the basic tools of Calculus of functions $f : \mathbb{R}^n \to \mathbb{R}^m$, that is Calculus of vector-valued functions of several real variables. The Euclidean norm of a vector $x = (x_1, ..., x_n) \in \mathbb{R}^n$ is denoted by $\|x\| = \sum_{i=1}^n x_i^2$.

96.2 Lipshitz Continuity

A function $f : A \to \mathbb{R}^m$ with a subset of $\mathbb{R}^n$ is Lipschitz continuous on A if there is a constant L such that

$$\|f(x) - f(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in A.$$

96.3 Differentiability

A function $f : A \to \mathbb{R}^m$ is *differentiable at* $\bar{x} \in A$, where A is an open subset of $\mathbb{R}^n$, if there is a $m \times n$ matrix $f'(\bar{x})$, called the *Jacobian* of the function $f(x)$ at $\bar{x}$, and a constant $K_f(\bar{x})$, such that for all $x \in A$ close to $\bar{x}$,

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}),$$

where $E_f(x,\bar{x})$ is an m-vector satisfying $\|E_f(x,\bar{x})\| \le K_f(\bar{x})\|x-\bar{x}\|^2$. We say that $f : A \to \mathbb{R}^m$ is *uniformly differentiable on* A if the constant $K_f(\bar{x}) = K_f$ can be chosen independently of $\bar{x} \in A$. We write $f' = \nabla f$ if $m = 1$ and call ∇f the gradient of f.

96.4 The Chain Rule

If $g : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\bar{x} \in \mathbb{R}^n$, and $f : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $g(\bar{x}) \in \mathbb{R}^m$ and further $g : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous, then the composite function $f \circ g : \mathbb{R}^n \to \mathbb{R}^p$ is differentiable at $\bar{x} \in \mathbb{R}^n$ with Jacobian

$$(f \circ g)'(\bar{x}) = f'(g(\bar{x}))g'(\bar{x}).$$

96.5 Mean Value Theorem for $f : \mathbb{R}^n \to \mathbb{R}$

If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable on $\mathbb{R}^n$ with a Lipschitz continuous gradient ∇f, then for given x and $\bar{x}$ in $\mathbb{R}^n$, there is $y = x + \bar{t}(x-\bar{x})$ with $\bar{t} \in [0,1]$, such that

$$f(x) - f(\bar{x}) = \nabla f(y) \cdot (x - \bar{x}).$$

96.6 A Minimum Point Is a Stationary Point

If $\bar{x} \in \mathbb{R}^n$ is a *local minimum point* of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, that is, $f(\bar{x}) \le f(x)$ for all x close to $\bar{x}$, then $\nabla f(\bar{x}) = 0$.

96.7 Taylor's Theorem

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable with Lipschitz continuous Hessian $H = (h_{ij})$ with elements $h_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$, then, for given x and $\bar{x} \in \mathbb{R}^n$, there is $y = x + \bar{t}(x-\bar{x})$ with $\bar{t} \in [0,1]$, such that

$$\begin{aligned} f(x) &= f(\bar{x}) + \nabla f(\bar{x}) \cdot (x-\bar{x}) + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(y)(x_i - \bar{x}_i)(x_j - \bar{x}_j) \\ &= f(\bar{x}) + \nabla f(\bar{x}) \cdot (x-\bar{x}) + \frac{1}{2}(x-\bar{x})^\top H(y)(x-\bar{x}). \end{aligned}$$

96.8 Contraction Mapping Theorem

If $g : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with Lipschitz constant $L < 1$, then the equation $x = g(x)$ has a unique solution $\bar{x} = \lim_{i\to\infty} x^{(i)}$, where $\{x^{(i)}\}_{i=1}^{\infty}$ is a sequence in $\mathbb{R}^n$ generated by Fixed Point Iteration: $x^{(i)} = g(x^{(i-1)})$, starting with any initial value $x^{(0)}$.

96.9 Inverse Function Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ and assume the coefficients of $f'(x)$ are Lipschitz continuous close to $\bar{x}$ and $f'(\bar{x})$ is non-singular. Then for y sufficiently close to $\bar{y} = f(\bar{x})$, the equation $f(x) = y$ has a unique solution x. This defines x as a function $x = f^{-1}(y)$ of y.

96.10 Implicit Function Theorem

If $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ with $f(x, y) \in \mathbb{R}^n$ and $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, $f(\bar{x}, \bar{y}) = 0$, and the Jacobian $f'_x(x, y)$ with respect to x is Lipschitz continuous for x close to $\bar{x}$ and y close to $\bar{y}$, and $f'_x(\bar{x}, \bar{y})$ is non-singular, then for y close to $\bar{y}$, the equation $f(x, y) = 0$ has a unique solution $x = g(y)$, which defines x as a function $g(y)$ of y.

96.11 Newton's Method

If $\bar{x}$ is a root of $f : \mathbb{R}^n \to \mathbb{R}^n$ such that $f(x)$ is uniformly differentiable with a Lipschitz continuous derivative close to $\bar{x}$ and $f'(\bar{x})$ is non-singular, then Newton's method $x^{(i+1)} = x^{(i)} - f'(x^{(i)})^{-1} f(x^{(i)})$ for solving $f(x) = 0$ converges quadratically if started sufficiently close to $\bar{x}$.

96.12 Differential Operators

Gradient of a function $u : \mathbb{R}^d \to \mathbb{R}$:

$$\text{grad } u = \nabla u = \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, ..., \frac{\partial u}{\partial x_d} \right).$$

Divergence of a vector function $u : \mathbb{R}^d \to \mathbb{R}^d$:

$$\text{div } u = \nabla \cdot u = \sum_{i=1}^{d} \frac{\partial u_i}{\partial x_i}.$$

Rotation of a vector function $u : \mathbb{R}^3 \to \mathbb{R}^3$:

$$\text{rot } u = \nabla \times u = \left(\frac{\partial u_3}{\partial x_2} - \frac{\partial u_2}{\partial x_3}, \frac{\partial u_1}{\partial x_3} - \frac{\partial u_3}{\partial x_1}, \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} \right).$$

Laplacian of a function $u : \mathbb{R}^d \to \mathbb{R}$:

$$\Delta u = \nabla \cdot (\nabla u) = \text{div (grad } u) = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}.$$

Identities:

$$\begin{aligned} \nabla \cdot (\nabla \times u) &= 0, \\ \nabla \times (\nabla u) &= 0, \\ \nabla \times (\nabla \times u) &= -\Delta u + \nabla(\nabla \cdot u). \end{aligned}$$

Laplacian in $\mathbb{R}^2$ in polar coordinates $x = (x_1, x_2) = (r\cos(\theta), r\sin(\theta))$:

$$\Delta u = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r} \right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}.$$

Laplacian in spherical coordinates
$x = (r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta), r\cos(\varphi))$:

$$\Delta u = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r} \right) + \frac{1}{r^2\sin(\theta)}\frac{\partial}{\partial \theta}\left(\sin(\theta)\frac{\partial u}{\partial \theta} \right) + \frac{1}{r^2\sin^2(\theta)}\frac{\partial^2 u}{\partial \varphi^2}.$$

The Laplacian is invariant under orthogonal coordinate transformations in $\mathbb{R}^d$.

96.13 Curve Integrals

If $\Gamma = s([a,b])$ is a curve in $\mathbb{R}^n$ given by the function $s : [a,b] \to \mathbb{R}^n$, and $u : \Gamma \to \mathbb{R}$, then

$$\begin{aligned} \int_\Gamma u\, ds &= \int_\Gamma u(x)\, ds(x) \equiv \int_a^b u(s(t))\|s'(t)\|\, dt, \\ \int_\Gamma u \cdot ds &= \int_a^b u(s(t)) \cdot s'(t)\, dt, \\ \int_\Gamma ds &= \int_a^b \|s'(t)\|\, dt = \text{length of } \Gamma. \end{aligned}$$

If $u = \nabla\varphi$, then

$$\int_\Gamma u \cdot ds = \varphi(s(b)) - \varphi(s(a)).$$

96.14 MultiD Integrals

Integral over the unit square: If $f : Q = [0,1] \times [0,1] \to \mathbb{R}$ is Lipschitz continuous, then

$$\int_Q f(x)\,dx = \int_0^1 \int_0^1 f(x_1,x_2)\,dx_1\,dx_2 = \lim_{n\to\infty} \sum_{i=1}^N \sum_{j=1}^N f(x_{1,i}^n, x_{2,j}^n) h_n h_n,$$

where $h_n = 2^{-n}$, $x_{j,i}^n = ih_n$, $N = 2^n$, and

$$\int_Q f(x)\,dx == \int_0^1 \big(\int_0^1 f(x_1,x_2)\,dx_2\big)\,dx_1 = \int_0^1 \big(\int_0^1 f(x_1,x_2)\,dx_1\big)\,dx_2.$$

Change of variables: If $y \to x = g(y)$ maps a domain $\tilde{\Omega}$ in $\mathbb{R}^d$ onto a domain Ω in $\mathbb{R}^d$, where the Jacobian of g is Lipschitz continuous and $f : \Omega \to \mathbb{R}$ be Lipschitz continuous, then

$$\int_\Omega f(x)\,dx = \int_{\tilde{\Omega}} f(g(y))|\det g'(y)|\,dy,$$

Polar coordinates:

$$\int_\Omega f(x_1,x_2)\,dx_1 dx_2 = \int_{\tilde{\Omega}} f(r\cos(\theta, r\sin(\theta))\,r dr\,d\theta,$$

where $(r,\theta) \to x$ is a one-to-one mapping of $\tilde{\Omega}$ onto Ω given by $x = (r\cos(\theta,\sin(\theta))$.
Spherical coordinates:

$$\int_\Omega f(x)\,dx$$
$$= \int_{\tilde{\Omega}} f\big(r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta), r\cos(\varphi)\big) r^2\sin(\varphi)\,dr\,d\theta\,d\varphi,$$

where $(r,\theta,\varphi) \to x$ is a one-to-one mapping of $\tilde{\Omega}$ onto Ω given by $x = (r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta), r\cos(\varphi))$.

96.15 Surface Integrals

If $S = s(\Omega)$ is a surface in $\mathbb{R}^3$ parameterized by the mapping $s : \Omega \to \mathbb{R}^3$, where Ω is a domain in $\mathbb{R}^2$, and $u : S \to \mathbb{R}$ is a real-valued function defined on S, then

$$\int_S u\,ds = \int_\Omega u(s(y))\|s'_{,1}(y) \times s'_{,2}(y)\|\,dy,$$

where $s'_{,i} = (\frac{\partial s_1}{\partial y_i}, \frac{\partial s_2}{\partial y_i}, \frac{\partial s_3}{\partial y_i})$.

96.16 Green's and Gauss' Formulas

If Ω is a domain in $\mathbb{R}^3$ with boundary Γ with outward unit normal $n = (n_1, n_2, n_3)$, and $u : \Omega \to \mathbb{R}^3$ and $v, w : \Omega \to \mathbb{R}$, then

$$\int_\Omega \frac{\partial v}{\partial x_i}\, dx = \int_\Gamma v\, n_i\, ds, \quad i = 1, 2, 3.$$

$$\int_\Omega \frac{\partial v}{\partial x_i} w\, dx = \int_\Gamma vw\, n_i\, ds - \int_\Omega v \frac{\partial w}{\partial x_i}\, dx, \quad i = 1, 2, 3.$$

$$\int_\Omega \nabla \cdot u\, dx = \int_\Gamma u \cdot n\, ds, \quad \text{(Gauss' Divergence theorem)}$$

$$\int_\Omega \nabla \times u\, dx = \int_\Gamma n \times u\, ds,$$

$$\int_\Omega \nabla v \cdot \nabla w\, dx = \int_\Gamma v \partial_n w\, ds - \int_\Omega v \Delta w\, dx,$$

$$\int_\Omega v \Delta w\, dx - \int_\Omega \Delta v\, w\, dx = \int_\Gamma v\, \partial_n w\, ds - \int_\Gamma \partial_n v\, w\, ds.$$

96.17 Stokes' Theorem

If S is a surface in $\mathbb{R}^3$ bounded by a closed curve Γ, n is a unit normal to S, Γ is oriented in a clockwise direction following the positive direction of the normal n, and $u : \mathbb{R}^3 \to \mathbb{R}^3$ is differentiable, then

$$\int_S (\nabla \times u) \cdot n\, ds = \int_\Gamma u \cdot ds.$$

97
Linear Algebra

97.1 Linear Algebra in $\mathbb{R}^2$

Scalar product of two vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^2$:

$$a \cdot b = (a, b) = a_1 b_1 + a_2 b_2.$$

Norm: $|a| = (a_1^2 + a_2^2)^{1/2}$.

Angle between two vectors a and b in $\mathbb{R}^2$: $\cos(\theta) = \frac{a \cdot b}{|a|\,|b|}$.
The vectors a and b are orthogonal if and only if $a \cdot b = 0$.

Vector product of two vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^3$:

$$a \times b = a_1 b_2 - a_2 b_1.$$

Properties of vector product: $|a \times b| = |a||b||\sin(\theta)|$, where θ is the angle between a and b. In particular, a and b are parallel if and only if $a \times b = 0$.

Volume of parallelogram spanned by two vectors $a, b \in \mathbb{R}^2$:

$$V(a, b) = |a \times b| = |a_1 b_2 - a_2 b_1|.$$

97.2 Linear Algebra in $\mathbb{R}^3$

Scalar product of two vectors $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ in $\mathbb{R}^3$:

$$a \cdot b = \sum_{i=1}^{3} a_i b_i = a_1 b_1 + a_2 b_2 + a_3 b_3.$$

Norm: $|a| = (a_1^2 + a_2^2 + a_3^2)^{1/2}$.

Angle between two vectors a and b in $\mathbb{R}^3$: $\cos(\theta) = \frac{a \cdot b}{|a|\,|b|}$.
The vectors a and b are orthogonal if and only if $a \cdot b = 0$.
Vector product of two vectors $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ in $\mathbb{R}^3$:

$$a \times b = (a_2 b_3 - a_3 b_2, a_3 b_1 - a_1 b_3, a_1 b_2 - a_2 b_1).$$

Properties of vector product: The vector product $a \times b$ of two non-zero vectors a and b in $\mathbb{R}^3$, is orthogonal to both a and b, and $|a \times b| = |a||b|| \sin(\theta)|$, where θ is the angle between a and b. In particular, a and b are parallel if and only if $a \times b = 0$.

Volume of parallelepiped spanned by three vectors $a, b, c \in \mathbb{R}^3$:

$$V(a, b, c) = |c \cdot (a \times b)|.$$

97.3 Linear Algebra in $\mathbb{R}^n$

Definition of $\mathbb{R}^n$: The set of ordered n-tuples, $x = (x_1, ..., x_n)$ with components $x_i \in \mathbb{R}$, $i = 1, ..., n$.

Vector addition and scalar multiplication: For $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ in $\mathbb{R}^n$ and $\lambda \in \mathbb{R}$, we define

$$x + y = (x_1 + y_1, x_2 + y_2, ..., x_n + y_n), \quad \lambda x = (\lambda x_1, ..., \lambda x_n).$$

Scalar product: $x \cdot y = (x, y) = \sum_{i=1}^{n} x_i y_i$. **Norm**: $|x| = (\sum_{i=1}^{n} x_i^2)^{1/2}$.

Cauchy's inequality: $|(x, y)| \leq |x|\,|y|$.

Angle of two vectors x and y in $\mathbb{R}^n$: $\cos(\theta) = \frac{(x,y)}{|x||y|}$.

Standard basis: $\{e_1, ..., e_n\}$, where $e_i = (0, 0, .., 0, 1, 0, ..., 0)$ with a single coefficient 1 at position i.

Linear independence: A set $\{a_1, ..., a_n\}$ of vectors in $\mathbb{R}^m$ is said to be *linearly independent* if none of the vectors a_i can be expressed as a linear combination of the others, that is, if $\sum_{i=1}^{n} \lambda_i a_i = 0$ with $\lambda_i \in \mathbb{R}$ implies that $\lambda_i = 0$ for $i = 1, ..., n$.

A basis for $\mathbb{R}^n$ is a linearly independent set of vectors whose linear combinations span $\mathbb{R}^n$. Any basis of $\mathbb{R}^n$ has n elements. Further, a set of n vectors in $\mathbb{R}^n$ span $\mathbb{R}^n$ if and only if it is linearly independent, that is, a set of n vectors in $\mathbb{R}^n$ that spans $\mathbb{R}^n$ or is independent, must be a basis. Also, a set of fewer than n vectors in $\mathbb{R}^n$ cannot span $\mathbb{R}^n$, and a set of more than n vectors in $\mathbb{R}^n$ must be linearly dependent.

97.4 Linear Transformations and Matrices

An $m \times n$ real (or complex) *matrix* $A = (a_{ij})$ is rectangular array with rows $(a_{i1}, ..., a_{in})$, $i = 1, ..., m$, and columns $(a_{1j}, ..., a_{mj})$, $j = 1, ..., n$, where $a_{ij} \in \mathbb{R}$ (or $a_{ij} \in \mathbb{C}$).

Matrix addition: Given two $m \times n$ matrices $A = (a_{ij})$ and $B = (b_{ij})$, we define $C = A+B$ as the $m \times n$ matrix $C = (c_{ij})$ with elements $c_{ij} = a_{ij}+b_{ij}$, corresponding to elementwise addition.

Multiplication by scalar Given a $m \times n$ matrix $A = (a_{ij})$ and a real number λ, we define the $m \times n$ matrix λA with elements (λa_{ij}), corresponding to multiplying all elements of A by the real number λ.

Matrix multiplication: Given a $m \times p$ matrix A and a $p \times n$ matrix B we define a $m \times n$ matrix AB with elements $(AB)_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj}$. Matrix multiplication is not commutative, that is, $AB \neq BA$ in general. In particular, BA is defined only if $n = m$.

A linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ can be expressed as $f(x) = Ax$, where $A = (a_{ij})$ is an $m \times n$ matrix with elements $a_{ij} = f_i(e_j) = (e_i, f(e_j))$, where $f(x) = (f_1(x), ..., f_m(x))$. If $g : \mathbb{R}^n \to \mathbb{R}^p$ and $f : \mathbb{R}^p \to \mathbb{R}^m$ are two linear transformations with corresponding matrices A and B, then the matrix of $f \circ g : \mathbb{R}^n \to \mathbb{R}^m$ is given by AB.

Transpose: If $A = (a_{ij})$ is a real $m \times n$ matrix, then the transpose $A^\top$ is an $n \times m$ matrix with elements $a_{ji}^\top = a_{ij}$, and $(Ax, y) = (x, A^\top y)$ for all $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$.

Matrix norms:

$$\|A\|_1 = \max_{j=1,...,n} \sum_{i=1}^{m} |a_{ij}|, \quad \|A\|_\infty = \max_{i+1,...,m} \sum_{j=1}^{n} |a_{ij}|, \quad \|A\| = \max_{x \in \mathbb{R}^n} \frac{\|Ax\|}{\|x\|}.$$

If $A = (\lambda_i)$ is a diagonal $n \times n$ matrix with diagonal elements $a_{ii} = \lambda_i$, then

$$\|A\| = \max_{i=1,...,n} |\lambda_i|.$$

Lipschitz constant of a linear transformation: The Lipschitz constant of a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ given by a $m \times n$ matrix $A = (a_{ij})$ is equal to $\|A\|$.

97.5 The Determinant and Volume

The **determinant** $\det A$ of an $n \times n$ matrix $A = (a_{ij})$, or the volume $V(a_1, ..., a_n)$ spanned by the column vectors of A, is defined by

$$\det A = V(a_1, ..., a_n) = \sum_{\pi} \pm a_{\pi(1)\,1} a_{\pi(2)\,2} \cdots a_{\pi(n)\,n},$$

where we sum over all permutations π of the set $\{1, ..., n\}$, and the sign indicates if the permutation is even (+) or odd (-). We have $\det A = \det A^\top$.

Volume $V(a_1, a_2)$ in $\mathbb{R}^2$:

$$\det A = V(a_1, a_2) = a_{11}a_{22} - a_{21}a_{12}.$$

Volume $V(a_1, a_2, a_3)$ in $\mathbb{R}^3$:

$$\begin{aligned}&\det A = V(a_1, a_2, a_3) = a_1 \cdot a_2 \times a_3\\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}).\end{aligned}$$

Volume $V(a_1, a_2, a_3, a_4)$ in $\mathbb{R}^4$:

$$\begin{aligned}\det A = V(a_1, a_2, a_3, a_4) &= a_{11}V(\hat{a}_2, \hat{a}_3, \hat{a}_4) - a_{12}V(\hat{a}_1, \hat{a}_3, \hat{a}_4)\\ &+a_{13}V(\hat{a}_1, \hat{a}_2, \hat{a}_4) - a_{14}V(\hat{a}_1, \hat{a}_2, \hat{a}_3),\end{aligned}$$

where the $\hat{a}_j$, $j = 1, 2, 3, 4$ are the 3-column vectors corresponding to cutting out the first coefficient of the a_j.

Determinant of a triangular matrix: If $A = (a_{ij})$ is a *upper triangular* $n \times n$ matrix, that is $a_{ij} = 0$ for $i > j$, then

$$\det A = a_{11}a_{22} \cdots a_{nn}.$$

This formula also applies to a *lower triangular* $n \times n$ matrix $A = (a_{ij})$ with $a_{ij} = 0$ for $i < j$.

The magic formula: $\det AB = \det A \det B$.

Test of linear independence: A set $\{a_1, a_2, ..., a_n\}$ of n vectors in $\mathbb{R}^n$ is linearly independent if and only if $V(a_1, ..., a_n) \neq 0$. The following statements are equivalent for an $n \times n$ matrix A: (a) The columns of A are linearly independent, (b) If $Ax = 0$, then $x = 0$, (c) $\det A \neq 0$.

97.6 Cramer's Formula

If A is a $n \times n$ non-singular matrix with $\det A \neq 0$, then the system of equations $Ax = b$ has a unique solution $x = (x_1, ..., x_n)$ for any $b \in \mathbb{R}^n$. given by

$$x_i = \frac{V(a1, ..., a_{i-1}, b, a_{i+1}, ..., a_n)}{V(a_1, a_2, ..., a_n)}, \quad i = 1, ..., n.$$

97.7 Inverse

A nonsingular $n \times n$ matrix A has a inverse matrix A^{-1} satisfying:

$$A^{-1}A = AA^{-1} = I,$$

where I is the $n \times n$ identity matrix.

97.8 Projections

The projection $Pv \in V$ of $v \in \mathbb{R}^n$, where V is a linear subspace of $\mathbb{R}^n$, is uniquely defined defined by $(v - Pv, w) = 0$ for all $w \in V$ and satisfies $|v - Pv| \leq |v - w|$ for all $w \in V$. Further, $PP = P$ and $P^\top = P$.

97.9 The Fundamental Theorem of Linear Algebra

If A is a $m \times n$ matrix with null space $N(A) = \{x \in \mathbb{R}^n : Ax = 0\}$ and range $R(A) = \{y = Ax : x \in \mathbb{R}^n\}$, then

$$N(A) \oplus R(A^\top) = \mathbb{R}^n \quad N(A^\top) \oplus R(A) = \mathbb{R}^m,$$

$$\text{dim } N(A) + \text{dim } R(A^\top) = n, \text{ dim } N(A^\top) + \text{dim } R(A) = m,$$

$$\text{dim } N(A) + \text{dim } R(A) = n, \text{ dim } N(A^\top) + \text{dim } R(A^\top) = m,$$

$$\text{dim } R(A) = \text{dim } R(A^\top),$$

The number of linearly independent columns of A is equal to the number of linearly independent rows of A.

97.10 The QR-Decomposition

An $n \times m$ matrix A can be expressed in the form

$$A = QR,$$

where Q is a $n \times m$ matrix with orthogonal columns and R is a $m \times m$ upper triangular matrix.

97.11 Change of Basis

A linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$, with matrix A with respect to the standard basis, has the following matrix in a basis $\{s_1, ...s_n\}$:

$$S^{-1}AS,$$

where the the coefficients s_{ij} of the matrix $S = (s_{ij})$ are the coordinates of the basis vectors s_j with respect to the standard basis.

97.12 The Least Squares Method

The least squares solution of the linear system $Ax = b$ with A an $m \times n$ matrix minimizing $|Ax - b|^2$ satisfies $A^\top Ax = A^\top b$, and is unique if the columns of A are linearly independent.

97.13 Eigenvalues and Eigenvectors

If A is an $n \times n$ matrix and $x =\in \mathbb{R}^n$ is a non-zero vector which satisfies $Ax = \lambda x$, where λ is a real number, then we say that $x \in \mathbb{R}^n$ is an *eigenvector* of A and that λ is a corresponding *eigenvalue* of A. The number λ is an eigenvalue of the $n \times n$ matrix A if and only if λ is a root of the characteristic equation $\det(A - \lambda I) = 0$.

97.14 The Spectral Theorem

If A is a symmetric $n \times n$ matrix A, then there is an orthonormal basis $\{q_1, ..., q_n\}$ of $\mathbb{R}^n$ consisting of eigenvectors q_j of A with corresponding real eigenvalues λ_j, satisfying $Aq_j = \lambda_j q_j$, for $j = 1, ..., n$. We have $D = Q^{-1}AQ$ and $A = QDQ^{-1}$, where Q is the orthogonal matrix with the eigenvectors q_j in the standard basis forming the columns, and D is the diagonal matrix with the eigenvalues λ_j on the diagonal. Further, $\|A\| = \max_{i=1,...,n} |\lambda_i|$.

97.15 The Conjugate Gradient Method for $Ax = b$

For $k = 0, 1, 2, ...$, with $r^k = Ax^k - b$, $x^0 = 0$ and $d^0 = b$, do

$$x^{k+1} = x^k + \alpha_k d^k, \quad \alpha_k = -\frac{(r^k, d^k)}{(d^k, Ad^k)},$$

$$d^{k+1} = -r^{k+1} + \beta_k d^k, \quad \beta_k = \frac{(r^{k+1}, Ad^k)}{(d^k, Ad^k)}.$$

Part VII

Sessions

98
Overview

> It was when I found out I could make mistakes that I knew I was on to something. (Ornette Coleman)
>
> Jazz is a mental attitude rather than a style. It uses a certain process of the mind expressed spontaneously through some musical instrument. I'm concerned with retaining that process. (Bill Evans)
>
> Jam Session = An informal gathering of musicians to play improvised or unrehearsed music. (Online Jam Session)

The BodyandSoul Sessions help you to master basic tools of Calculus and Linear Algebra in interaction with a computer, including the following topics presented in Leibniz World of Mathematics:

1. functions,
2. Lipschitz continuity,
3. derivatives,
4. Fundamental Theorems of Calculus and Linear Algebra,
5. elementary functions
6. geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$,
7. fixed point iteration and Newton's method,

8. time stepping and adaptive error control
9. gradient, divergence and Laplacian,
10. piecewise linear interpolation.
11. finite element programming,

Each Session contains simple Python codes which you can use as templates to get a quick start.

A set of Sessions using Matlab is available as Sessions A-F.

A new set based on Python is presented in the following chapters.

Further below you will find additional Sessions related to the material of

- Part VII World of Differential Equations.
- Part VIII World of Finite Elements.

The material covered by the Sessions is summarized in

- 1D Calculus
- MultiD Calculus
- Geometry and Linear Algebra.

98.1 Python Code

The Sessions includes writing Python code for

- Functions.
- Computing derivatives, analytically and computationally.
- Time stepping of $\dot{u} = f(t)$ and $\dot{u} = f(u)$. Quadrature in 1d.
- Computing elementary functions.
- Computing maximum of a Lipschitz continuous function.
- Fixed point iteration for $x = g(x)$.
- Newton's method for $f(x) = 0$.
- Transformation of matrix to row and column echelon form.
- Gaussian elimination.
- Jacobi iteration for linear system $Ax = b$ with A diagonally dominant.

- Conjugate gradient method for linear system Ax with A positive definite.
- Least squares method for linear system $Ax = b$.
- Geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$: scalar and vector product, projection, reflection, rotation,..
- Level curves of $f : \mathbb{R}^2 \to \mathbb{R}$.
- Level surfaces of $f : \mathbb{R}^3 \to \mathbb{R}$.
- Graphics: curves, surfaces and volumes.
- Quadrature in 2d and 3d.

99
Functions

> All things are subject to interpretation whichever interpretation prevails at a given time is a function of power and not truth. (Friedrich Nietzsche)
>
> Although all the good arts serve to draw man's mind away from vices and lead it toward better things, this function can be more fully performed by this art, which also provides extraordinary intellectual pleasure. (Nicolaus Copernicus)

99.1 To Read

- Functions
- What Is a Function?

99.2 To Do

Do the following:

- Plot some polynomial and rational functions $f : \mathbb{Q} \to \mathbb{Q}$.
- Construct different functions in Python code.

FIGURE 99.1. Educational road bridge in Beijing: Do you get the message? ($F = m_1 m_2/r^2$, $E = MC^2$, $f(b) - f(a) = f'(\xi)(b-a)$)

FIGURE 99.2. (Jam) Session with Jan Johansson (piano) and Stan Getz (sax)

- Contemplate how functions are specified in Python. Does it tell us what a function is?
- Is $t \to f(t) = \sin(t)$ a function? How is $f(t)$ determined for a given value of t? By table or computation? How is $\sin(\sqrt{2})$ specified?
- Reflect over symbolic vs constructive digital specification of functions.

100
Derivatives and Lipschitz Continuity

It is incontestable and deplorable that Negroes have committed crimes; but they are derivative crimes. They are born of the greater crimes of the white society. (Martin Luther King, Jr)

The power of kings and magistrates is nothing else but what is only derivative; transformed and committed to them in trust from the people to the common good of them all, in whom the power yet remains fundamentally, and cannot be taken from them without a violation of their natural birthright. (John Milton)

Continuous eloquence wearies. Grandeur must be abandoned to be appreciated. Continuity in everything is unpleasant. Cold is agreeable, that we may get warm. (Blaise Pascal)

100.1 To Read

- Lipschitz Continuity
- Definition of Derivative

FIGURE 100.1. The concept of derivative.

100.2 To Do

Recall that the derivative $\dot{u} : \mathbb{R} \to \mathbb{R}$ of a function $u : \mathbb{R} \to \mathbb{R}$ is defined by the relation

$$|u(t+k) - u(t) - \dot{u}(t)k| \leq Ck^2 \quad \text{for } k > 0. \tag{100.1}$$

where C is a constant.

- Compute $\dot{u}(t)$ using (**??**) for different $u(t)$ and k.
- Compare with analytical formulas for $u(t)$ polynomial.
- Study what time step k to best use with respect to the precision of function values $u(t)$. (Hint)
- Study the relation between Lipschitz constant and derivative. (Hint)
- Give examples of functions which are (i) Lipschitz continuous but not differentiable, (ii) not Lipschitz continuous.
- Reflect over how a Lipschitz continuous function $f : \mathbb{Q} \to \mathbb{Q}$ can be thought to be extended to $f : \mathbb{R} \to \mathbb{R}$.
- Which functions $u(t)$ satisfy $|u(t+k) - u(t)| \leq Ck^\theta$ for all $k > 0$, where $\theta > 1$ a constant?
- Find a function $u(t)$ satisfying $|u(t+k) - u(t)| \leq Ck^\theta$ for all $k > 0$, where $0 < \theta < 1$ is a constant.
- Compare with Weierstrass functions. (optional)

101
FundThm Calculus: $\dot{u}(t) = f(t)$

> Discipline in art is a fundamental struggle to understand oneself, as much as to understand what one is drawing. (Henry Moore)
>
> Fundamental progress has to do with the reinterpretation of basic ideas. (Alfred North Whitehead)

101.1 To Read

- Fundamental Theorem of Calculus
- Proof of Fundamental Theorem of Calculus
- The Integral (optional)

101.2 To Do

Consider the IVP

$$\dot{u}(t) = f(t) \quad \text{for } t > 0, \quad u(0) = u^0, \tag{101.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a given Lipschitz continuous function with Lip constant L, and u^0 a given initial value. Compute $u(t)$ by time stepping:

$$u((n+1)k) = u(nk) + kf(nk) \quad \text{for } n = 0, 1, ..., \text{ with } u(0) = u^0, \tag{101.2}$$

FIGURE 101.1. What is (so) Fundamental?

with some given time step k.

Do the following:

- Write Python code implementing (101.2).
- Compute with $f(t)$ polynomial and compare with analytical formulas.
- Compute the difference $e_k(t) = u(t) - \bar{u}(t)$ between $u(t)$ computed with time step k and $\bar{u}(t)$ computed with time step $\frac{k}{2}$.
- Study computationally the dependence of $e_k(t)$ on k, t and L.
- Estimate $e_k(t)$ analytically in terms of k, t and L.
- Compare analytical theory and computational experience.
- Show that convergence follows if $|e_k(t)| \leq Ck$ for some constant C, by comparing with Achilles and the Tortoise computing with time steps $k, \frac{k}{2}, \frac{k}{4}, ...$
- Interprete time-stepping as quadrature of integral as area. (Hint)
- Compare using Midpoint Euler, and motivate why this method also is called the trapezoidal method.
- Compare the accuracy of Forward and Midpoint Euler.
- Argue that the error in Midpoint Euler should be proportional to k^2.

102
FundThm Calculus: $\dot{u} = f(u)$

102.1 To Read

- General Fundamental Theorem of Calculus

102.2 To Do

Consider the IVP

$$\dot{u}(t) = f(u(t)) \quad \text{for } t > 0, \quad u(0) = u^0, \tag{102.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a given Lipschitz continuous function with Lip constant L, and u^0 a given initial value. Compute $u(t)$ by time stepping:

$$u((n+1)k) = u(nk)+kf(u(nk)) \quad \text{for } n = 0, 1, ..., \text{ with } u(0) = u^0, \tag{102.2}$$

with some given time step k.

- Write Python code implementing (102.2).
- Compute the difference $e_k(t) = u(t) - \bar{u}(t)$ between $u(t)$ computed with time step k and $\bar{u}(t)$ computed with time step $\frac{k}{2}$.
- Study computationally the dependence of $e_k(t)$ on k, t and L.
- Estimate $e_k(t)$ analytically in terms of k, t and L (e.g. in the case $f(u) = u$).

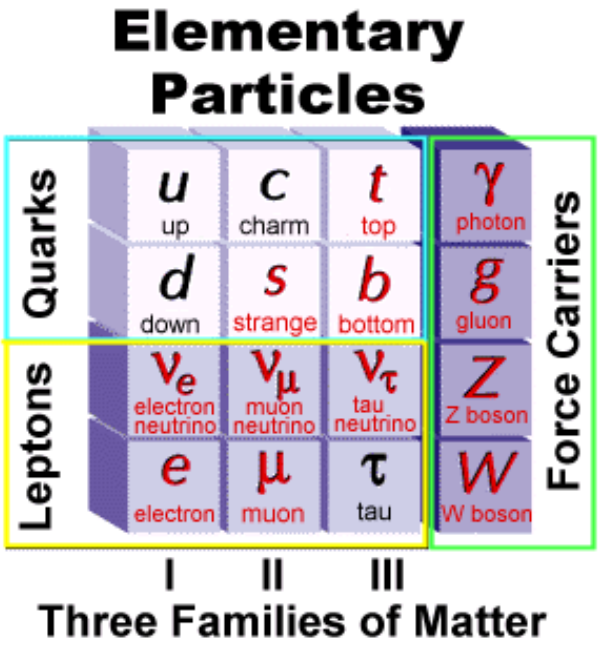


FIGURE 102.1. Fundamental physics?.

- Compare theory and experience.
- Compare with BodySoul Session E.
- Compare the accuracy of Forward and Midpoint Euler in terms of powers of k.

103
Fundamental Theorem Games

103.1 To Do

Design computer games based on the Fundamental Theorem e.g. along the following lines (see also the next Session):

1. Game of Keeping Even Pace: Let $U : [0,1] \to \mathbb{R}$ be a given function with $U(0) = 0$ and seek to control the input $v(t)$ of the IVP $\dot{u}(t) = v(t)$ so that

$$|U(t) - u(t)| \leq \delta \quad \text{for } 0 < t < 1, \tag{103.1}$$

where $\delta > 0$ is a given tolerance. This is like keeping even pace with somebody by controling your velocity.

2. Game of Fastest Integrator 1: Given $v : [0,T] \to \mathbb{R}$ compute as fast as possible, up to a certain tolerance, the solution of the IVP: $\dot{u}(t) = v(t)$ for $0 < t < 1$ and $u(0) = 0$.

2. Game of Fastest Integrator 2: Given $f : [0,T] \to \mathbb{R}$ and u^0 compute as fast as possible, up to a given tolerance, the solution of the IVP: $\dot{u}(t) = f(u(t))$ for $0 < t \leq T$ and $u(0) = u^0$.

104
Elementary Functions

> From time immemorial, man has desired to comprehend the complexity of nature in terms of as few elementary concepts as possible. (Abdus Salam)
>
> What I try to do in the book is to trace the chain of relationships running from elementary particles, fundamental building blocks of matter everywhere in the universe, such as quarks, all the way to complex entities, and in particular complex adaptive system like jaguars. (Murray Gell-Mann)
>
> Elementary, my dear Watson.

104.1 To Read

- exponential function
- trigonometric functions
- exponential complex
- logarithm

104.2 To Do

The elementary functions are solutions of elementary IVPs such as

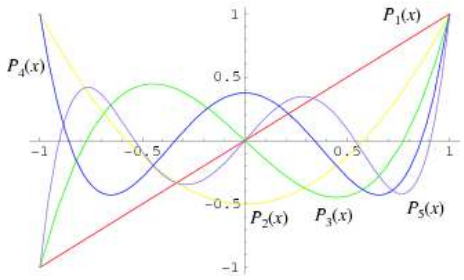


FIGURE 104.1. Legendre polynomials $P_1(x)$,...,$P_4(x)$.

- $\dot{u} = u, \quad (u(t) = \exp(t))$,
- $\dot{u} = v$, $\dot{v} = -u, \quad (u(t) = \sin(t)$ or $u(t) = \cos(t))$,
- $\dot{u}(t) = \frac{1}{t}, \quad (u(t) = \log(t))$.

Do the following:

- Compute elementary functions by time stepping.
- Derive the basic properties of $\exp(t)$, $\log(t)$, $\sin(t)$ and $\cos(t)$ from their defining IVPs:
- $\exp(a+b) = \exp(a)\exp(b)$, $(\exp(a))^r = \exp(ra)$,
- $\log(ab) = \log(a) + \log(b)$, $\log(a^r) = r\log(a)$,
- $\frac{d}{dt}\sin(t) = \cos(t)$, $\frac{d}{dt}\cos(t) = -\sin(t)$,
- $\sin(t+\pi) = -\sin(t)$, $\sin(t+\frac{\pi}{2}) = \cos(t)$,...
- Define $\exp(it) = \cos(t) + i\sin(t)$ where $i^2 = -1$.
- Prove that $\frac{d}{dt}\exp(it) = i\exp(it)$.
- Prove formulas like $\sin(2t) = 2\sin(t)\cos(t)$ (using that $\exp(i2t) = \exp(it)\exp(it)$).
- Study in the same spirit some other elementary functions, like Bessel functions (optional).
- Solve some Separable IVPs computationally and compare with analytical solutions. Consider e.g.: $\dot{u} = -u^2$, $\dot{u} = u^2$, $\dot{u} = u(1-u)$.
- Write Python code for analytical differentiation of combinations of elementary functions.

105
Geometry in $\mathbb{R}^2$

> All one's inventions are true, you can be sure of that. Poetry is as exact a science as geometry. (Gustave Flaubert)
>
> Geometry is not true, it is advantageous. (Henri Poincare)
>
> I am coming more and more to the conviction that the necessity of our geometry cannot be demonstrated, at least neither by, nor for, the human intellect. (Carl Friedrich Gauss)

105.1 To Read

- Geometry in $\mathbb{R}^2$

105.2 To Do

Write Python code for

- length or norm of a vector,
- distance between two points,
- scalar product of two vectors,
- $\times$-product of two vectors

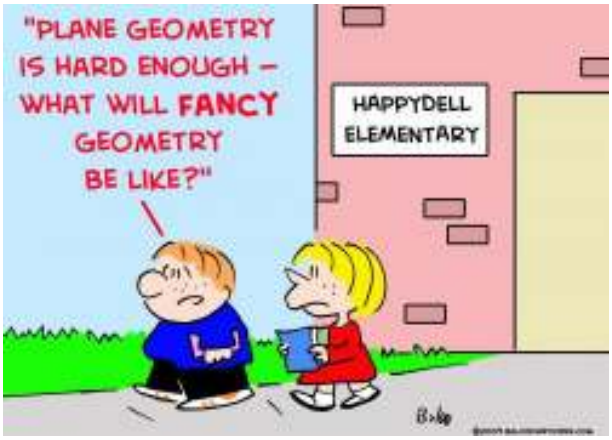


FIGURE 105.1. Math Fear.

- projection of one vector on another,
- orthogonalizing two non-colinear vectors,
- angle between two vectors,
- area of a romb spanned by two vectors
- rigid rotation and translation.

106
Geometry in $\mathbb{R}^3$

> The early study of Euclid made me a hater of geometry. (James Joseph Sylvester)

106.1 To Read

- Geometry in $\mathbb{R}^3$

106.2 To Do

Write Python code for

- length or norm of a vector,
- distance between two points,
- scalar product of two vectors,
- vector product of two vectors,
- triple-product of three vectors,
- projection of one vector on another,
- projection of a vector on a plane,

FIGURE 106.1. Geometry of Ales Stenar.

- orthogonalizing three non-coplanar vectors,
- angle between two vectors,
- volume of a parallelepiped spanned by three vectors,
- rigid rotation and translation.

107
FundThm of Linear Algebra

Algebra is generous; she often gives more than is asked of her. (D'Alembert)

I do not believe there is anything useful which men can know with exactitude that they cannot know by arithmetic and algebra. (Nicholas Malebranche)

107.1 To Read

- Geometry in $\mathbb{R}^n$
- vector, linear combination, scalar product, Cauchy's inequality,
- matrix, transpose, matrix product, linear independence, basis, volume,...
- orthogonal matrix, inverse matrix.

107.2 To Do: Fundamental Theorem

- Write Python codes for determining the column and row echelon forms of a matrix A.

FIGURE 107.1. Gilbert Strang with a generic $-1\ 2\ -1$ diagonal Cake Matrix.

- Use the code to determine the range $R(A)$ and null space $N(A)$ of different matrices A.
- Check the Fundamental Theorem of Linear Algebra computationally for different matrices.
- Check the Proof of the Fundamental Theorem of Linear Algebra.

107.3 To Do: Gaussion Elimination

- Write Python code for Gaussian elimination.

107.4 Watch

- Gilbert Strang on Linear Algebra

108
Contraction Mapping

108.1 To Read

- Fixed Point Iteration

108.2 To Do

Consider an equation of the form

$$x = g(x) \tag{108.1}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a Lipschitz continuous function with Lip constant L. A number x satisfying this equation is said to a fixed point of g. Consider the iteration

$$x(n+1) = g(x(n)), \quad , n = 0, 1, 2, ... \tag{108.2}$$

with $x(0)$ given.

Do the following

- Study computationally the convergence of the sequence $x(0), x(1), x(2), ..$ with different functions g with different L.
- Experience convergence if $L < 1$.
- Study the dependence of $|x(n+1) - x(n)|$ on $L < 1$.

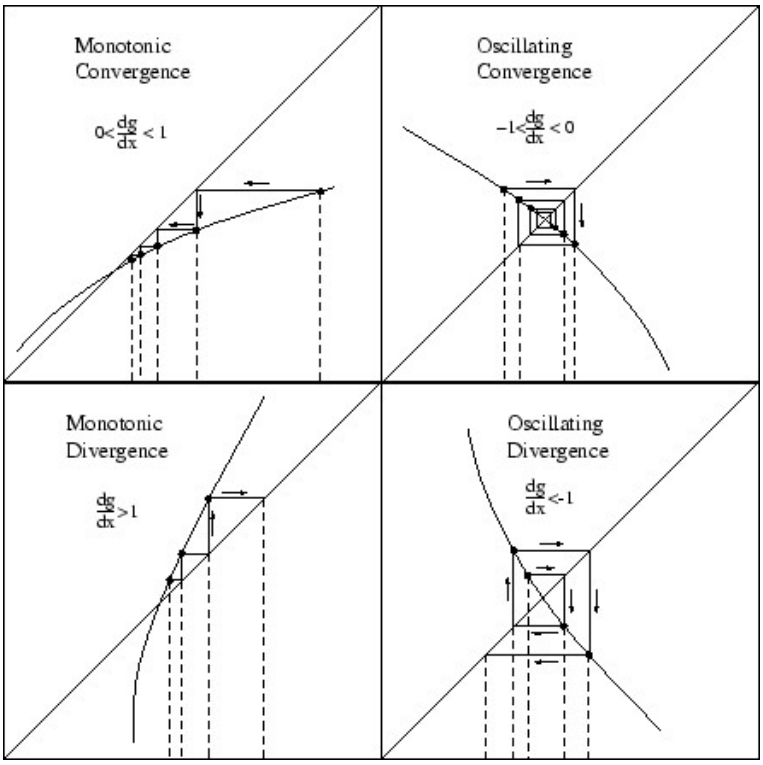


FIGURE 108.1. Fixed Point Iteration in images.

- Study the convergence to a fixed point x satisfying $x = g(x)$.
- Study the role of the derivative $g'(x)$.
- What is Banach's Contraction Mapping Theorem?

109
Newton's Method

109.1 To Read

- Newton's method

109.2 To Do

Consider Newton's method for an equation $f(x) = 0$:

$$x(n+1) = x(n) - \frac{f(x(n))}{f'(x(n))} \quad \text{for } n = 0, 1, 2, ..., \tag{109.1}$$

which has the form of fixed point iteration for the function $g(x) = x - \frac{f(x)}{f'(x)}$.

Do the following

- Study computationally the convergence of the sequence $x(0), x(1), x(2), ..$ with different functions f.
- Experience convergence and divergence.
- Study the role of the condition $f'(x) \neq 0$.
- Study the rate of convergence.
- Show that $g'(x) = 0$ if $f'(x) \neq 0$ and f' is differentiable.
- Show quadratic convergence. (Hint)

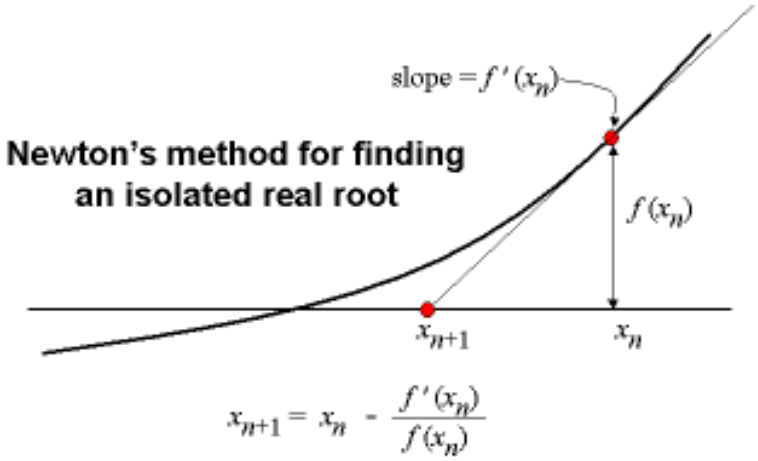


FIGURE 109.1. Newton's method.

110
Root Functions

110.1 To Do

- Show that for $a > 0$ the equation $x^2 = a$ has a unique positive solution $x = \sqrt{a} = a^{\frac{1}{2}}$.
- Compute $\sqrt{a}$ by Newton's method.
- Compute the solution of the equation $x^{12} = 2$ using Newton's method. Connect to well-tempered musical scales.

Prove the following rules for computing with the square root function (Hint):

- $\sqrt{a}\sqrt{b} = \sqrt{ab}$, that is $a^{\frac{1}{2}}b^{\frac{1}{2}} = (ab)^{\frac{1}{2}}$.
- $\sqrt{\frac{1}{a}} = \frac{1}{\sqrt{a}}$
- What about $a^{\frac{1}{m}}$ for $m = 3, 4, ..$?

FIGURE 110.1. Cylindrical Root Beer.

111
Maximum of a Continuous Function

Consider a Lipschitz continuous function $u : [a, b] \to \mathbb{R}$ defined on a closed bounded interval $[a, b]$.

111.1 To Do

- Does $u(x)$ attain a maxium value at $u(\xi)$ at some point $\xi \in [a, b]$?
- How to determine $u(\xi)$ and ξ?
- Write Python code for computing $u(\xi)$ and ξ?
- Can the code profit from knowledge of:
- the Lip constant of $u(x)$?
- if $u(x)$ is differentiable?
- What is the number of operations required to determine $u(\xi)$ to a certain precision?
- The same for ξ?

FIGURE 111.1. Reaching a maxium point.

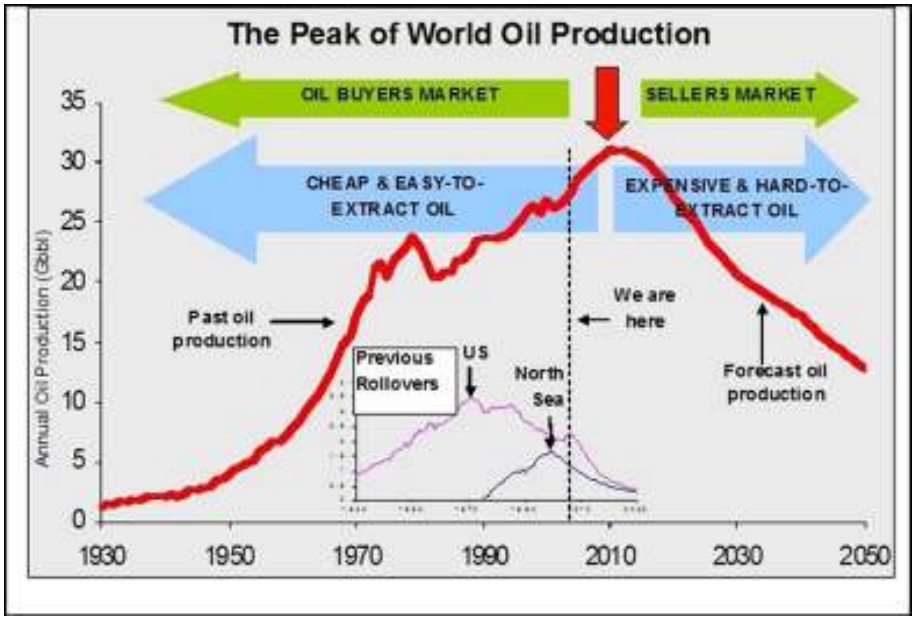


FIGURE 111.2. The production of oil is about to peak.

112
$\mathbb{R}^N$ as Vector Space

112.1 To Do

Write Python code for $\mathbb{R}^N$ as vector space including the operations

- vector addition,
- multiplication of vector by scalar,
- scalar product, length of vector,
- projection on subspace,
- orthogonal decomposition,
- linear independence, volume?

Assume that operations on real numbers are defined.

113
Kepler vs Newton

113.1 To Contemplate

Kepler got famous by showing experimentally that the planets move in elliptical orbits around the Sun satisfying Keplers three laws of motion. Newton got even more famous by proving mathematically that Kepler's three laws all follow from one single inverse square law of gravitation.

Why did Newton become more famous than Kepler? Theory vs practice? Is one law better than three?

113.2 To Do

Consider the basic two-body problem with one light body (the Earth) orbiting a heavy body (th Sun) under the action of a gravitational force, assuming the heavy body to be fixed (at the origin say).

- Verify Kepler's Laws computationally.
- Try to verify Kepler's Laws analytically.

113.3 Hint: Newton's Strike og Genius

According to Lagrange the *two-body* problem for a small mass orbiting around a heavy mass, such as the Earth moving around the Sun neglecting

FIGURE 113.1. Kepler dressed for party.

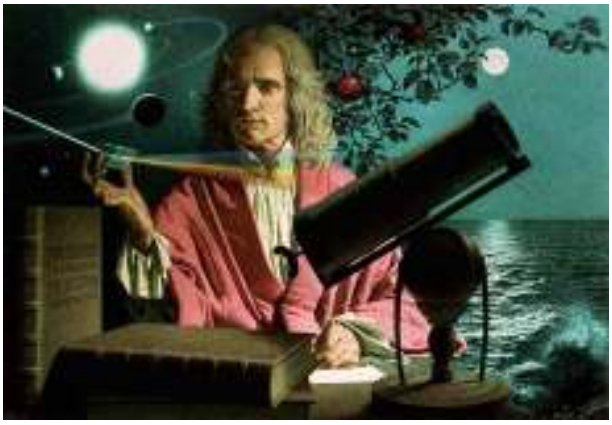


FIGURE 113.2. Newton controling the planets by brain power.

the influence of the other planets can be formulated as follows in polar coordinates (r, θ) with the origin at the center of the heavy mass:

$$\begin{cases} \ddot{r} - r\dot{\theta}^2 = -\frac{1}{r^2}, & t > 0, \\ \frac{d}{dt}(r^2\dot{\theta}) = 0, & t > 0, \end{cases} \tag{113.1}$$

complemented with initial values for position and velocity. Introducing the change of variables $u = r^{-1}$, show that $\dot{\theta} = cu^2$ for c constant. Use this relation together with the fact that the chain rule implies that

$$\frac{dr}{dt} = \frac{dr}{du}\frac{du}{d\theta}\frac{d\theta}{dt} = -c\frac{du}{d\theta} \quad \text{and} \quad \ddot{r} = -c^2u^2\frac{d^2u}{d\theta^2}$$

to rewrite the system (113.1) as

$$\frac{d^2u}{d\theta^2} + u = c^{-2}. \tag{113.2}$$

Show that the general solution of this equation is

$$u = \frac{1}{r} = \gamma\cos(\theta - \alpha) + c^{-2},$$

where γ and α are constants. Here γ controls the eccentricity with $\gamma = 0$ corresponding to the zero eccentricity of a circular orbit, and we can choose $\alpha = 0$ without loss og generality.

Finally, show that the solution is either an ellipse, parabola, or hyperbola, using the fact that these curves can be described as the loci of points for which the ratio of the distance to a fixed point and to a fixed straight line, is constant. Polar coordinates are suitable for expressing this relation.

Round off by proving Kepler's three laws for planetary motion using the experience you have gained.

113.4 Insolation and Glacial cycles

To connect to climate, consider the amount of insolation from the Sun depending on the distance to the Sun. The insolation scales like r^-2dt assuming the radiation from the Sun spreads spherically. Why? Because the the surface area of a sphere scales like r^2 and the area of the Earth absorbing radiation is fixed. Thus the insolation intensity $I(t)$ per unit of time is proportional to u^2, that is $I(t) = Cu^2$ for som constant C. The total insolation over a year period $[0, T]$ is thus given by

$$\int_0^T I(t)\,dt = C\int_0^T u^2\,dt = \frac{C}{c}\int_0^T \frac{d\theta}{dt}dt = \frac{C}{c}\int_0^{2\pi} d\theta = 2\pi\frac{C}{c}, \tag{113.3}$$

independent of the eccentricty parameter γ. The total amount of received radiation over a year is thus independent of the eccentricity.

This connects to the blog post Glacial Cycles and Eccentricty: Both glacial cycles and eccentricty shows a periodicity of 100.000 years (over the last million years), with glaciation coupled to small eccentricty. It appears that the stronger North Hemisphere Summer heating with larger eccentricty keeps the Ice Age away, presumably because the climate dynamics is stronger with hotter shorter NH Summer and more heat transfered North from the Equator during the long NH Winter.

114
Separable IVPs

114.1 To Read

- Separable IVPs

114.2 To Do: Take-Off

Consider

$$\dot{v} = 1 - v^2 \tag{114.1}$$

as model for take-off of an aircraft.

- Motivate the model.
- Find an analytical solution formula.
- Compare analytical and computational solution.
- Compare the time required to find analytical vs computational solution.

114.3 To Do: Fox-Rabbit Model

Consider

$$\dot{u} = au - buv, \quad \dot{v} = -\alpha + \beta uv \tag{114.2}$$

FIGURE 114.1. To live or not live?

where a, b, α and β are constant cofficients, as a predator-prey model.

- Motivate the model.
- Find an analytical solution formula.
- Compare analytical and computational solution.
- Compare the time required to find analytical vs computational solution.

115
Elementary Arithmetics

115.1 To Read

Recall the algorithms for addition, subtraction, multiplication and long division of integers, from elementary school education.

115.2 To Do

- Write Python code for integer computation with multidigit integers based on one-digit computation of addition,subtraction, multiplication and long division.

116
Iterative Methods for Linear Systems

116.1 To Browse

- Numerical Linear Algebra

116.2 To Do: Jacobi

Consider a linear system of equations

$$Ax = b, \tag{116.1}$$

where $A = (a_{ij})$ is a $d \times d$ matrix. Consider Jacobi iteration

$$a_{ii}x_i(n+1) = b - \sum_{j \neq i} a_{ij}x_j(n),\ i = 1, ..., d, \tag{116.2}$$

or more generally, damped Jacobi:

$$x(n+1) = x(n) - \alpha(Ax(n) - b), \tag{116.3}$$

where $\alpha = (\alpha_i)$ is a diagonal matrix with positive coefficients (chosing $\alpha_i = \frac{1}{a_{ii}}$ gives back Jacobi).

Do the following:

- Study convergence of Jacobi iteration for different A.

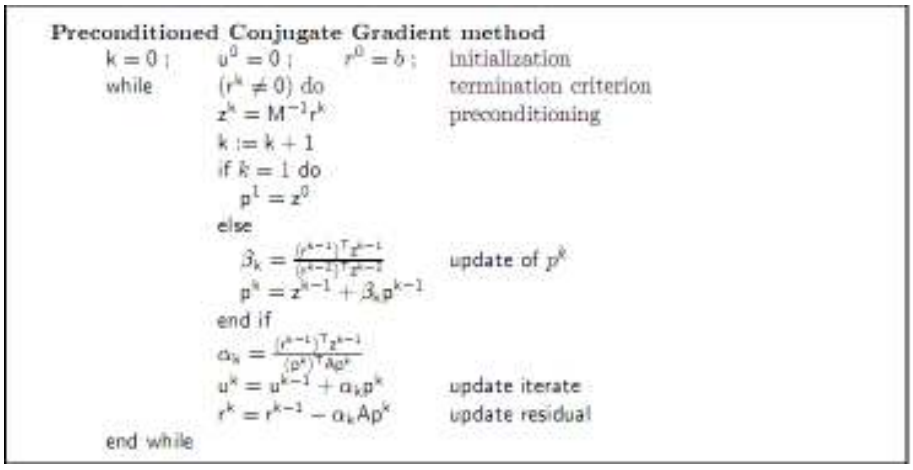


FIGURE 116.1. Preconditioned conjugate gradient method.

- Study convergence of damped Jacobi for different A.
- Study the IVP $\dot{x} + Ax = 0$ for $t > 0$, and connect to eigenvalues of A.

116.3 To Do: Conjugate Gradient

- Write Python code for the conjugate gradient method for a positive definite linear system.

117
Least Squares Method for $Ax = b$

117.1 To Read

- Least Squares Methods

117.2 To Do

Solve different (overdetermined) linear systems $Ax = b$ by a least squares method, where A is an $m \times n$ matrix with $m > n$. In other words, solve the positive (semi-definite) symmetric $n \times n$ system

$$A^\top A = A^\top b \tag{117.1}$$

by an iterative method or Gaussian elimination.

117.3 Connection to Singular Value Decomposition

Recall that the spectral decomposition of a symmetric $n \times n$ matrix B

$$Q^\top B^\top B Q = \Lambda, \tag{117.2}$$

where Λ is a diagonal $n \times n$ matrix with non-negative elements λ_i and Q is an orthogonal matrix (with $Q^\top Q = I$), is closely connected to a Singular Value Decomposition SVD of $A = U\Sigma V^\top$, where U and V are

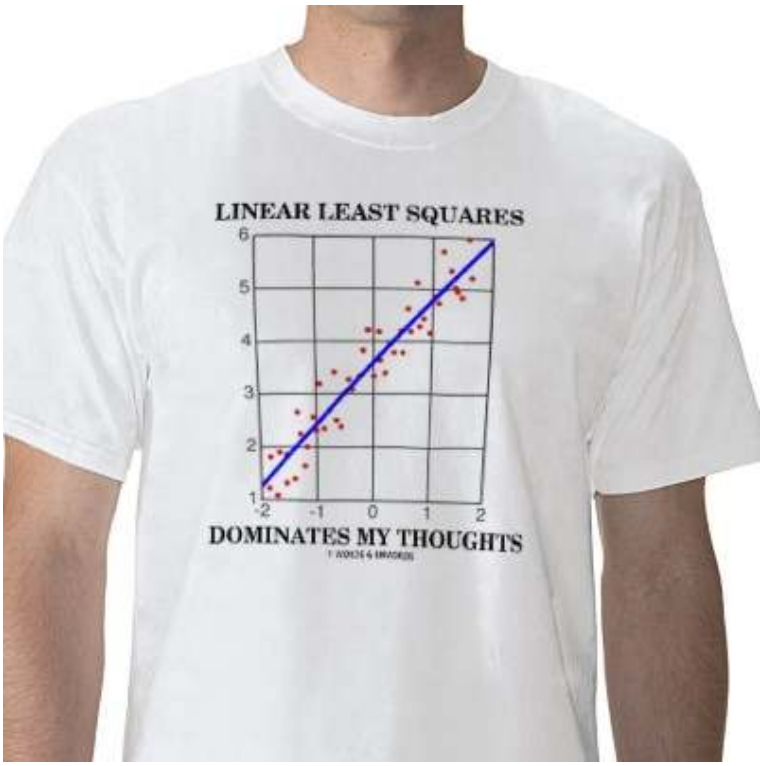


FIGURE 117.1. Obsession.

orthogonal matrices and Σ a diagonal $m \times n$ matrix with elements (singular values) $\sqrt{\lambda_i}$. Watch:

- Strang on SVD
- Strang Review of Linear Algebra

117.4 Principal Component Analysis

The eigenvectors corresponding to largest singular values can be used to give information about principal features of A, see:

- Singular Values and Principal Components.
- Singular Value Decomposition and Data Mining.

118
Calculus in Several Dimensions

118.1 To Read

- Integration in Several Dimensions
- The Divergence Theorem
- Green's and Stokes' Theorems

118.2 To Browse

- Calculus in Several Dimensions
- Divergence, Rotation and Laplacian
- Curve Integrals
- Surface Integrals
- Multiple Integrals
- Gauss' and Green's Theorems
- Stokes' Theorem

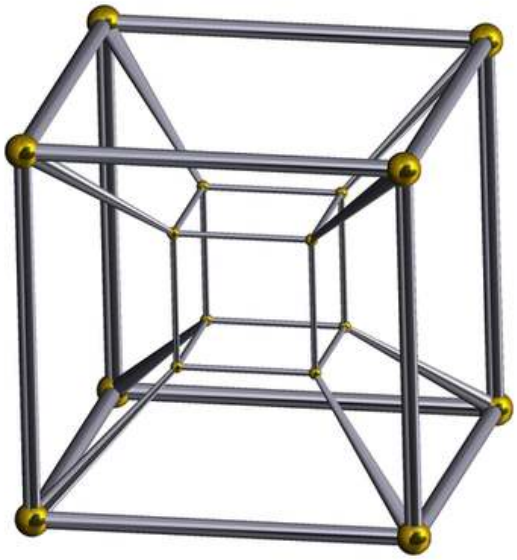


FIGURE 118.1. 4-dimensional cube.

118.3 To Do

Compute analytically

- gradient, divergence, rotation, Laplacian of selected functions,
- selected multiple integrals by iterated 1d integration,
- selected curve and surface integrals.

Show that

- level curves of $u : \mathbb{R}^d \to \mathbb{R}$ are orthogonal to ∇u,
- divergence and Laplacian are invariant under orthogonal coordinate transformations.

Compute by quadrature

- selected curve, surface and multiple integrals.

119
Piecewise Linear Interpolation

119.1 Defining the Interpolant

To estimate finite element discretization errors (in time and space), we are led to estimate the interpolation error between a given function $u(x)$ defined on a domain Ω and its *piecewise linear interpolant* u_h taking on the same values as $u(x)$ at the nodes of a triangulation. Here x represents time or a space coordinate. To estimate the interpolation error over the domain Ω it is sufficient to consider the error over each finite element separately, because the interpolant is uniquely defined by the nodal values for each element (interval, triangle or tetrahedron).

Recall that Midpoint Euler (or The Trapezoidal Method) constructs a solution to $\dot{u}(t) = f(u(t))$ for $t > 0$ as a continuous piecewise linear function $u(t)$ of time t.

119.2 To Do 1d

Consider a differentiable function $u(x)$ defined on the interval $[0, h]$ and let u_h be a linear function interpolating $u(x)$ at the end points, that is $u_h(0) = u(0)$ and $u_h(h) = u(h)$. We seek to estimate the interpolation error

$$e_h = u(x) - u_h(x) \quad \text{for } x \in [0, h]. \tag{119.1}$$

Step 1: Reduce to the case $u(0) = u(h) = 0$, by changing $u(x)$ and $u_h(x)$ by the same linear function. Notice that in this case $u_h(x) \equiv 0$ for $x \in [0, h]$.

Step 2: Assume that $u_h(x) \neq u(x)$ for some $x \in (0,h)$. Motivate that $u(x)$ takes on a maximum or minimimum value at some point $\xi \in (0,h)$ and show that $u'(\xi) = 0$.

Step 3: Use the differentiability of $u(x)$ to show that for $x \in [0,h]$

$$\begin{aligned} |u(x) - u(\xi)| = |u(x) - u(\xi) - u'(\xi)(x-\xi)| \leq C|x-\xi|^2 \leq Ch^2, \\ |u'(x) - u'(\xi)| \leq C|x-\xi| \leq Ch. \end{aligned} \tag{119.2}$$

Alternatively, use Taylor's formula or Taylor series with expansion around $x = \xi$.

Step 4: Conclude that

$$\begin{aligned} |u(x) - u_h(x)| &\leq C_0 C h^2, \quad (C_0 \approx \frac{1}{8}) \\ |u'(x) - u_h'(x)| &\leq C_1 C h, \quad (C_1 \approx \frac{1}{2}) \end{aligned} \tag{119.3}$$

where C bounds $|u''(x)|$ for $x \in [0,h]$.

Alternative proof: Let $x \in (0,h)$ and consider the function $g(y)$ defined for $y \in [0,h]$ by

$$g(y) = u(y) - u(0)\frac{h-y}{h} - u(h)\frac{y}{h} - \gamma(x)y(h-y), \tag{119.4}$$

where $\gamma(x)$ is so chosen that $g(x) = 0$. Notice that $g(0) = g(x) = g(h) = 0$ and use the mean-value theorem (first for $g(y)$ twice and then for $g'(y)$ once) to show that $g''(\xi) = 0$ for some $\xi \in (0,h)$ and thus that $\gamma(x) = -\frac{1}{2}u''(\xi)$. Then show that $C_0 = \frac{1}{8}$

119.3 Direct Computation of Interpolation errors

Let $\bar{u}_h$ be a piecewise quadratic interpolant of $u(x)$ interpolating at the endpoints and midpoint of each element. Use

$$\max_{x\in[0,h]} |\bar{u}_h(x) - u_h(x)|, \quad \max_{x\in[0,h]} |\bar{u}_h'(x) - u_h'(x)| \tag{119.5}$$

as direct quantitative estimates of the interpolation errors

$$\max_{x\in[0,h]} |u(x) - u_h(x)|, \quad \max_{x\in[0,h]} |u'(x) - u_h'(x)|. \tag{119.6}$$

Use this technique for estimation of errors in piecewise linear interpolation of different functions.

119.4 To Do in 2d and 3d

Extend to 2d and 3d.

119.5 Compare

- Piecewise Polynomials 1d
- Piecewise Polynomials 2d and 3d

119.6 Piecewise Constant Approximation

In *piecewise constant approximation* the interpolant u_h is defined as a constant on each finite element, e.g, as the mean-value over the element. Recall from Time Stepping Error Analysis

$$\max_{x\in[0,h]} |u(x) - u_h(x)| \leq h \max_{x\in[0,h]} |u'(x)|. \tag{119.7}$$

119.7 L_2-projection onto Piecewise Constants

Define $u_h(x)$ on $[0,h]$ as the mean-value of $u(x)$, that is,

$$u_h(x) = \frac{1}{h}\int_0^h u(y)\,dy \quad x \in [0,h]. \tag{119.8}$$

Show that $u_h(x)$ can be defined as the constant Pu defined by the orthogonality relation

$$\int_0^h (u(y) - Pu)v(y)\,dy = 0 \tag{119.9}$$

for all constant functions $v(y)$ on $[0,h]$. Show that the constant Pu is a best approximation of $u(y)$ in the sense that

$$\int_0^h (u(y) - Pu)^2\,dy \leq \int_0^h (u(y) - v(y))^2\,dy \tag{119.10}$$

for all constant functions $v(y)$. Hint: Write

$$\begin{aligned}\int_0^h (u - Pu)^2\,dy &= \int_0^h (u - Pu)(u - Pu)\,dy + \int_0^h (u - Pu)(Pu - v)\,dy \\ &= \int_0^h (u - Pu)(u - v)\,dy\end{aligned} \tag{119.11}$$

and use Cauchy's inequality for integrals.

Extend to piecewise constant approximation on a partition of an interval. Extend to $2d$ and $3d$.

Note: This is the basic step in the basic error analysis of the finite element method.

120
Quadrature

120.1 Quadrature by Piecewise Polynomial Interpolation

An integral $\int_I u(x)\,dx$ of a function $u(x)$ over an interval I, van be computed by replacing (interpolating) $u(x)$ by a piecewise polynomial interpolant (constant, linear, quadratic,...) u_h on some partition of I into subintervals, and computing the integral $\int_I u_h(x)\,dx$ analytically (as a sum of analytically computable integrals over subintervals). This is called (numerical) *quadrature*. The quadrature is then exact if $u(x)$ is a piecewise polynomial in question.

120.2 Trapezoidal Rule by Linear Approximation

Let $[0, h]$ be an interval and consider the quadrature formula (Trapezoidal Rule):

$$\int_0^h u(x)\,dx \approx \frac{h}{2}(u(0) + u(h)) \tag{120.1}$$

obtained by replacing the given function $u(x)$ by its linear interpolant

$$u_h = (1 - \frac{x}{h})u(0) + \frac{x}{h}u(h) \tag{120.2}$$

and computing the integral of u_h analytically.

The quadrature error can be estimated by reducing to the case $u(0) = u(h) = 0$ and then assuming that $u(x)$ is quadratic (linear plus one) so that $u(x) = x(h - x)$. We compute

$$\int_0^h x(h-x)\,dx = \frac{h^3}{6}, \tag{120.3}$$

suggesting the quadrature error estimate (since the second derivative of $x(h - x)$ equals -2):

$$|\int_0^h u(x)\,dx - \frac{h}{2}(u(0) + u(h))| \leq \frac{h^3}{12}\max_{[0,h]}|u''| \tag{120.4}$$

The accurcay of the Trapezoidal rule is thus of order h^2, when normalizing for the length h of the interval, so that over an interval I of unit length partitioned into subintervals of length h, the quadrature error is bounded by $\frac{h^2}{12}\max_I |u''|$.

120.3 To Do

- Estimate the quadrature error in the rectangle and trapezoidal quadrature rules corresponding to piecewise constant and linear linear approximation.
- Compare with Forward/Backward/Midpoint Euler time stepping.
- Contemplate adaptive quadrature with variable subinterval length.

120.4 To Read

- Adaptive Quadrature

121
Residual vs Output Error

121.1 To Read

Consider a linear system of equations

$$A\bar{x} = b \tag{121.1}$$

where A is a $d \times d$ matrix, $b \in \mathbb{R}^d$ a given vector and $\bar{x}$ an exact solution. Consider an approximate solution x with residual $R(x) = Ax - b \neq 0$, computed by some iterative method, (or by Gaussian elimination using single precision). Suppose we want to estimate the solution error $x - \bar{x}$ in terms of the residual $R(x)$ (computed in double precision).

Suppose we then choose a vector $\psi \in \mathbb{R}^d$, e.g. $\psi = (1, 1, ..., 1)$, and consider the output error

$$E(x) = (x - \bar{x}, \psi) \quad (= \sum_{i=1}^{d} (x_i - \bar{x}_i)). \tag{121.2}$$

Let then φ solve the dual problem $A^\top \varphi = \psi$ and note that

$$(x - \bar{x}, \psi) = (x - \bar{x}, A^\top \varphi) = (Ax - A\bar{x}, \varphi) = (R(x), \varphi) \tag{121.3}$$

which offers a representation of the output error $(x - \bar{x}, \psi)$ in terms of the scalar product of the residul $R(x)$ with the dual solution φ.

FIGURE 121.1. Residual is a cross-platform 3D game interpreter.

121.2 To Do

Use the error representation to estimate the output error $E(x)$ by solving the dual equation $A^\top \varphi = \psi$ approximatively by iteration (or single precision Gaussian elimination). Consider in particular:

- different weights ψ corresponding to mean-values, point-values and norms,
- A pos def with eigenvalues bounded away from zero (well-conditioned),
- A pos def with small eigenvalues (ill-conditioned)
- A anti-symmetric.

122
Adaptive Time-Step Error Control

122.1 To Read

- Time Stepping Error Analysis

122.2 To Do

- Extend the a posteriori error estimate for Midpoint Euler to variable time step.
- Formulate an algorithm for control of the error at final time in terms of variable time step.
- Test the algorithm on different problems with different characteristics.
- Prove an a posteriori error estimate for Backward Euler and formulate a corresponding algorithm.

123
Stability Analysis

123.1 To Read

- Time Stepping Error Analysis

123.2 To Do

Consider an IVP of the form

$$\dot{u} + Au = 0 \quad \text{for } t > 0, \quad u(0) = u^0, \tag{123.1}$$

where A is a $d \times d$ constant matrix. Consider the following two basic cases

- (1) A symmetric positive (semi-)definite,
- (2) A anti-symmetric ($A^\top = -A$),
- (3) A symmetric negative definite,
- (4) A anything.

Show in case (1) the following stability estimate:

$$|u(t)|^2 + 2\int_0^t (Au, u)\, dt = |u^0|^2. \tag{123.2}$$

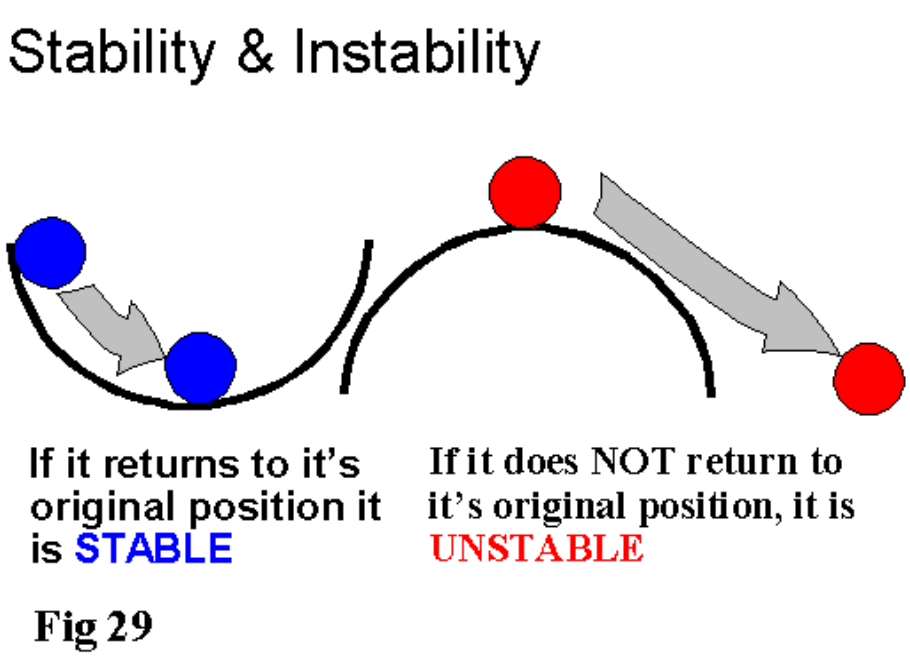


FIGURE 123.1. The difference between stable and unstable.

In particular

$$S_d(T) \equiv \max_{0\le t\le T} \frac{|u(t)|}{|u^0|} \le 1 \tag{123.3}$$

Show in case (2) the following stability estimate:

$$|u(t)|^2 = |u^0|^2, \tag{123.4}$$

that is $S_d(T) = 1$ for $T > 0$.

Show in the case (3) and (4) that it is possible that $S_d(T)$ grows exponentially with T and thus may be very large.

Show that (1) is representative of diffusion. Show that (2) is representative of wave propagation. Compute solutions in concrete illustrating cases.

124
Analytical Mechanics

> Rotation is a universal phenomenon; the Earth and all the other members of the solar system rotate on their axes, the satellites revolve aoround the planets, the planets revolve around the Sun, and the Sun itself is a member of the galaxy or Milky Way system which revolves in a very remarkable way. How did all these rotary motions come into being? And what part do they play in the system of the world? (E. T. Whittaker in a lecture on *Spin in the Universe*, author of A Treatise on the Analytical Dynamics of Particles and Rigid Bodies, 1904)

124.1 Degrees of Freedom

In analytical mechanics one seeks to describe a mechanical system with few cleverly chosen degrees of freedom, in order to allow analytical solution of the equations of motion. The starting point is always Newton's laws, with the 2nd Law expressing conservation of (linear) momentum directly connecting to the translation of the center of mass, which imply conservation of also *angular momentum* connecting to rotation around the center of mass.

The motion of an N-body (N-particle system connected by springs) can be described by $6N$ degrees of freedom (3 for position and 3 for velocity for each particle). If the springs are very stiff, the N-body is close to a *rigid body* which does not change form under external forcing. The motion of a

rigid N-body can be described the motion of its center of mass together with rigid rotations around the center of mass, together at most 6 degrees of freedom.

124.2 To Read

- Lagrange's equations
- N-body mechanics

124.3 Conservation of Linear Momentum

Consider an N-body B consisting of set of particles of mass m_i located at $x^i(t)$ at time t, $i = 1, ..., N$, somehow connected by springs. Newton's 2nd Law for each particle reads

$$m_i\ddot{x}^i = F^i \equiv F^i_{int} + F^i_{ext}, \tag{124.1}$$

where F^i_{int} is the internal spring force and F^i_{ext} an exterior force acting on particle i. Assuming that exterior forces vanish and using that by Newton's 3rd Law the internal spring forces sum to zero, we have by summation

$$\frac{d^2}{dt^2}\sum_1^J m_i x^i = \frac{d^2}{dt^2}MX = 0 \tag{124.2}$$

where $M = \sum_1^J m_i$ and

$$X = \frac{1}{M}\sum_1^J m_i x^i \tag{124.3}$$

is the position of the *center of mass* of B. In other words, without exterior forcing the center of mass of B moves along a straight line $X = Vt + X^0$, where

$$V = \dot{X} = \frac{1}{M}\sum_1^J m_i \dot{x}^i \tag{124.4}$$

is a constant velocity, and X^0 the position of the center of mass at $t = 0$.

124.4 Conservation of Angular Momentum

The *angular momentum* L of B with respect to the origin is defined by

$$L = \sum_1^J x^i \times m_i\dot{x}^i. \tag{124.5}$$

We have since by the 3rd Law $\sum_i x^i \times F^i_{int} = 0$ (verify this!),

$$\dot{L} = \sum_i \dot{x}^i \times m_i \dot{x}^i + \sum_i x^i \times m_i \ddot{x}^i = \sum_i x^i \times F^i_{ext} = 0 \tag{124.6}$$

if $F^i_{ext} = 0$ for $i = 1, .., J$. Angular momentum is thus conserved if exterior forces vanish.

If we write $x^i = X + \bar{x}^i$ and $\dot{x}^i = V + \bar{v}^i$ with $\bar{x}^i$ and $\bar{v}^i$ position and velocity with respect to the center of mass, we have

$$\sum_i m_i \bar{x}_i = 0, \quad \sum_i m_i \bar{v}^i = 0, \tag{124.7}$$

and we can thus write

$$L = \sum_i (X + \bar{x}^i) \times m_i (V + \bar{v}^i) = X \times MV + \sum_i \bar{x}^i \times m_i \bar{v}^i, \tag{124.8}$$

eapressing the angular momentum L as the sum of the angular momentum of the center of mass and the angular momentum with respect to the center of mass.

124.5 Moment of Inertia of a Rigid Body

We now assume that B is a *rigid body* connected by very stiff springs maintaing the form of B. If B is rotating with angular velocity ω around a unit vector n, the velocity $\dot{x}^i$ of particle i is given by

$$\dot{x}^i = \omega n \times x^i \tag{124.9}$$

and thus the angular momentum L_n with respect to the direction n, is given by

$$L_n = n \times L = \sum_i m_i n \times x^i \times (\omega n \times x^i) = \omega \sum m_i r_i^2 \equiv \omega I_n, \tag{124.10}$$

where

$$I_n \sum m_i r_i^2 \tag{124.11}$$

is the *moment of inertia* with respect to n with r_i the distance of particle i to the axis of rotation. The angular velosity ω is thus conserved in the absence of exterior forcing. Accordingly, decreasing I_n will correspond to increasing ω in the absence of exterior forcing, as shown by the spinning ice-skater.

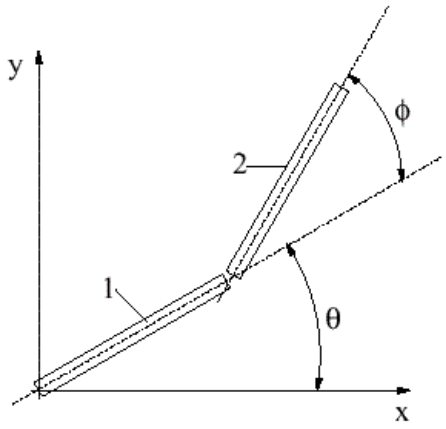


Figure 1: Two link system in horizontal plane.

FIGURE 124.1. A simple rigid body system described by two angular variables. Can you write down the equations of motion?

124.6 To Do

- Compute the center of mass and the moments of inertia with respect to different axes of rotation for various rigid bodies.
- Formulate equations of motion as conservation of linear momentum with respect to the center of mass, and conservation of angular momentum.
- Solve the equations of motion analytically in some simple cases, and computationally else.

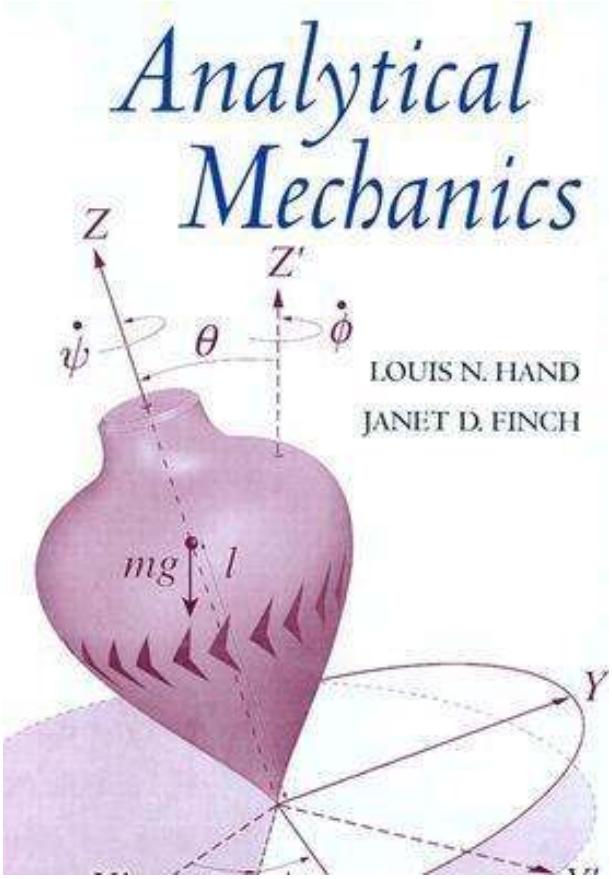


FIGURE 124.2. The masterpiece by Langarnge.

125
Finite Element Programming

See BodySoul Session F.

126
Tool Bag Proof Inspection

To test your level of theoretical knowledge you can take on to prove the analytical results stated in Tool Bags including

- rules for differentiation,
- rules for integration,
- Fundamental Theorems.
- convergence of fixed point iteration and Newton's method,
- ...

The idea is that knowledge of the construction of a tool, is helpful when using the tool: Understanding how a plyer is designed can be helpful when figuring out how to apply it.

126.1 To Do

- Prove (some of) the analytical results listed in Tool Bags.

127
Climate Sensitivity

Some aspects of climate have not been observed to change.
(IPCC Summary for Policymakers 2007)

127.1 To Browse

- On Climate Sensitivity 1, 2, 3, 4, 5, 6.
- The Incorrect Postulate of Climate Alarmism 1, 2.
- What is the GreenHouse Effect, Really?
- The Maximal GreenHouse Effect

127.2 A Simple Model

Consider the following model for the vertical heat transfer in the atmosphere:

$$\begin{aligned} \dot{T} + \beta T' + \alpha T - \epsilon T'' = q \quad \text{for } t > 0,\ 0 < x < 1, \\ -\epsilon T'(0,t) = Q(t), \quad -\epsilon T'(1,t) = 0 \quad \text{for } t > 0 \end{aligned} \tag{127.1}$$

where $x \in [0,1]$ is a vertical coordinate, $T(x,t)$ is atmosphere temperature at x at time $t > 0$, $\alpha(x,t)$ is a coefficient of net outgoing radiation, $\beta(x,t)$

is a vertical convection velocity, ϵ a heat conduction/diffusion coefficient, $Q(t)$ is incoming heat flux from the ocean (originating from insolation), and $q(x,t)$ is a internal heat source from evaporation/condensation. As usual, $\dot{T} = \frac{\partial T}{\partial t}$ and $'T = \frac{\partial T}{\partial x}$.

127.3 Without Convection-Radiation-Evaporation-Condensation

The basic stationary case is $Q(t) = 1$ constant, $\beta = \alpha 0$, $q = 0$ and ϵ constant, which gives $T = \frac{1-x}{\epsilon}$, with corresponding temperature sensitivity $T(0) = \frac{1}{\epsilon}$ (which is large if ϵ is small). This a case of high temperature sensitivity connected to heat conduction/radiation driven by negative temperature gradients.

127.4 With Convection-Radiation-Evaporation-Condensation

Consider now the case has $q = -1$ for $0 < x < 0.5$ (evaporation) and $q = 1$ for $0.5 < x < 1$ (condensation), ϵ small and $\beta = 1$, which gives $T(x) \approx -x$ for $0 < x < 0.5$ and $T(x) \approx x - 1$ for $0.5 < x < 1$, with corresponding temperature sensitivity $T(0) = 0$. This is a case of small temperature sensitivity connected to convection combined with evaporation/condensation with temperature gradients of varying sign.

We conclude that convection coupled with evaporation/condensation, like in the real atmosphere, can change temperature sensitivity drastically, and since climate sensitivity is related to temperature sensitivity, also climate sensitivity can be drastically reduced as compared to that of the simplest radiative model. Compare with Greenhouse Gas. Conclusion?

Compare with a common green house, which has high temperature sensitivity because convection is prevented by the glass enclosure. Compare also with a boiling pot where increasing the forcing results in more vigorous boiling while the temperature stays the same, resulting in zero temperature sensitivity is zero.

127.5 To Do

- Study the dynamics of the model under varying forcing.
- Estimate global climate sensitivity from the model.

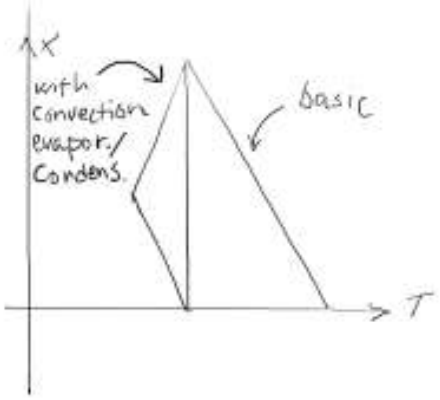


FIGURE 127.1. Model temperature distribution in the atmosphere without and with convection and evaporation/condensation.

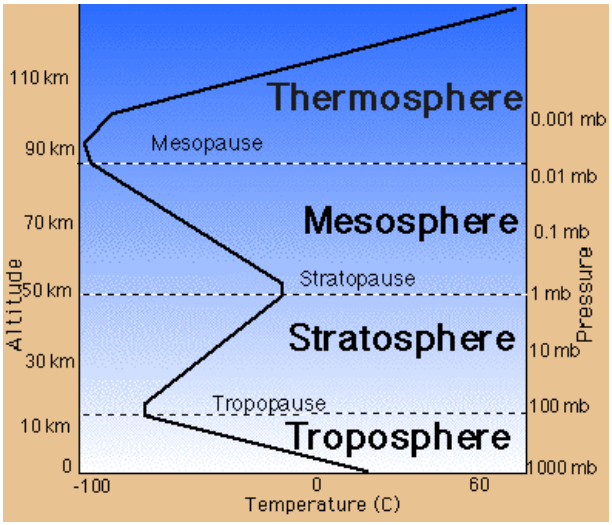


FIGURE 127.2. Real temperature distribution in the atmosphere. Notice similarity in troposphere-stratosphere with model.

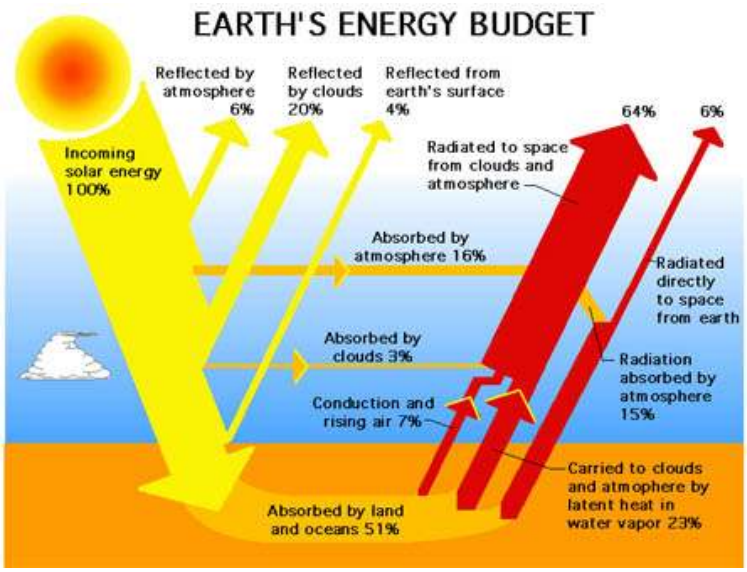


FIGURE 127.3. The energy budget of the Earth according to NASA.

128
From Google to Googol

128.1 To Read

One googol is equal to 10^{100}. The name Google is inspired by googol as a very large number. How large isthen a googol? Let us compare with

- 10^{80} = estimated number of atoms in the Universe
- 2×10^{16} bytes processed by Google each day (is Google close to googol?)
- 10^{14} number of neural connectins in a human brain
- 10^{14} number of cells in a human body
- 6×10^{23} number of atoms in a mole
- 4×10^{17} age of the Universe in seconds
- 10^{30} all possible passwords of 40 characters
- 10^{13} US budget deficit in Swedish Crowns
- 6×10^{-34} Planck's constant, supposedly the smallest quantum of energy
- 4×10^{9} clock rate of Pentium4 microprocessor.
- $\exp(LT) \approx 10^{100}$ if $L = 20, T = 10$: Compare Illusion of exponential growth

FIGURE 128.1. Google servers.

128.2 To Do

- Compute $S_N = \sum_{n=1}^{N} \frac{1}{n} = 1 + \frac{1}{2} + ... + \frac{1}{N}$ for different increasing N.
- Argue that $S_N \approx \log(N)$. How large does S_N become, computationally?
- Compare computing S_N summing from left to right (smaller and smaller terms) with summing from right to left (larger and larger terms). Explain the discrepancy.
- What is the machine epsilon (precision) of your computer?
- What is the range of integers your computer is willing to deal with?

129

Equivalence of Inertial and Gravitational Mass

> Then let the beginning of our reflections be the consideration that whatever motion comes to be attributed to the Earth must necessarily remain imperceptible to us and as if nonexistent, so long as we look only at terrestrial objects; for as inhabitants of the Earth, we consequently participate in the same motion.... (Salvatio in Galileos Dialogue, Day 2)

129.1 To Read: The Origin of Newton's Law

Einstein got famous by stating that *inertial and gravitational mass cannot be distinguished*. Let's see if this statement is any reason to be famous, and let us at the same time seek to understand from where Newton's 2nd Law comes. Newton's 2nd Law states

$$M\frac{dv}{dt} = F \tag{129.1}$$

where M is the *inertial mass* of a body B moving with velocity v (with respect to a certain coordinate system) subject to a force F We note that Newton's 2nd Law is *Galilean invariant*, that is it takes exactly the same form in all coordinate systems which move with uniform velocity with respect to each other and use the same length (and time) scale. This is because the accelleration $A = \frac{dv}{dt}$ is invariant under coordinate transformations with constant velocity.

Newton's law $F = MA$ expresses superposition of inertial mass as linearity in M: If we put two masses together to form one mass, then the corresponding forces add up, which expresses conservation of mass: Putting two masses into a basket does not make part of the combined mass mysteriously disappear (as it does in a fusion reactor like the Sun, where part of the mass is turned into energy). Newton's law $F = MA$ is thus linear in inertial mass M, as an expression of conservation of mass.

Secondly, if we apply the double force 2F by first applying the F to give the accelleration A with respect to an initial rest state, and then the remaining F to give accelleration with respect to the current state, then the accellerations will add up to 2A with respect to the initial rest state, as a consequence of Galilean invariance. Newton's 2nd Law is thus linear also in accelleration A, as an expression of Galilean invariance.

Newton's law $F = MA$ thus can be viewed to express Galilean invariance and conservation of mass. We can agree to define one of the three variables F, A and M in terms of the other: For example, we can agree to define inertial mass M in terms of F and $A = dv/dt$ as $M = F/A$.

129.2 To Read: Gravitational Mass

Let now M_g be the gravitational mass of the body B of nertial mass M. This means that the body is acted upon by a force $F = M_g f$ where f is a normalized gravitational force. Newton's 2nd Law now takes the form

$$M\dot{v} = F = M_g f \tag{129.2}$$

in a fixed x-coordinate system. The body B is thus falling freely under the gravitational force $F = M_g f$. Let us now consider a different x' coordinate system with origin O', which we can imagine being in free fall under the normalized gravitational force f, thus with velocity v_0' in the x system satisfying

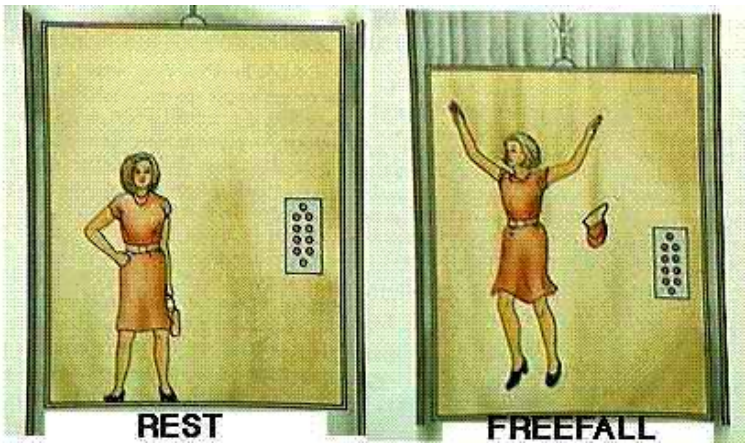
$$\dot{v}_0' = f \tag{129.3}$$

Let us now write $x = x_0' + x''$ and $v = v_0' + v''$, where x_0' and v_O' are the position and velocity of O' in the x-system, with x'' and v'' being position and velocity of B in the x'-system. Now, observe that

$$\frac{M_g}{M} f = \dot{v} = \dot{v}_0' + \dot{v}'' = f \tag{129.4}$$

where $v'' = 0$, because both the body and the coordinate system are falling freely in the same way (this is the assumption). This shows that $M_g = M$, thus that gravitational mass is the same as inertial mass. In other words, there is just one mass, that is, inertial mass.

How do we understand that all bodies fall frely the same way? Well, if not bodies would be torn apart under free fall, because different parts would



tend to fall differently. For example, the person in the free falling elevator, would bump the head into the ceiling or still feel pressure under the feet. And this is not what seems to be the case. Convinced? What about doing an elevator experiment?

129.3 To Read: How to Determine Mass?

How do determine the mass M of a body B? Since inertial and gravitational mass are the same, we have two possibilities:

- Measure aceeleration A under a given force F, and determine $M = \frac{F}{A}$.
- Hang the body in a spring and measure spring force F, and determine $M = \frac{F}{f}$, where f is strength of the gravitational field.

129.4 To Browse

- Does the Earth Rotate?

129.5 To Do

- Is Einstein's Equivalence prinicple a deep new physical law worthy of Nobel Prize or "just" a trivial consequence of Newton's 2nd Law?
- Did Galileo understand that all bodies fall freely the same way?

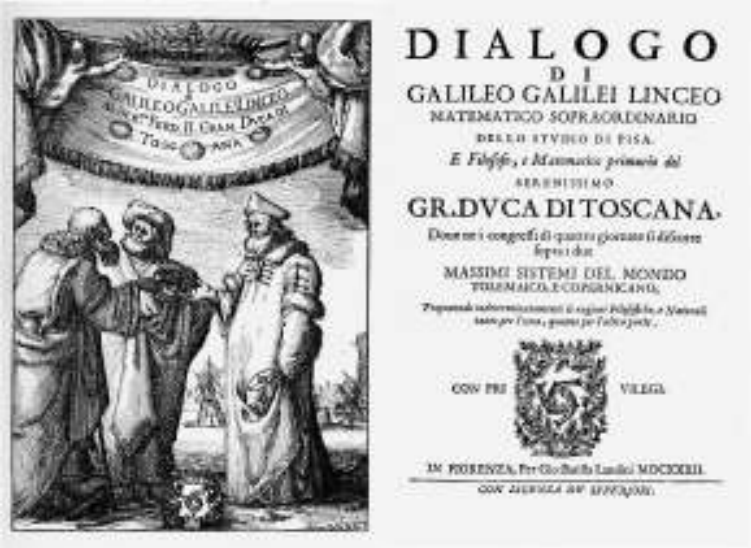


FIGURE 129.1. Title page of Galileo's Dialogue Concerning the Two Chief World Systems. See also Summary.

> For consider: Motion, in so far as It is and acts as motion, to that extent exists relatively to things that lack it; and among things which all share equally in any motion, it does not act, and is as if It did not exist. Thus the goods with which a ship is laden leaving Venice, pass by Corfu, by Crete, by Cyprus and go to Aleppo. Venice, Corfu, Crete, etc. stand still and do not move with the ship; but as to the sacks, boxes, and bundles with which the boat is laden and with respect to the ship itself, the motion from Verflice to Syria is as nothing, and in no way alters their relation among themselves. This is so because it is common to all of them and all share equally in it. If, from the cargo in the ship, a sack were shifted from a chest one single inch, this alone would be more of a movement for it than the two-thousand-mile journey made by all of them together. (Salvatio in Galileos Dialogue, Day 2)

Part VIII

World of Differential Equations

130
Conservation Laws

> A State without the means of some change is without the means of its conservation. (Edmund Burke)

130.1 Basic Laws of Solid/Fluid Mechanics

The basic laws describing mechanics of material bodies in the form of gases-fluids-solids, express

- *conservation laws* expressing conservation of mass, momentum and total energy,
- *constitutive laws* connecting stress to material motion/deformation.

The form of the conservation laws are given with conservation of momentum expressing Newton's 2nd Law, conservation energy reflecting the definition of total energy as the sum of kinetic and heat energy, and conservation of mass reflecting that matter is non-destructible (in the absense of nuclear fusion).

The modeling thus concerns the constitutive laws, which range in difficulty from complex non-linear to simple linear, like modeling the stress-strain relation of a spring.

The conservation and constitutive laws are expressed as Partial Differential Equations or PDEs using combinations of the following differential operators:

- time derivative $\frac{\partial}{\partial t}$
- gradient ∇
- divergence $\nabla\cdot$
- rotation $\nabla\times$
- Laplacian Δ .

with the space derivates referring to a fixed orthogonal coordinate system in $\mathbb{R}^3$. We will refer to this fixed coordinate system, which does not change with the motion of material bodies, as *Eulerian*, to be distinguished from a *Lagrangian* system deforming with material motion.

130.2 Read More

- Laplacian Models.

130.3 Conservation of Mass

Consider matter of density $\rho(x,t)$ moving in a volume $\Omega \subset \mathbb{R}^3$ with velocity $v(x,t)$ where x are coordinates in a fixed coordinate system in $\mathbb{R}^3$. Consider a fixed small volume $V \subset \Omega$ with boundary B. The amount of material flowing into of V through the B per unit time step, is given by

$$-\int_B \rho v \cdot n\, ds = -\int_V \nabla\cdot(\rho v)\, dx, \tag{130.1}$$

where n is the outward unit normal to B and we used the Divergence Theorem. The following equation expresses that mass is conserved: What goes in increases what's inside:

$$\frac{\partial}{\partial t}\int_V \rho\, dx = -\int_V \nabla\cdot(\rho v)\, dx, \tag{130.2}$$

that is

$$\int_V (\dot\rho + \nabla\cdot(\rho v))\, dx = 0 \tag{130.3}$$

and thus since V is an arbitrary volume in Ω:

$$\dot\rho + \nabla\cdot(\rho v) = 0 \quad \text{in } \Omega, \tag{130.4}$$

which is the conservation law expressing conservation of mass of density ρ moving with velocity v.

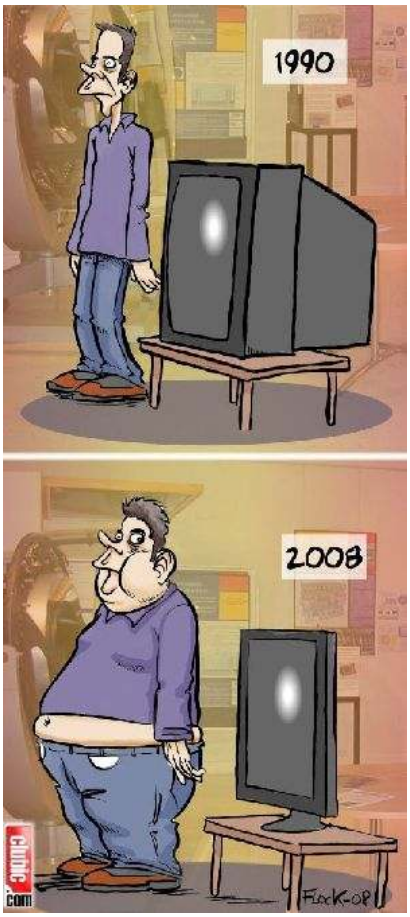


FIGURE 130.1. Conservation of mass.

130.4 Conservation of Momentum

Conservation of momentum $m = \rho v$ of matter subject to force ∇p from a pressure p, is expressed by replacing ρ in (130.4) by each of the components of momentum $m_i = \rho v_i$ and balancing by the force according to Newton's 2nd law, to get

$$\dot{m}_i + \nabla \cdot (m_i v) = -\frac{\partial p}{\partial x_i} \quad \text{in } \Omega, \tag{130.5}$$

or in short vector notation

$$\dot{m} + \nabla \cdot (mv) + \nabla p = 0 \quad \text{in } \Omega. \tag{130.6}$$

130.5 Conservation of Total Energy

Conservation of total energy ϵ is expressed by replacing ρ in (130.4) by ϵ and compensating with the the rate of work $-pv \cdot n$ performed on the boundary B of the control volume V according to the Divergence Theorem

$$-\int_B pv \cdot n \, ds = -\int_V \nabla \cdot (pv) \, dx, \tag{130.7}$$

to get the conservation law expressing conservation of total energy:

$$\dot{\epsilon} + \nabla \cdot (\epsilon v + pv) = 0 \quad \text{in } \Omega. \tag{130.8}$$

130.6 Conservation of Mass, Momentum, Energy

We collect the basic conservation laws of mass, momentum and energy:

$$\begin{aligned} \dot{\rho} + \nabla \cdot (\rho v) &= 0, \\ \dot{m} + \nabla \cdot (mv) + \nabla p &= 0, \\ \dot{\epsilon} + \nabla \cdot (\epsilon v + pv) &= 0, \end{aligned} \tag{130.9}$$

which is a very concise efficient mathematical formulation of mechanics in terms of mass ρ, momentum $m = \rho v$ with v velocity, and total energy $\epsilon = \frac{1}{2}\rho|v|^2 + e$ with e heat energy. A constitutive law for the pressure p completes the model for a gas/fluid: For a compressible perfect gas $p = (\gamma - 1)e$ with γ and gas constant. For a nearly incompressible gas/fluid $\delta \Delta p = \nabla \cdot v$ with δ a small positive (regularization) parameter. Effects from viscosity, heat conductivity and mass diffusion are here not included, and will be added later.

Constitutive laws for solids are given in Chapter 132.

130.7 To Think About

- Suppose $\rho v \cdot n = 0$ on the boundary of Ω. What can you say about the total mass $\int_\Omega \rho(x,t) \, dx$ as a function of time t?
- The same about m and ϵ?

131
Initial and Boundary Conditions

Science is a differential equation. Religion is a boundary condition. (Alan Turing)

We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes. (Laplace)

131.1 Dirichlet, Neumann and Robin Boundary Conditions

The conservation and constitutive differential equations are formulated over domains Ω in $\mathbb{R}^n$, $n = 1, 2, 3$, and are complemented by initial conditions and boundary conditions according to the following specifications, in the presence of viscous or diffusive effects:

- If an unknown $u(x,t)$ appears with a time derivative $\dot{u}$, then an initial condition $u(x,0) = u^0(x)$ for $x \in \Omega$ is specified. If a second time derivative $\ddot{u}$ appears, then also $\dot{u}(x,t)$ requires an initial condition.

- The boundary conditions for an unknown $u(x,t)$ specified on (part of) the boundary Γ of Ω, may take one of the following forms:

Dirichlet condition:

$$u(x,t) = \hat{u}(x,t) \quad \text{for } x \in \Gamma, t > 0, \tag{131.1}$$

where $\hat{u}$ is given on Γ.

Neumann condition:

$$\nu\frac{\partial u}{\partial n}(x,t) = g(x,t) \quad \text{for } x \in \Gamma, t > 0, \tag{131.2}$$

where g is given on Γ and

$$\frac{\partial u}{\partial n} = n \cdot \nabla u \tag{131.3}$$

is the normal derivative in the outward normal direction n to Γ. We will see that the Neumann condition is related to the presence of the term $-\nu\Delta u$ in a conservation law. This is the reason ν appears as a coefficient of the normal derivative.

Robin condition:

$$\alpha u(x,t) + \nu\frac{\partial u}{\partial n}(x,t)\alpha = \alpha\hat{u}(x,t) + g(x,t) \quad \text{for } x \in \Gamma, t > 0, \tag{131.4}$$

where $\alpha \geq 0$ is a coefficient. With $\alpha = 0$ we get the Neumann condition, and with α large and/or ν small, we get the Dirichlet condition.

For Maxwell's equations of electromagnetics involving the operator $\nabla\times$, also the boundary conditions $u \times n$ and $u \cdot n$ appear.

131.2 Essential and Natural Boundary Conditions

We shall see that a Neumann boundary condition reflects the presence of the Laplacian in the differential equation (modeling viscous or diffusive effects), and therefore in variational formulations do not have to be explicitely enforced like a Dirichlet condition. For this reason a Neumann, and also a Robin, condition is called a *natural boundary condition*, while a Dirichlet condition is called an *essential boundary condition*.

131.3 How Many Boundary Conditions to Specify?

As a general rule, we expect to have to specify either Dirichlet or Neumann/Robin boundary conditions on a given portion of the boundary. We

FIGURE 131.1. Non-natural boundary condition.

shall see this requires the presence of a viscous term of the principal form $-\nu\Delta u$ with $\nu > 0$. We can always assume the presence of such a term just by assuming ν to be sufficiently small, but we must then be prepared to see *boundary layers* developing at Dirichlet boundaries where material particles leave the computational domain, with a boundary layer being narrow zone along the boundary where a variable changes rapidly to satisfy a specied Dirichlet condition. Neumann conditions also give rise to boundary layers, but with much smaller change of variable values, since it is the normal derivative which may have to change quickly.

We will discover that at boundary portions where material particles enter the computational domain, Dirichlet boundary conditions will have to be imposed, while the choice is free at other parts.

131.4 To Think About

- What interpretation can a Robin boundary condition have?

> If we knew exactly the laws of nature and the situation of the universe at the initial moment, we could predict exactly the situation of that same universe at a succeeding moment. but even if it were the case that the natural laws had no longer any secret for us, we could still only know the initial situation approximately. If that enabled us to predict the succeeding situation with the same approximation, that is all we require, and we should say that the phenomenon had been predicted, that it is governed by laws. But it is not always so; it may happen that

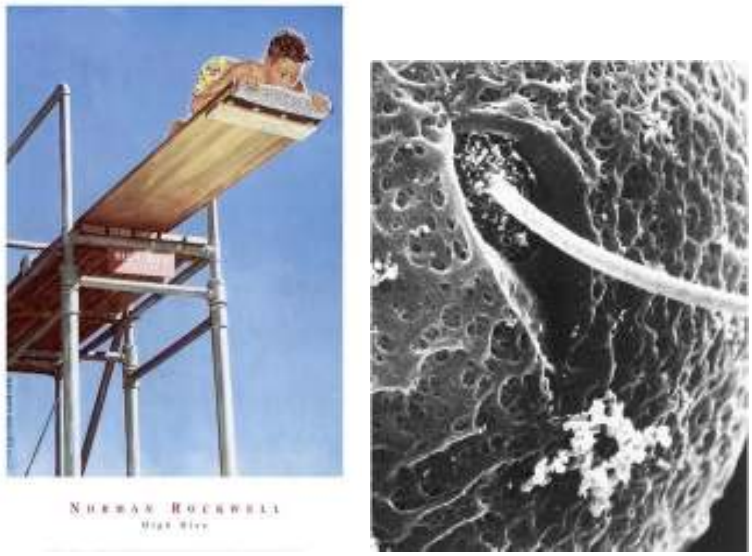


FIGURE 131.2. Initial conditions.

small differences in the initial conditions produce very great ones in the final phenomena. A small error in the former will produce an enormous error in the latter. Prediction becomes impossible, and we have the fortuitous phenomenon. (Poincaré in a 1903 essay *Science and Method*)

132
Constitutive Laws: Fluids/Solids

To those who ask what the infinitely small quantity in mathematics is, we answer that it is actually zero. Hence there are not so many mysteries hidden in this concept as they are usually believed to be. (Euler)

Madam, I have come from a country where people are hanged if they talk. (Euler to Queen Mother of Prussia, on his lack of conversation in his meeting with her, on his return from Russia)

The reader will find no figures in this work. The methods which I set forth do not require either constructions or geometrical or mechanical reasonings: but only algebraic operations, subject to a regular and uniform rule of procedure. (Lagrange)

It took the mob only a moment to remove his head; a century will not suffice to reproduce it. (Lagrange about the chemist Lavoisier)

132.1 Eulerian and Lagrangian Descriptions

We can describe natural phenomena in concise mathematical form using differential equations expressing laws of balance in material *continuous media* modeling fluids and solids. To model fluids it is efficient to use an Eulerian description with observers tied to fixed points in space which do not

FIGURE 132.1. Euler after losing one eye, and Lagrange.

move with material particles. To model solids it is often convenient to use a Lagrangian description with observers tied to moving material particles.

Eulerian coordinates refer to a fixed space-time coordinate system, while Lagrangian coordinates refer to a coordinate system which deforms with the motion of material particles.

In a Lagrangian description individual material particles have markers making it possible to directly follow their motion. In an Eulerian description, particles have no markers; instead velocities are recorded at fixed points in space of particles which happen to pass. From such velocities it is possible to follow particle trajectories in secondary step.

If you float along with the flow of in a river, then your observations of the flow will be Lagrangian. If you tie yourself to the shore, then your observations will be Eulerian.

The World emerges in an interaction of fluids and solids we need to find a unified formulation combining Eulerian and Lagrangian descriptions. We show that this is possible expressing Newton's 2nd Law in Eulerian form, and allowing the computational mesh to deform with the solid in Lagrangian form.

We start with fluid dynamics in Eulerian form, proceed to solid mechanics in total Lagrangian and updated Lagrangian form, and then present a unified formulation of so-called *Arbitary Eulerian-Lagrangian ALE* form.

132.2 Fluids vs Solids

In a viscous fluid the *viscous stress* σ (with pressure subtracted) is related to *velocity strain* $\epsilon(v) = (\epsilon_{ij}(v))$ as the symmetric form of the gradient ∇v

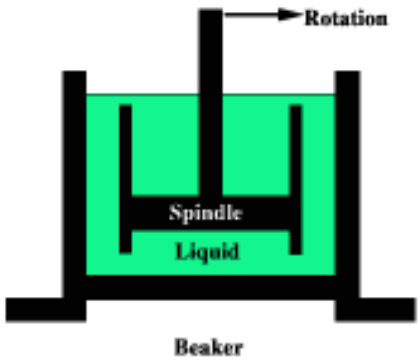


FIGURE 132.2. Viscosimeter for measuring viscosity. How does it work?

of the velocity $v = (v_1, v_2, v_3)$,

$$\epsilon(v) = \frac{1}{2}(\nabla v + \nabla v^\top) = (\epsilon_{ij}(v)); \quad \epsilon_{ij}(v) = \frac{1}{2}(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_i}{\partial x_j}). \tag{132.1}$$

In the basic case of a *Newtonian fluid* by a linear constitutive law

$$\sigma = 2\nu\epsilon(v), \tag{132.2}$$

where $\nu > 0$ a viscosity coefficient.

In an elastic solid the *stress* $\sigma = (\sigma_{ij})$ is related to the strain of *displacement* u, In the case of small displacements/strains and linear elasticity by *Hooke's Law*:

$$\sigma = E\epsilon(u) \equiv \lambda\nabla \cdot u\delta + 2\nu\epsilon(u), \tag{132.3}$$

where λ and μ are positive *Lamé coefficients*, and $\delta = (\delta_{ij})$ with $\delta_{ij} = 1$ if $i = j$ and $= 0$ else. The coefficient λ is referred to as *Young's modulus* which for a 1d problem measures stress increase per unit strain increase.

For large discplacements/strains different strain measures are relevant: A strain measure with reference to the initial configuration of an elastic solid before loading, is given by

$$\frac{1}{2}(FF^\top - I) \tag{132.4}$$

where $F = \frac{\partial x}{\partial X}$ with $x(t, X)$ the position at time t of a particle at position X at initial time. A corresponding Hooke's Law can take the form

$$\sigma = \frac{E}{2}(FF^\top - I) \tag{132.5}$$

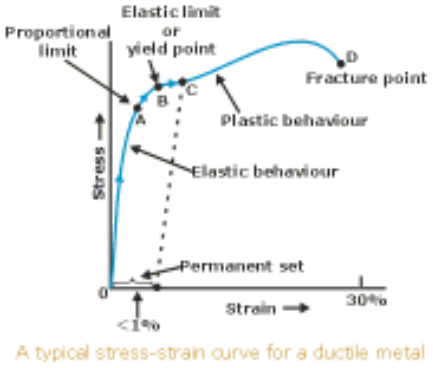


FIGURE 132.3. Stress-strain curve for elasto-plastic material.

with E acting as above. A corresponding constitutive law with the strain measure referring instead to the current configuration after loading, is given by

$$\sigma = \frac{E}{2}(I - (FF^\top)^{-1}). \tag{132.6}$$

We note that the displacement u is defined by $x(t,X) = X + u(t,X)$ or $u = x(t,X) - X$, which gives

$$F = I + \frac{\partial u}{\partial X} \tag{132.7}$$

and thus for small displacements/strains

$$\frac{1}{2}(FF^\top - I) \approx \epsilon(u), \tag{132.8}$$

since

$$FF^\top - I \approx (I + \frac{\partial u}{\partial x})(I + \frac{\partial u}{\partial x}^\top) \approx \nabla u + \nabla u^\top = 2\epsilon(u).$$

We refer to (132.5) and (132.6) as *total Lagrangian formulations* following particle trajectories $x(t,X)$ with $x(0,X) = X$, and computing $F = \frac{\partial x}{\partial X}$.

Note that if $x = QX$, where Q is an orthogonal matrix, then $F = Q$ and $FF^\top - I = QQ^\top - I = 0$, which fits with the inituitive idea that a rigid rotation of an elastic body does not generate strains and thus no stresses (if we negelect dynamic accelleration forces from the rotation, which is possible for slow rotation).

An alternative time-differentiated form of the constitutive law (132.5) can be derived as follows: The material time derivative $\frac{D\sigma}{Dt}$ (following material particles) computed using (132.5) combined with the fact that (proved below)

$$\frac{DF}{Dt} = \nabla v\, F, \tag{132.9}$$

is given by

$$\frac{D\sigma}{Dt} = \nabla v FF^\top + FF^\top \nabla v^\top = \nabla v(FF^\top - I) + (FF^\top - I)\nabla v^\top = \nabla v\sigma + \sigma \nabla v^\top.$$

We thus obtain a constitutive law connecting a certain stress rate $\dot{\sigma}$ to velocity strain $\epsilon(v)$, assuming $E = 2I$ for simplicity, of the form

$$\dot{\sigma} \equiv \frac{D\sigma}{Dt} - \nabla v\sigma - \sigma \nabla v^\top = (\nabla v + \nabla^T) = 2\epsilon(v). \tag{132.10}$$

which we refer to as an *updated Lagrangian formulation*.

The constitutive law for a solid/fluid continuum can thus be written on the form

$$\begin{aligned} &\sigma = 2\nu\epsilon(v) \quad \text{in fluid (Eulerian)} \\ &\sigma = \frac{E}{2}(FF^\top - I) \quad \text{in solid (total Lagrangian)} \\ &\dot{\sigma} = E\epsilon(v) \quad \text{in solid (updated Lagrangian),} \\ &\sigma = E\epsilon(u) \quad \text{in solid if displacements/strains are small.} \end{aligned} \tag{132.11}$$

Note that F in the total Lagrangian formulation can be computed by updating F accordig to (132.9).

We can also combine into a hybrid between solid and fluid in a constitutive equation of the form

$$\alpha_1\dot{\sigma} + \alpha_2\sigma = E\epsilon(v), \tag{132.12}$$

with the α_i nonnegative coefficients.

For a fluid contained in a fixed volume, we do not have to trace fluid particles, since what appears in the constitutive equations $\sigma = 2\nu\epsilon(v)$ is the velocity $v(x,t)$ of particles at the point x at time t, irrespective of from where they came.

For a solid which deforms, in general we have to follow material particles to record the current configuration of the body, and to compute $F = \frac{\partial x}{\partial X}$ appearing in (132.5) in a total Lagrangian formulation.

In the updated Lagrangian formulation $\dot{\sigma} = E\epsilon(v)$, the velocity v is also expressed in Eulerian coordinates, and the time stepping of σ by the definition of $\dot{\sigma}$ automatically follows the motion of material particles.

In the case of displacement/strain is small, the initial configuration can be used as reference throughout and an Eulerian formulation is possible also for a solid.

We sum up our observations concerning the formulation of constitutive laws

- In a viscous fluid stress is related to velocity strain. Eulerian description is efficient.

- In an elastic solid stress is related to displacement strain. A total or updated Lagrangian description is efficient, which for small displacements/strains reduces to Eulerian in initial configuration.

132.3 Cauchy Stresses in Eulerian Coordinates

For a solid in equilibrium, the differential equation expressing conservation of momentum is referred to as the *equilibrium equation* expressing equlibrium of internal and exterior forces. Like conservation of momemtum, the equlibrium equation can be expressed in Eulerian coordinates in terms of stresses measured in the current deformed configuration, which are referred to as *Cauchy stresses*.

We shall find that formulating the equilibrium equation in Eulerian coordinates is efficient, in particular for fluid-solid interaction.

It is also possible to express the equlibrium equation using other stress measures referring to the initial undeformed configuration, but this is more complicated.

132.4 Fluid-Solid in a Nutshell

The constitutive laws for fluids and solids and visco-elastic hybrids can be summarized as

$$\begin{aligned} \sigma &= E\epsilon(v) \quad \text{in viscous fluid} \\ \dot{\sigma} &= E\epsilon(v) \quad \text{in elastic solid} \\ \sigma + \dot{\sigma} &= E\epsilon(v) \quad \text{in visco-elastic fluid-solid,} \end{aligned} \tag{132.13}$$

noting that "only a dot" makes a viscous fluid into an elastic solid. Neat!

Extension to include effects of plasticity is also possible in this formulation.

132.5 Proof that $\frac{DF}{Dt} = \nabla v\, F$

To prove (132.9) we change order of differentiation to get

$$\frac{DF}{Dt} = \frac{\partial}{\partial t}\frac{\partial x(t,X)}{\partial X} = \frac{\partial}{\partial X}v(x(t,X),t) = \frac{\partial v}{\partial x}\frac{\partial x}{\partial X} = \nabla v F. \tag{132.14}$$

132.6 Watch

- Designing a bridge

- Twin Towers collapse
- Solid mechanics FEM simulation
- Solid mechanics FEM demo

132.7 To Think About

- Motivation of (132.4) as strain measure?
- Motivation of (132.5) as stress-strain law?
- How to measure viscosity?
- How to measure a spring constant?
- How to measure coefficients of elasticity?

132.8 Unicorn Simulations

Unicorn is a FEniCS module for fluid, structure and fluid-structure interaction based on the unified material model (132.13) combined with conservation of mass, momentum and energy in Eulerian coordinates (130.9). For orientation browse the following prototype Unicorn simulations:

- Simple Flapping Bird
- Flapping Flag
- 2d Benchmark

to which we will return in more detail below.

FIGURE 132.4. Eulerian observers tied to the shore.

FIGURE 132.5. Lagrangian observers floating freely.

133
Diffusion

> Available energy is energy which we can direct into any desired channel. Dissipated energy is energy which we cannot lay hold of and direct at pleasure, such as the energy of the confused agitation of molecules which we call heat. Now, confusion, like the correlative term order, is not a property of material things in themselves, but only in relation to the mind which perceives them. A memorandum-book does not, provided it is neatly written, appear confused to an illiterate person, or to the owner who understands it thoroughly, but to any other person able to read it appears to be inextricably confused. Similarly the notion of dissipated energy could not occur to a being who could not turn any of the energies of nature to his own account, or to one who could trace the motion of every molecule and seize it at the right moment. It is only to a being in the intermediate stage, who can lay hold of some forms of energy while others elude his grasp, that energy appears to be passing inevitably from the available to the dissipated state. (James Clerk Maxwell)

133.1 Model

Heat conduction is a *diffusion process* which is modeled by the conservation law

$$\kappa \dot{u} + \nabla \cdot q = f \tag{133.1}$$

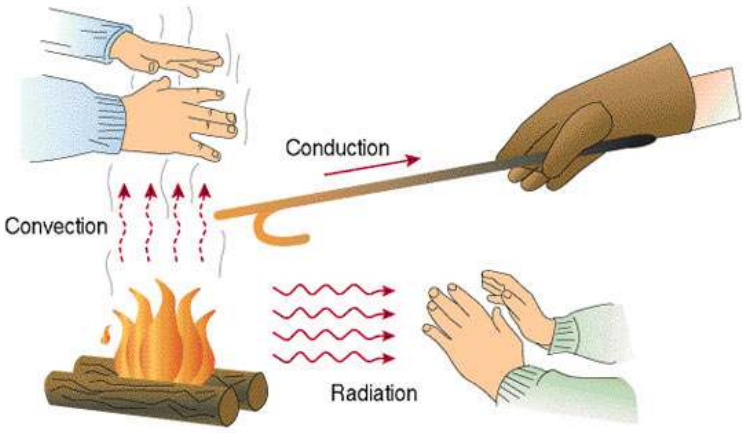


FIGURE 133.1. Heat conduction/diffusion vs convection and radiation.

where $e(x,t)$ is temperature, q is heat flow, κ is a given coefficient of *heat capacity*, and f a given heat source, which expresses conservation of heat energy. This law is complemented by a constitutive law relating the heat flow q to the temperature gradient ∇e, in the simplest case by *Fourier's Law*

$$q = -\mu \nabla e. \tag{133.2}$$

With $\kappa = \mu = 1$, we obtain the *heat equation*

$$\dot{e} - \Delta e = f \tag{133.3}$$

with the following stationary form

$$-\Delta e = f \tag{133.4}$$

which is *Poisson's equation*.

133.2 Simulations

- 2d heat equation model problem.
- Transient heat conduction.
- Heat transfer in brake disc.

133.3 Read More

- Poisson's equation: stationary diffusion

- The Power of Abstraction
- Heat equation: time-dependent diffusion

133.4 To Think About

- How to motivate Fourier's Law (232.3)?

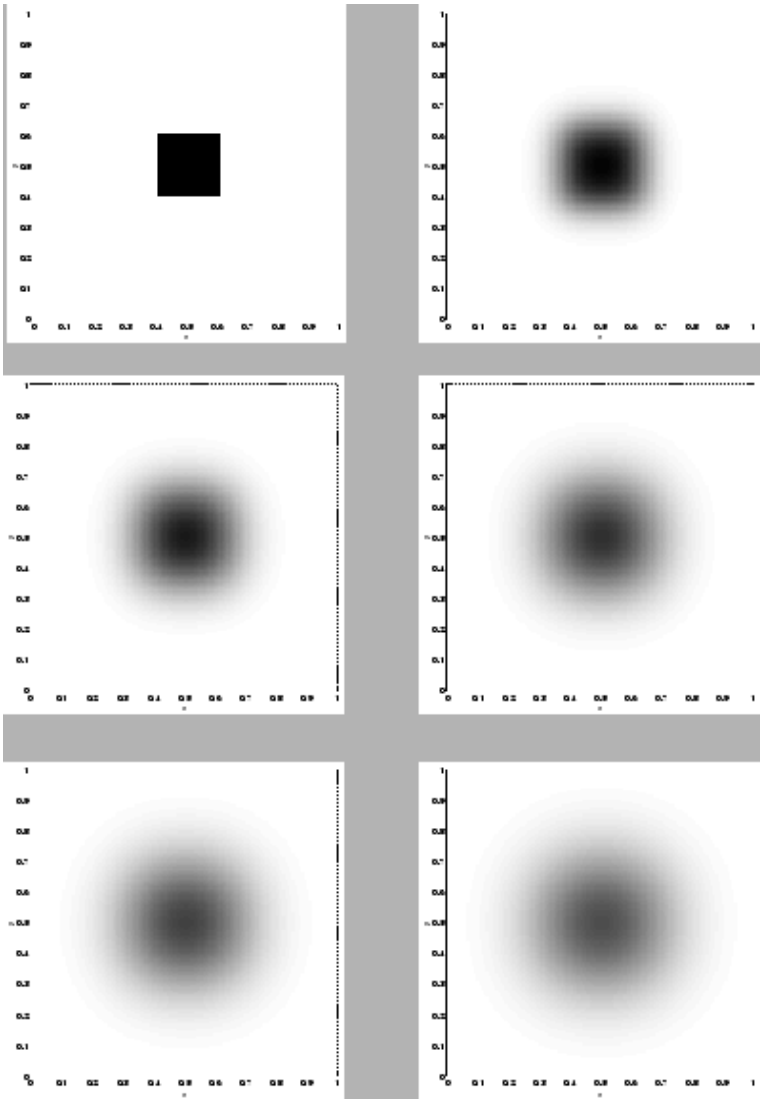


FIGURE 133.2. Diffusion tends to smooth sharp gradients and spread out.

134
Diffusion-Convection-Reaction

On the diffusion of education among the people rest the preservation and perpetuation of our free institutions. (Daniel Webster)

When they come to model Heaven And calculate the stars, how they will wield The mighty frame, how build, unbuild, contrive To save appearances, how grid the sphere With centric and eccentric scribbled o'er Cycle and epicycle, orb in orb. (John Milton in Paradise Lost)

134.1 Convection of Heat

Heat can also convect with a given velocity v, in which case conservation of heat takes the same form as conservation of mass, letting temperature e replace density ρ:

$$\dot{e} + \nabla \cdot (ve) = 0, \tag{134.1}$$

assuming unit heat capacity, which we refer to as a (linear) *convection equation* if the velocity v is given.

134.2 Convection-Diffusion of Heat

Adding diffusion, we obtaoin the *convection-diffusion equation* of the form

$$\dot{e} + \nabla \cdot (ve) - \Delta e = f, \tag{134.2}$$

where we added a source term f.

134.3 Convection-Diffusion-Reaction of Anything

If we add also reaction we get the following convection-diffusion-reaction equation for any quantity which is convecting-diffusing subject to change from reaction and a source, like a chemical substance in a fluid:

$$\dot{e} + \alpha e + \nabla \cdot (ve) - \Delta = f, \tag{134.3}$$

where α is a reaction rate coefficientl.

134.4 Read More

- Stationary Reaction-Diffusion-Convection
- Time-dependent Reaction-Diffusion-Convection

134.5 Watch

- Convection-diffusion experiment
- Ice cube experiment
- Convection-diffusion in Spanish

134.6 To Think About

- How to motivate the reaction term αu? What does the sign of α signify?

FIGURE 134.1. To make good beer requires a proper combination of diffusion, convection and reaction.

135
Compressible Euler

> Remote from human passions, remote even from the pitiful facts of nature, the generations have gradually created an ordered cosmos, where pure thought can dwell as in its natural home and where one, at least, of our nobler impulses can escape from the dreary exile of the actual world. (Bertrand Russell)

135.1 Model

The *compressible Euler equations* express conservation of mass, momentum (Newton's 2nd Law) and total energy of a compressible perfect inviscid gas/fluid in the form: Find the density ρ, velocity v, total energy $\epsilon = \rho\frac{|v|^2}{2} + e$ with e internal energy as functions of (x,t), such that

$$\begin{aligned} \dot\rho + \nabla\cdot(\rho v) &= 0, \\ \dot m + \nabla\cdot(mu + p) &= f \\ \dot\epsilon + \nabla\cdot(\epsilon v + pv) &= 0, \end{aligned} \tag{135.1}$$

combined with the constitutive law $p = (\gamma - 1)e$ with $1 < \gamma < 2$ a gas constant, for the pressure of a *perfect gas*. Here f is a given volume force.

Whoops, haven't we seen these formulas before?

135.2 To Think About

- How to motivate the gas law $p \sim e$, with $e \sim \rho T$ and T temperature?

136
Incompressible Euler

136.1 Model

For an incompressible fluid the density ρ does not change under motion, which means that the *material time derivative* $\frac{D\rho}{Dt}$, the rate of change following particle trajectories $x(t)$ satisfying $\dot{x}(t) = v(x(t),t)$ with $v(x,t)$ the velocity, vanishes. In matheamtical notation this means that

$$\frac{D\rho}{Dt} \equiv \frac{\partial}{\partial}\rho(x(t),t) = (\dot{\rho} + v\cdot\nabla\rho)(x(t),t) = 0 \tag{136.1}$$

which by conservation of mass using the Chain Rule:

$$0 = \dot{\rho} + \nabla\cdot(\rho v) = \dot{\rho} + v\cdot\nabla\rho + \rho\nabla\cdot v = \rho\nabla\cdot v, \tag{136.2}$$

that is,

$$\nabla\cdot v = 0. \tag{136.3}$$

This leads to the *incompressible Euler equations* assuming $\rho \equiv= 1$:

$$\begin{aligned} \dot{v} + v\cdot\nabla v + \nabla p &= f,\\ \nabla\cdot v &= 0, \end{aligned} \tag{136.4}$$

with the law of conservation of energy decoupled assuming (v,p) determined by (136.4).

We shall discover the the Euler equations are formal mathematical models which have to complemented with small viscous terms to make sense.

We note that mass conservation can be written

$$\frac{D\rho}{Dt} = -\rho\nabla\cdot v \tag{136.5}$$

which expresses that ρ increases where $\nabla\cdot v < 0$, which expresses that fluid particles are concentrating, as we expect. Accumulation of fluid particles increases density and spreading fljuid particles apart decreases density. Elementary, my Dear Watson.

137
Incompressible Navier-Stokes

137.1 Model

Adding Newtonian viscosity to the Euler equations changes the momentum equation to the principal form

$$\dot{v} + v \cdot \nabla v + \nabla p - \nu \Delta v = f, \tag{137.1}$$

where $\nu > 0$ is a viscosity coefficient and the Laplacian term $-\nu\Delta v$ acts like diffusion on the velocity v.

The viscous effect in Newtonian fluid is more precisely modeled as

$$-\nabla \cdot \sigma \tag{137.2}$$

where σ is the *viscous stress* given by

$$\sigma = \mu\epsilon(v) \equiv \mu\frac{1}{2}(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}). \tag{137.3}$$

Note that σ is a *shear stress* because $\sum_{i=1}^{3} \sigma_{ii} = \mu\nabla \cdot v = 0$.

This leads to the incompressible Navier-Stokes equations for a unit density Newtonian fluid with viscosity μ: Find the velocity $v(x,t)$ velocity, pressure $p(x,t)$ and viscous shear stress $\sigma(x,t)$ such that

$$\begin{aligned} \dot{v} + v \cdot \nabla v + \nabla p - \nabla \cdot \sigma &= 0, \\ \nabla \cdot v &= 0 \\ \sigma &= \mu\epsilon(v) \end{aligned} \tag{137.4}$$

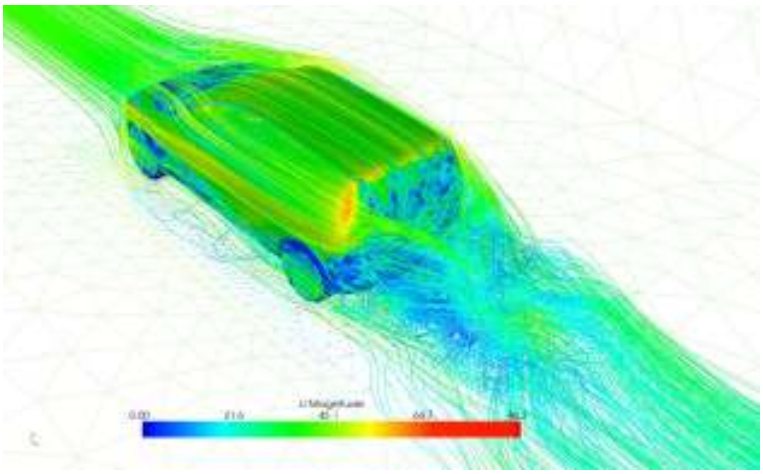


FIGURE 137.1. Turbulent incompressible flow around a Volvo by FEniCS/Unicorn.

where

$$\epsilon(v) = \frac{1}{2}(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}), \tag{137.5}$$

and $\nu > 0$ is a constant viscosity coefficient. The case with small viscosity, modeling the flow of air and water at macroscales, is of particular relevance in applications and include the phenomenon of turbulence. We shall discover that computational solution of the Navier-Stokes equations is possible, which opens to understanding and predicting turbulent flow.

137.2 Simulations

- Turbulent flow around a car by Unicorn.
- Turbulent flow 1
- Turbulent flow 2
- Turbulent boundary layer flow

137.3 To Browse

- Computational Turbulent Incompressible Flow

138
Nearly Incompressible Navier-Stokes

138.1 Model

Water is a very nearly incompressible fluid, and air at speeds much lower than the speed of sound is nearly incompressible, but no fluid is exactly incompressible. We shall discover that in computational simulation it is preferable to change the incompressibility condition $\nabla \cdot v = 0$ to the following constititutive law for the pressure:

$$\delta_1 \dot{p} - \delta_2 \Delta p = -\nabla \cdot v, \tag{138.1}$$

where δ_1 and δ_2 are small positive *regularization* parameters. This relation expresses that where div $\cdot v < 0$, that is where fluid density is increasing by accumulating fluid particles, the pressure will increase and generate a pressure gradient ∇p tending to spread fluid particles apart. Similarly, where fluid density is decreasing with $\nabla \cdot v > 0$, a counter-balancing pressure gradient will be generated. Altogether, the pressure law (138.1) will seek to maintain any initial density: Fluid is neither compressed nor stretched, just squeezed.

A common choice of the regularization parameters is $\delta_1 = 0$ and $\delta_2 = h$ where h is the mesh size. With this choice the regularized nearly incompressible Navier-Stokes equations take the form

$$\begin{aligned} \dot{v} + v \cdot \nabla v + \nabla p - \nabla \cdot \sigma &= 0, \\ -h\Delta p + \nabla \cdot v &= 0 \\ \sigma &= \mu\epsilon(v). \end{aligned} \tag{138.2}$$

138.2 Simulations

- Incompressible turbulent cavity flow
- Sloshing blood
- Navier-Stokes waterfall
- Interactive real-time Navier-Stokes

139
Compressible Navier-Stokes

139.1 Model

The comnpressible Navier-Stokes equations expand the compressible Eulr equations to include viscous forces σ, which for a Newtonian fluid connect to velocity strains $\epsilon(v)$ and velocity divergence $\nabla \cdot v$ by a linear relation of the form

$$\sigma = 2\mu\epsilon(v) + \kappa\nabla \cdot v\delta \tag{139.1}$$

where μ and κ are positive constants and $\delta = (\delta_{ij})$ Kronecker's delta with $\delta_{ij}) = 1$ if $i = j$ and $\delta_{ij} = 0$ else. The momentum equation is then modified to

$$\dot{m} + \nabla \cdot (mu + p) - \nabla \cdot \sigma = f. \tag{139.2}$$

139.2 Simulations

- Implosion with reflections by Unicorn.
- Falling at supersonic speeds
- Supersonic bullet
- High speed compressor
- Diesel engine intake flow

FIGURE 139.1. Compressible Flow.

139.3 To Browse

- Secret of Thermodynamics
- Computational Thermodynamics

140
Navier/Lagrange: Solid Mechanics

140.1 Model

The equations of solid mechanics are the conservation laws of mass, momentum and energy combined with constitutive laws for stresses in terms of strains, in total or updated Lagrangian formulations.

In Lagrangian formulations material particles (of given mass) are traced and conservation of mass is then automatic (if particles are not lost). If heat effects are not included this reduces the description to conservation of momentum (equlibrium equation) in Eulerian coordinates combined with stress-strain constitutive laws.

In Lagrangian FEM the finite element mesh moves with the material particles, and thus is distorted with the motion. Remeshing may be necessary to avoid mesh collapse.

140.2 Simulations

- Elastic circus cow by Unicorn.
- Elastic truss
- Golden Gate Bridge
- Earthquake simulation.
- WTC collapse simulation

- Commercial software: Nastran.

140.3 To Browse

- Automated computational modeling

141
Fluid-Structure Interaction

141.1 Model

A incompressible unit density solid interacting with an incompressible unit-density fluid, is described by

$$\begin{aligned} \dot{v} + v\cdot\nabla v + \nabla p - \nabla\cdot\sigma &= 0, \\ \nabla\cdot v &= 0, \\ \sigma &= 2\mu\epsilon(v) \quad \text{in fluid}, \\ \dot{\sigma} &= E\epsilon(v) \quad \text{in solid}, \end{aligned} \tag{141.1}$$

where $v(x,t)$ is the velocity of material particles at x at time t, and σ is the Cauchy stress, and σ is a stress rate.

141.2 Simulations

- Fluid-structure interaction bench-mark by Unicorn.
- Flapping prehistoric bird by Unicorn.
- Flapping plate by Unicorn.
- Flapping-wing micro air vehicle.
- Breaking wine glass.

FIGURE 141.1. Fluid-structure interaction.

141.3 To Browse

- Unified fluid-structure modeling.

142
Wave Equation

> Go ahead and faith will come to you. (d'Alembert)
>
> Souls act according to the laws of final causes, through apparitions, ends and means. Bodies act according to the laws of efficient causes or of motions. And these two kingdoms, that of efficient causes and that of final causes, are in harmony with each other. (Leibniz)
>
> Those beautiful laws of physics are a marvellous proof of an intelligent and free being against the system of absolute and brute necessity. (Leibniz)

142.1 Model

Recall from World of Games

- elastic string
- elastic net
- elasticbody

modeled by the wave equation

$$\ddot{u} - \Delta u = f \tag{142.1}$$

in dimension 1,2 and 3, with Dirichlet, Neumann or Robin boundary conditions, and initial conditions for u and $\dot{u}$.

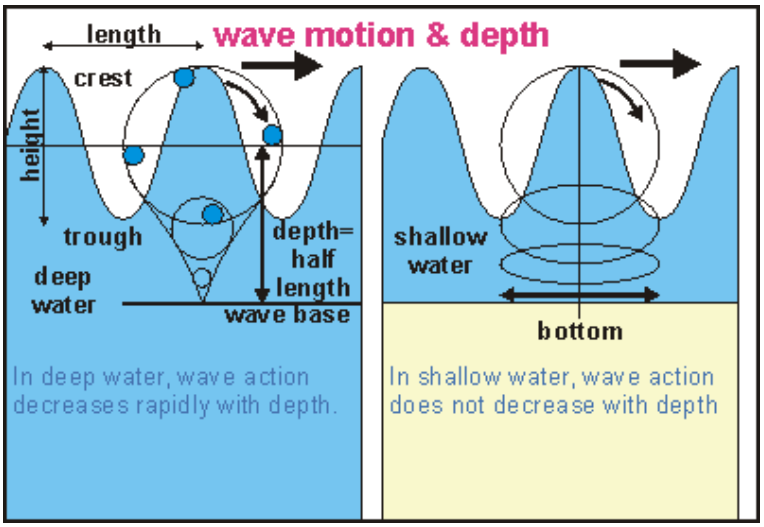


FIGURE 142.1. Particle motion in deep and shallow water waves.

142.2 Simulations

- Wave relections in a pond
- Acoustics simulations

142.3 Read More

- Wave Equation

143

Maxwell: Electromagnetics

143.1 Introduction

143.2 Faraday, Ampére, Coulomb, Gauss, Ohm

The interaction between electric and magnetic fields are described by *Maxwell's equations*:

$$\begin{cases} \dfrac{\partial B}{\partial t} + \nabla \times E = 0, \\ -\dfrac{\partial D}{\partial t} + \nabla \times H = J, \\ \nabla \cdot B = 0, \quad \nabla \cdot D = \rho, \\ B = \mu H, \quad D = \epsilon E, \quad J = \sigma E, \end{cases} \tag{143.1}$$

where E is the *electric field*, H is the *magnetic field* , D is the *electric displacement*, B is the *magnetic flux* , J is the *electric current*, ρ is the *charge*, μ is the *magnetic permeability*, ϵ is the *dielectric constant* of *electric permittivity*, and σ is the *electric conductivity*. The first equation is referred to as *Faraday's law*, the second is *Ampère's law*, $\nabla \cdot D = \rho$ is *Coulomb's law*, *Gauss law* $\nabla \cdot B = 0$ expresses the absence of "magnetic charge", and $J = \sigma E$ is *Ohm's law*. Maxwell, see Fig. 232.7, included the term $\partial D/\partial t$ for purely mathematical reasons and then used Calculus to

predict the existence of electromagnetic waves before these had been observed experimentally.

FIGURE 143.1. Maxwell (1831-1879), inventor of the mathematical theory of electromagnetism: "We can scarcely avoid the conclusion that light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena".

Typical boundary conditions include various combinations of $E \cdot n$ (perfect insulator), $E \times n$ (perfect conductor), $H \cdot n$ and $H \times n$.

Maxwell's equations describe the whole world of electromagnetic phenomena with an astounding economy of notation and accuracy of modelling. See Chart of Electromagnetic Spectrum. Our modern information society builds on electromagnetic waves. We shall now pick out a couple of Laplace equation models from Maxwell's equations by considering some basic particular cases.

143.3 To Read

For a presentation of Maxwell's equations including the crucial aspect of relative motion, e.g. between a magnetic field and electrical current, see

- Maxwell's Equations for Bodies in Motion

143.4 Watch

- Electromagnetic waves
- Faraday and Maxwell
- Feynman on electromagnetic waves
- Feynman on confusion
- Dipole antenna
- Wave equation for E and B
- How an electric motor works

143.5 Simulations

- Radiation from dipole antenna
- Electromagnetic simulation software
- More software
- Direct current electric engine principle

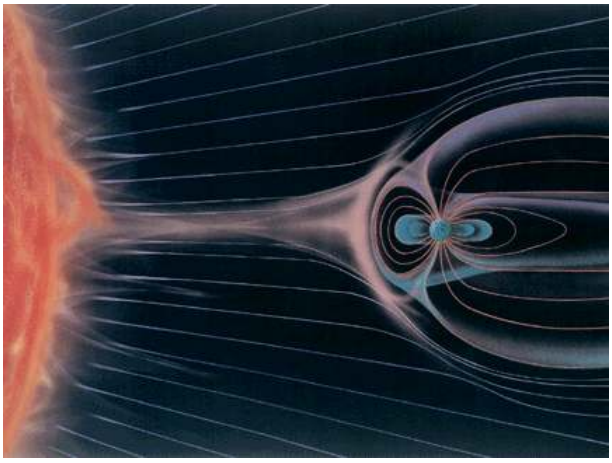


FIGURE 143.2. Solar wind of charged particles interacting with magnetic field of the Earth.

144
Schrödinger: Quantum Mechanics

144.1 Introduction

- Intro 1.
- Intro 2...
- Intro 6.
- Wave function and wave-particle duality??
- Quamtum confusion?
- BBC Illusion of reality
- The reality does not exist??

144.2 Read

- Many-Minds Quantum Mechanics

144.3 Simulations

- 1d particle in various potentials

FIGURE 144.1. Schrödinger's equation.

144.4 To Think About

- Schrödinger's cat.

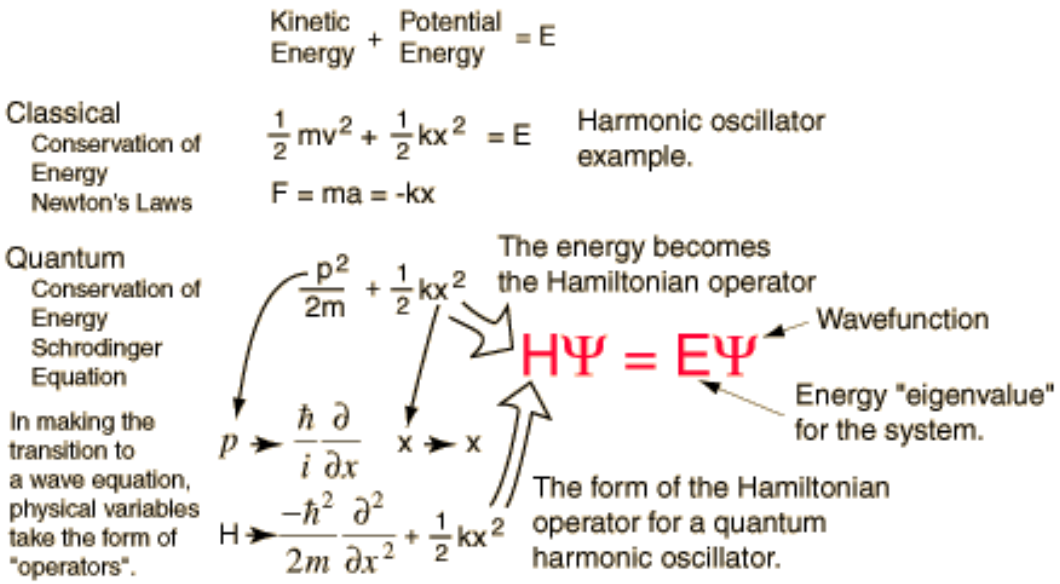


FIGURE 144.2. Schrödinger's equation vs harmonic oscillator.

145
Kohn-Sham: Quantum Chemistry

145.1 Simulations

- Carbon Nanotube 1
- Carbon nanotube 2
- Fullerne impact on nanotube

146
Black-Scholes: Options

146.1 Model

- The Formula

146.2 The Differential Equation

The Formula gives the solution to the Black-Scholes differential equation:

$$\begin{aligned} \frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0, \quad 0 < t < T, \quad S > 0, \\ V(S,T) = V_T(S) \quad S > 0, \end{aligned} \tag{146.1}$$

where $V(S,t)$ is the asset value, the variable S represents the underlying stock price, the constants σ and r represent volatility and interest rate, and $V_T(S)$ and $V(0,t)$ are given data. Note that time t is running backwards with initial data being given at $t = T$ and final time at $t = 0$. This is a 1d convection-diffusion-reaction equation which alternatively can be (quickly) solved by FEM.

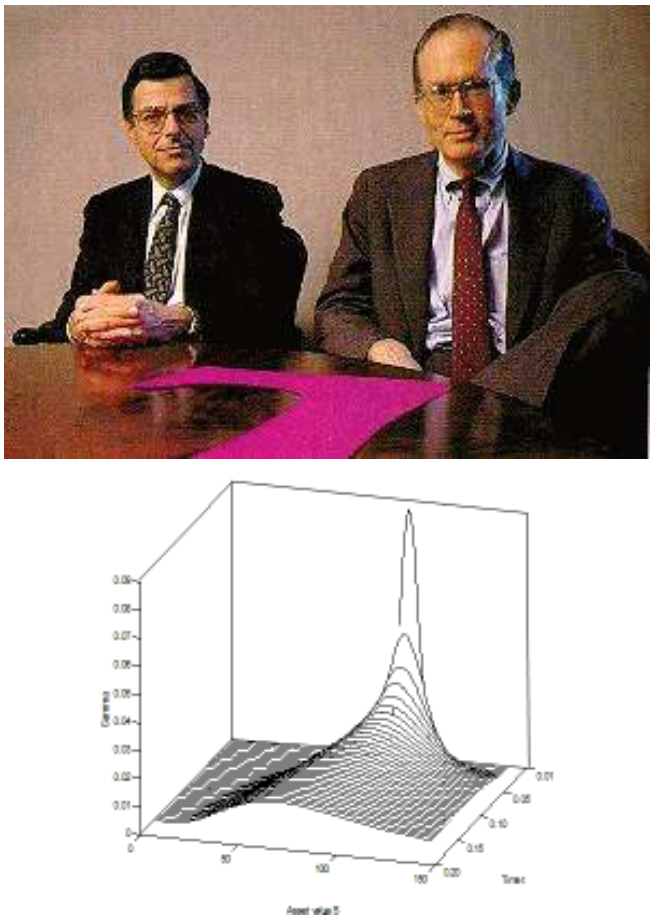


FIGURE 146.1. Myron Scholes and Fischer Black and one of their solutions (they all look alike).

147
Differential Equations Tool Bag

It seems to me that there are at least four different viewpoints– or extremes of viewpoint– that one may reasonably hold:

1. All thinking is computation; in particular, feelings of conscious awareness are evoked merely by the carrying out of appropriate computations. (Hard AI)

2. Awareness is a feature of the brain's physiological action; and whereas any physical action can be simulated computationally, computational simulation cannot by itself evoke awareness. (Soft AI)

3. Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally. (Penrose's view)

4. Awareness cannot be explained by physical, computational, or any other scientific terms.

(R. Penrose in *Shadows of the Mind*)

147.1 Introduction

We here collect basic facts about solving differential equations analytically and numerically.

147.2 The Equation $u'(x) = \lambda(x)u(x)$

The solution to the scalar initial value problem

$$u'(x) = \lambda(x)u(x) \quad \text{for } x > a, \quad u(a) = u_a,$$

where $\lambda(x)$ is a given function of x, and u_a a given initial value, is

$$u(x) = \exp(\Lambda(x))u_a = e^{\Lambda(x)}u_a,$$

where $\Lambda(x)$ is a primitive function of $\lambda(x)$ such that $\Lambda(a) = 0$. In particular, if λ is a constant, then $u(x) = \exp(\lambda x)u_a$.

147.3 The Equation $u'(x) = \lambda(x)u(x) + f(x)$

The solution the scalar initial value problem

$$u'(x) = \lambda(x)u(x) + f(x) \quad \text{for } x > a, \quad u(a) = u_a,$$

where $\lambda(x)$ and $f(x)$ are given functions of x, and u_a a given initial value, can be expressed using Duhamel's principle in the form

$$u(x) = e^{\Lambda(x)}u_a + e^{\Lambda(x)}\int_a^x e^{-\Lambda(y)}f(y)\,dy.$$

where $\Lambda(x)$ is a primitive function of $\lambda(x)$ such that $\Lambda(a) = 0$.

147.4 The Differential Equation $\sum_{k=0}^n a_k D^k u(x) = 0$

A solution to the constant coefficient differential equation

$$p(D)u(x) = \sum_{k=0}^n a_k D^k u(x) = 0, \quad \text{for } x \in I,$$

where I is an interval of real numbers, has the form

$$u(x) = \alpha_1 \exp(\lambda_1) + ... + \alpha_n \exp(\lambda_n),$$

where the α_i are arbitrary constants and the λ_i are the roots of the polynomial equation $p(\lambda) = 0$ with $p(\lambda) = \sum_{k=0}^n a_k\lambda^k$, assuming there are n distinct roots. If $p(\lambda) = 0$ has a multiple root λ_i of multiplicity r, then the solution is the sum of terms of the form form $q(x)\exp(\lambda_i x)$, where $q(x)$ is a polynomial of degree at most $r-1$. For example, if $p(D) = (D-1)^2$, then a solution of $p(D)u = 0$ has the form $u(x) = (a_0 + a_1 x)\exp(x)$.

147.5 The Damped Linear Oscillator

A solution $u(t)$ to

$$\ddot{u} + \mu\dot{u} + ku = 0, \text{for } t > 0,$$

where μ and k are constants, has the form

$$u(t) = ae^{-\frac{1}{2}(\mu+\sqrt{\mu^2-4k})t} + be^{-\frac{1}{2}(\mu-\sqrt{\mu^2-4k})t},$$

if $\mu^2 - 4k > 0$, and

$$u(t) = ae^{-\frac{1}{2}\mu t}\cos(\frac{t}{2}\sqrt{4k-\mu^2}) + be^{-\frac{1}{2}\mu t}\sin(\frac{t}{2}\sqrt{4k-\mu^2}),$$

if $\mu^2 - 4k < 0$, and

$$u(t) = (a+bt)e^{-\frac{1}{2}\mu t},$$

if $\mu^2 - 4k = 0$, where a and b are arbitrary constants.

147.6 The Matrix Exponential

The solution to the initial value problem linear system

$$u'(x) = Au(x) \quad \text{for } 0 < x \leq T,\ u(0) = u_0,$$

where A is a *constant* $d \times d$ matrix, $u_0 \in \mathbb{R}^d$, $T > 0$, is given by

$$u(x) = \exp(xA)u_0 = e^{xA}u_0.$$

If A is diagonalizable so that $A = SDS^{-1}$, where S is nonsingular and D is diagonal with diagonal elements d_i (the eigenvalues of A), then

$$\exp(xA) = S\exp(xD)S^{-1}.$$

where $\exp(xD)$ be the diagonal matrix with diagonal elements equal to $\exp(xd_i)$.

The solution to the initial value problem

$$u'(x) = Au(x) + f(x) \quad \text{for } 0 < x \leq 1,\ u(0) = u_0,$$

where $f(x)$ is a given function, is given by Duhamel's principle:

$$u(x) = \exp(xA)u_0 + \int_0^x \exp((x-y)A)f(y)\,dy.$$

147.7 Fundamental Solutions of the Laplacian

The function $\Phi(x) = \frac{1}{4\pi}\frac{1}{\|x\|}$ for $x \in \mathbb{R}^3$ satisfies the differential equation $-\Delta\Phi = \delta_0$ in $\mathbb{R}^3$, where δ_0 represents a point mass at the origin. The function $\Phi(x) = \frac{1}{2\pi}\log(\frac{1}{\|x\|})$ for $x \in \mathbb{R}^2$ satisfies the differential equation $-\Delta\Phi = \delta_0$ in $\mathbb{R}^2$, where δ_0 represents a point mass at the origin.

147.8 The wave equation in 1d

The general solution to the one-dimensional wave equation

$$\ddot{u} - u'' = 0 \quad \text{for } x, t \in \mathbb{R},$$

is given by $u(x,t) = v(x-t) + w(x+t)$ where $v, w : \mathbb{R} \to \mathbb{R}$ are arbitrary functions.

147.9 Numerical Methods for IVPs

The dG(O), the discontinuous Galerkin method with piecewise constants, for the initial value problem $\dot{u}(t) = f(u(t),t)$ for $t > 0$, $u(0) = u^0$, with $f : \mathbb{R}^{d+1} \to \mathbb{R}^d$, takes the form

$$U^n = U^{n-1} + \int_{t_{n-1}}^{t_n} f(U^n, t)\, dt, \quad n = 1, 2, ...,$$

where $U(t)$ is piecewise constant on a partition $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} < \cdots$, with $U(t) = U^n$ for $t \in (t_{n-1}, t_n]$ and $U(0) = u^0$. With right-end point quadrature we obtain the implicit backward-Euler method:

$$U^n = U^{n-1} + k_n f(U^n, t_n)\, dt, \quad n = 1, 2, ...,$$

where $k_n = t_n - t_{n-1}$. The explicit forward Euler method reads:

$$U^n = U^{n-1} + k_n f(U^{n-1}, t_{n-1})\, dt, \quad n = 1, 2, ...,$$

The cG(1), the continuous Galerkin method with continuous piecewise linear functions, takes the form

$$U(t_n) = U(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(U(t), t)\, dt, \quad n = 1, 2, ...,$$

where $U(t)$ is continuous piecewise linear with nodal values $U(t_n) \in \mathbb{R}^d$ and $U(0) = u^0$.

147.10 cg(1) for Convection-Diffusion-Reaction

The cG(1) finite element method for the scalar convection-diffusion-reaction problem

$$-\nabla \cdot (a\nabla u) + \nabla \cdot (ub) + cu = f \quad \text{in } \Omega,$$
$$a\frac{\partial u}{\partial n} + \kappa u = g \quad \text{on } \Gamma,$$

with Robin boundary conditions, where f and g are given data, and $a > 0$, b, c and $\kappa \geq 0$ are given coefficients, and Ω is a given domain in $\mathbb{R}^2$ with boundary Γ, takes the form: Find $U \in V_h$ such that

$$\int_\Omega a\nabla U \cdot \nabla v \, dx + \int_\Omega \nabla \cdot (ub)v \, dx + \int_\Omega cuv \, dx + \int_\Gamma \kappa uv \, ds$$
$$= \int_\Omega fv \, dx + \int_\Gamma gv \, ds,$$

where V_h is a space of continuous piecewise linear functions on a triangulation of Ω with no restriction on the nodal values on the boundary.

147.11 Svensson's Formula for Laplace's Equation

$$U_{i,j} = \frac{1}{4}(U_{i-1,j} + U_{i+1,j} + U_{i,j-1} + U_{i,j+1}), \quad \text{for } i, j \in \mathbb{Z},$$

where $U_{i,j}$ approximates $u(ih, jh)$ with $h > 0$ and $u : \mathbb{R}^2 \to \mathbb{R}$ solves $\Delta u = 0$.

147.12 Optimal Control

The stationary equations for the saddle point problem $\min_{v,q} \max_\mu L(v, q, \mu)$, with

$$L(v, q, \mu) = \frac{1}{2}\|v - \hat{u}\|^2 + \frac{\alpha}{2}\|q\|^2 + (\dot{v} + f(v, q), \mu)$$

with $(v, w) = \int_0^T v \cdot w \, dt$ and (v, q, μ) varying freely (with $v(0) = u^0$ and $\mu(T) = 0$), take the form:

$$\dot{u} + f(u, p) = 0 \quad \text{on } [0, T], \quad u(0) = u^0, \tag{147.1}$$
$$-\dot{\lambda} + f'_v(u, p)^\top \lambda = \hat{u} - u \quad \text{on } [0, T], \quad \lambda(T) = 0. \tag{147.2}$$
$$f'_q(u, p)^\top \lambda + \alpha p = 0 \quad \text{on } [0, T], \tag{147.3}$$

where $\top$ denotes transpose. Here (237.1) is the state equation, (237.2) is the *costate equation*, and (237.3) is the *feed back control* coupling the *optimal control* p to the *costate* λ.

148
Applications Tool Bag

148.1 Introduction

In this section we collect the basic models of engineering and science expressed as differential equations. For specification of boundary and initial values we refer to the text.

148.2 Malthus' Population Model

$$\dot{u} = \lambda u - \mu u,$$

where $u(t)$ is the population at time t, $\lambda \geq 0$ the birth rate and $\mu \geq 0$ the death rate.

148.3 The Logistics Equation

$$\dot{u} = u(1 - u)$$

148.4 Mass-Spring-Dashpot System

$$m\ddot{u} + \mu\dot{u} + ku = f, \quad \text{((force balance)},$$

where $u(t)$ is the displacement, m is the mass, μ the viscosity, and k the spring constant.

148.5 LCR-Circuit

$$L\ddot{u} + R\dot{u} + \frac{u}{C} = f, \quad \text{((balance of potentials),}$$

where $u(t)$ is a primitive function of the current, L is the inductance, R the resistance, C the capacitance, and f a potential.

148.6 Laplace's Equation for Gravitation

$$-\Delta u = \rho,$$

where $u : \mathbb{R}^3 \to \mathbb{R}$ is the gravitational potential and $\rho(x)$ the mass density.

148.7 The Heat Equation

$$\dot{u} - \nabla \cdot q = f, \quad q = k\nabla u \quad \text{(heat balance and Fourier's law)}$$

where $u(x,t)$ is a temperature, $q(x,t)$ a heat flux, $k(x,t) > 0$ a conduction coefficient and $f(x,t)$ a heat source. If $k = 1$, then we get the heat equation: $\dot{u} - \Delta u = f$.

148.8 The Wave Equation

$$\ddot{u} - \Delta u = f.$$

148.9 Convection-Diffusion-Reaction

$$\dot{u} + \nabla \cdot (\beta u) + \alpha u - \nabla \cdot (\epsilon \nabla u) = f.$$

where $u(x,t)$ a concentration, $\beta(x,t)$ is a convection velocity, $\alpha(x,t)$ a reaction coefficient, $\epsilon(x,t)$ a diffusion coefficient, and $f(x,t)$ a production rate.
.

148.10 Maxwell's Equations

$$\begin{cases} \dfrac{\partial B}{\partial t} + \nabla \times E = 0, & \text{(Faraday's law)} \\ -\dfrac{\partial D}{\partial t} + \nabla \times H = J, & \text{(Ampère's law)} \\ \nabla \cdot B = 0, \quad \nabla \cdot D = \rho, & \text{(Gauss' and Coulomb's laws} \\ B = \mu H, \quad D = \epsilon E, \quad J = \sigma E, & \text{(constitutive laws and Ohm's law)} \end{cases}$$

where E is the electric field, H is the magnetic field , D is the electric displacement, B is the em magnetic flux , J is the electric current, ρ is the charge, μ is the magnetic permeability , ϵ is the dielectric constant, and σ is the electric conductivity.

148.11 The Incompressible Navier-Stokes Equations

$$\frac{\partial u}{\partial t} + (u \cdot \nabla)u + \nabla p - \nu \Delta u = f, \quad \nabla \cdot u = 0,$$

where $u(x,t)$ is the fluid velocity, $p(x,t)$ the pressure, $f(x,t)$ a given force and $\nu > 0$ a constant viscosity.

148.12 Schrödinger's Equation

$$i\frac{\partial \varphi}{\partial t} = \big(-\frac{1}{2}\sum_j \Delta_j + V(r_1, ..., r_N)\big)\varphi(r_1, ..., r_N), \quad r_j \in \mathbb{R}^3.$$

$$i\frac{\partial \varphi}{\partial t} = \big(-\frac{1}{2}\Delta + \frac{1}{|x|}\big)\varphi(x), \quad x \in \mathbb{R}^3, \quad \text{(Hydrogen atom)}.$$

Part IX

World of Finite Elements

149
The Finite Element Method

> Mathematics as an expression of the human mind reflects the active will, the contemplative reason, and the desire for aesthetic perfection. Its basic elements are logic and intuition, analysis and construction, generality and individuality. (Richard Courant)

149.1 FEM as Discretization of PDEs

The *Finite Element Method* or *FEM* is a general mathematical methodology for *discretizing* Partial Differential Equations or PDEs into systems of algebraic equations which can be fed into a computer and solved by numerical linear algebra.

FEM thus transforms PDEs with infinitely small space and time steps, into systems of algebraic equations with finite space and time steps, which can be solved by a computer.

More precisely, FEM is a general methodoloy for generating discrete approximations of derivates on general meshes in space and time. FEM is a central tool in automation of computational mathematical modeling being realized in the FEniCS Project closely connected to BodyandSoul.

FEM discretizes with respect to both space and time, based on

- (a) variational formulation of IBVPs,
- (b) piecewise polynomial approximation.

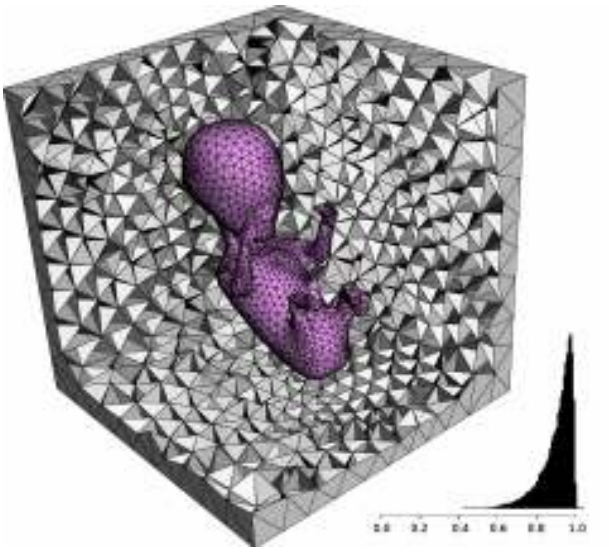


FIGURE 149.1. The birth of FEM.

Recall that Lagrangian formulations naturally can be viewed in discrete form as mass-spring systems, while Eulerian formulations most naturally start with continuous PDE formulations which are then discretized by FEM.

149.2 More Detailed Presentations

We present below a shortcut to FEM focussing a skeleton with a minimum of mathematical notation. Once you have chartered this introduction you are ready to get a fuller picture in

- Piecewise polynomials 1d
- FEM in 1d or FEM for twopoint BVP
- Piecewise polynomials 2d/3d
- FEM in 2d and 3d and FEm for Poisson
- Abstract FEM
- FEM for scalar IVP and FEM for system IVP
- FEM for heat equation
- FEM for wave equation

- FEM for Convection-Diffusion-Reaction
- FEM for Navier-Stokes equations.
- Computational Turbulent Incompressible Flow
- Computational Compressible Flow.
- Computational Thermodynamics

The BodyandSoul Sessions E-F gives an introduction to the mathematics and programming of FEM.

149.3 Why FEM Modeling is Efficient

We have seen that many phenomena of the real world can be efficiently modeled as differential equations, and with FEM as an automatic discretizer of differential equations, we get an efficient tool of generating discrete systems of equations which can be solved by computers, and act as simulators of complex phenomena.

PDEs allow an efficient formulation of basic laws of physics, because complex discrete matter is replaced by a fictional simple continuum. But there is a hook: The "simple continuum" is closed to inspection by analytical solution and thus the continuum has to be discretized to allow computational solution and FEM is a very flexible and efficient tool for discretization. Altogether PDEs discretized by FEM makes it possible for you to uncover secrets of science and technology.

We first consider FEM discretization of PDEs in space, leading to systems of time-dependent ODEs (Ordinary Differential Equations), and then FEM discretization in time, leading to formally implicit time stepping methods with algebraic equations for the computer to solve. Altogether, you will find FEM to be general method for discretizing PDEs in space-time on general adaptive meshes into algebraic equations solvable by computers.

149.4 Connection to Particle-Spring Models

We recall that the 1d wave equation $\ddot{u} = u''$ came out from a discrete particle-spring system modeled by

$$\ddot{u}_i(t) = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}, \tag{149.1}$$

with $u_i(t) = u(ih, t)$ (using here subindices) and

$$u'' = \frac{d^2u}{dx^2} \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}, \tag{149.2}$$

resulting from repeated differentiation with

$$u' = \frac{du}{dx} \approx \frac{u_{i+1} - u_i}{h}, \tag{149.3}$$

where h is the mesh size.

We shall now recover these basic derivative approximations, also referred to as *difference quotients*, using simple finite elements (piecewise linear functions) on regular (equally-spaced) meshes.

In general, a FEM discretization can be viewed as a form of particle-spring model, where particle masses and spring constants are determined by FEM from the coefficients of the underlying differential equation. FEM is thus a general method for discretizing PDEs which can be automated as shown in the FEniCS Project closely connected to the authors.

A particle-spring model requires input of particle masses and spring constants, and to do this by hand for a system of many particles and springs is impossible. FEM automates this procedure.

149.5 Watch FEM and Get Inspired

- Mesh refinement in FEniCS Unicornwith Corresponding supersonic flow
- Nastran demo
- Heart: Healthy vs Sick
- The Origin of Drag
- CTLabs TV Channel

149.6 Leibniz Solution of the Brachistochrone Problem

The brachistochrone problem was one of the earliest problems posed to test the potential of the new Calculus: Find the shape of the curve down which a bead sliding from rest and accelerated by gravity will slip (without friction) from one point to another in the least time. Watch an experiment and a discussion.

The term derives from the Greek (brachistos) "the shortest" and (chronos) "time, delay." Leibniz easily solved the problem using an argument based on piecewise linear approximation, which can be viewed as an early application of FEM, cf. Fig 149.6.

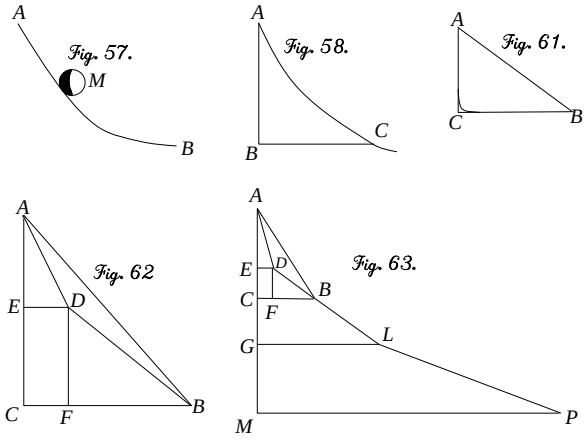


FIGURE 149.2. The first(?) application FEM by Leibniz: The Brachistochrone.

FIGURE 149.3. Books about FEM.

FIGURE 149.4. Richard Courant inventing FEM in 1943.

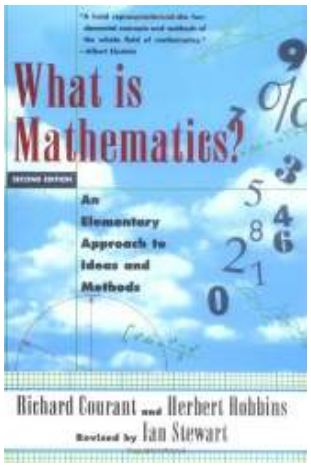


FIGURE 149.5. Richard Courant's version of BodyandSoul.

150
FEM Wave: $\ddot{u} - u'' = f$

> Science is a differential equation. Religion is a boundary condition. (Alan Turing)
>
> We are not interested in the fact that the brain has the consistency of cold porridge. (Alan Turing)

We now proceed to recover mass-spring models by FEM discretizations of PDEs and start with the wave equation for an elastic string which we met in World of Games: Find $u(x,t)$ such that

$$\begin{aligned} &\ddot{u} - u'' = f \quad \text{for } x \in (0,1),\ t > 0, \\ &u(0,t) = u(1,t) = 0 \quad \text{for } t > 0, \\ &u(x,0) = u^0(x), \quad \dot{u}(x,0) = \dot{u}^0(x) \quad \text{for } x \in (0,1), \end{aligned} \tag{150.1}$$

where $f(x,t)$ is a given function and $u^0(x)$ and $\dot{u}^0(x)$ are given initial value functions. Here $u(x,t)$ denotes the transversal deflection at location x at time t of an elastic string of unit tension covering the interval $[0,1]$ in its reference configuration and being fixed at its ends, and $f(x,t)$ is a given transversal force. One may think of a horsiontal string with $u(x,t)$ being its vertical displacement from straight reference configuration and $f(x,t)$ a vertical force.

150.1 Experience Vibrating Strings

- Piano string harmonics and chords

- String vibration
- Traveling waves as sums of stationary waves
- Wave propagation
- Slow motion: Snare drum.

150.2 Linear Combination of Tent-Functions

We seek a solution $u(x,t)$ as a *linear combination*

$$u(x,t) = \sum_{j=1}^{J} u_j(t)\varphi_j(x), \tag{150.2}$$

of J given *basis functions* $\varphi_1(x), ..., \varphi_J(x)$ depending on x, with unknown coefficients $u_1(t), ..., u_J(t)$, depending on t.

The basis functions $\varphi_j(x)$ are chosen as the following *continuous piecewise linear functions* defined on $[0,1]$ by

$$\varphi_i(jh) = 1 \quad \text{if } j = i, \quad \varphi_i(jh) = 0 \quad \text{else}, \quad i,j = 1, ..., J, \tag{150.3}$$

where $h = \frac{1}{J+1}$ is the mesh size of a mesh covering $[0,1]$ with mesh points or *nodes* jh, $j = 0, ..., J+1$.

Each function $\varphi_j(x)$ has the shape of a "tent" spanned by a pole of unit length at $x = jh$ and tied down to zero at the neighboring points $(j-1)h$ and $(j+1)h$, and is thus referred to as a *tent function* or a *hat function* as depicted in Fig. 215.5.

The functions $\varphi_1, ..., \varphi_J$ are basis functions in the sense that an arbitrary continuous piecewise linear function $v(x)$ on the given mesh, satisfying the boundary conditions $v(0) = v(1) = 0$, can be uniquely expressed as a linear combination of the basis functions:

$$v(x) = \sum_{j=1}^{J} v_j\varphi_j(x), \quad \text{with } v_j = v(jh). \tag{150.4}$$

In the representation (150.2) the coefficient functions thus are uniquely given by the *nodal values* $u_j(t) = u(jh,t)$.

We may compare with a Fourier series expansion with the basis functions being trigonometric functions, e.g. $\sin(jx)$, $j = 1, 2, ..., J$, of the form

$$u(t,x) = \sum_{j=1}^{J} u_j(t)\sin(j\pi x), \quad u_j(t) = 2\int_0^1 u(x,t)\sin(j\pi x)\,dx, \tag{150.5}$$

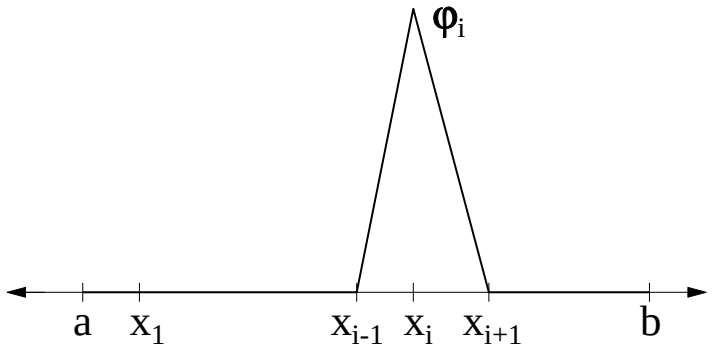


FIGURE 150.1. Basis function as tent or hat function.

where as above each term $u_j(t)\sin(jx)$ is written as a product of one functions depending on t only and one function depending on x only, thus using what is called *separation of variables*. FEM is based on piecewise polynomial (linear) approximation instead of trigonometric functions used by Fourier. Modern image or information processing use a blend of Fourier and finite element functions in the form of wavelets including in particular a Mexican hat wavelet.

Fourier's great idea was to express a general function as a sum of trigonometric functions with global support, and one of the ingredients of FEM is to write a general function as a sum of piecewise polynomial functions with local support.

We note that each tent function is non-zero on at most two subintervals; a whole tent on two and a half tent on one. In other words the *support* of each basis function consists of at most two intervals, and thus the finite element basis functions have *local support*, to be compared with trigonometric basis functions of Fourier series with *global support* like the trigonometric function $\sin(j\pi x)$ on the interval $[0,1]$). Since finite element basis functions have local support they only interact locally, which we shall see is a major advantage. Fourier series basis fuhctions compensate the global support be being orthogonal (in a certain sense), while wavelet basis functions combine local support with orthogonality.

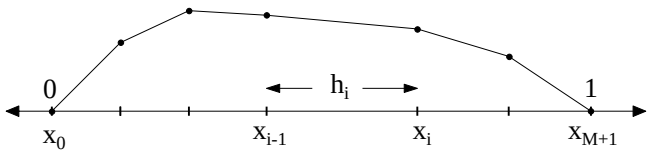


FIGURE 150.2. A continuous piecewise linear function in V_h.

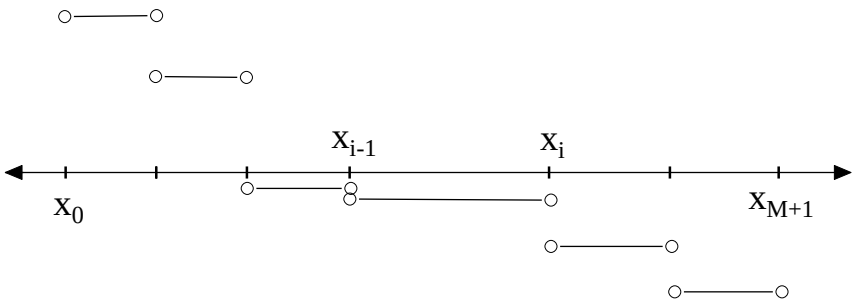


FIGURE 150.3. The derivative of the continuous piecewise linear function in Fig. 150.2.

150.3 FEM as Galerkin's Method

We now seek to determine the coefficient functions $u_1(t), ..., u_J(t)$, using *Galerkin's Method* by inserting the Ansatz (150.2) into the differential equation $\ddot{u} - u'' = f$, then multiplying by $\varphi_i(x)$ for $i = 1, ..., J$, and integrating with respect to x, to get moving derivatives inside the sum:

$$\begin{aligned}&\int_0^1 (\sum_{i=1}^J \ddot{u}_j(t)\varphi_j(x))\varphi_i(x)\,dx + \int_1^1 (\sum_{i=0}^J u_j(t)\varphi_j'(x))\varphi_i'(x)\,dx \\ &= \int_0^1 f(x,t)\varphi_i(x)\,dx, \quad i = 1, ..., J.\end{aligned} \tag{150.6}$$

We here also used integration by parts, recalling the boundary condition $\varphi_i(0) = \varphi_i(1) = 0$, to replace

$$-\int_0^1 u''(x)\varphi_i(x)\,dx \quad \text{by} \quad \int_0^1 u'(x)\varphi_i'(x)\,dx, \tag{150.7}$$

motivated by the fact that $u(x,t)$ as a continuous piecewise linear function is only differentiable once, with a piecewise constant derivative. Formally, we thus distribute the second derivative in $-\int u''\varphi_i dx$ carried by u alone, into $\int u'\varphi_i' dx$ as first derivatives carried by both u and φ_i in partnership.

Using next linearity to bring the summation outside the integration, we obtain for $i = 1, .., J$,

$$\sum_{i=1}^J \int_0^1 \varphi_i(x)\varphi_j(x)\,dx\,\ddot{u}_j(t) + \sum_{j=1}^J \int_0^1 \varphi_i'(x)\varphi_j'(x)\,dx\,u_j(t) = \int_0^1 f(x,t)\varphi_i(x)\,dx, \tag{150.8}$$

which is system of ODEs in the coefficient vector function $u = (u_1, u_2, ..., u_J)$ of the form: Find $u(t)$ such that

$$\begin{aligned} M\ddot{u}(t) + Au(t) &= b(t) \quad \text{for } t > 0, \\ u(0) &= u^0, \end{aligned} \tag{150.9}$$

where $M = (m_{ij})$ is a *mass matrix*, $A = (a_{ij})$ a *stiffness matrix* and $b(t) = (b_i(t)$ is a *load vector* with coefficients given by

$$\begin{aligned} m_{ij} &= \int_0^1 \varphi_i(x)\varphi_j(x)\,dx \quad a_{ij} = \int_0^1 \varphi_i'(x)\varphi_j'(x)\,dx, \\ b_i(t) &= \int_0^1 f(x,t)\varphi_i(x)\,dx, \quad i,j = 1, ..., J. \end{aligned} \tag{150.10}$$

Direct analytical evaluation of the integrals with piecewise polynomial

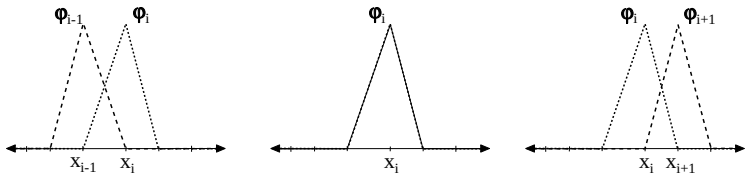


FIGURE 150.4. Three possibilities to obtain a non-zero element in the stiffness matrix.

integrands, gives

$$\begin{aligned} m_{ii} &= \frac{2h}{3},\ m_{i\,i+1} = m_{i\,i-1} = \frac{h}{6},\ m_{ij} = 0 \text{ else},\ i,j = 1, ..., J, \\ a_{ii} &= \frac{2}{h},\ a_{i\,i+1} = -\frac{1}{h},\ a_{i\,i-1} = -\frac{1}{h},\ a_{ij} = 0 \text{ else},\ i = 1, ..., J. \end{aligned} \tag{150.11}$$

We thus obtain the following system of ODEs: Find $u(t) = (u_1(t), ..., u_J(t))$ such that for $t > 0$:

$$\frac{h}{6}\ddot{u}_{i-1} + \frac{2h}{3}\ddot{u}_i + \frac{h}{6}\ddot{u}_{i+1} - \frac{u_{i+1} - 2u_i + u_{i-1}}{h} = b_i(t), \quad i = 1, ..., J, \tag{150.12}$$

with $u(0)$ and $\dot{u}(0)$ given by initial data u^0 and $\dot{u}^0$.

If we *lump* the mass matrix moving the off-diagonal coefficients $\frac{1}{6}$ to the diagonal, then we obtain the system

$$h\ddot{u}_i - \frac{u_{i+1} - 2u_i + u_{i-1}}{h} = b_i, \quad \text{for } i = 1, ..., J, \tag{150.13}$$

or dividing by h

$$\ddot{u}_i - \frac{u_{i+1} - u_i + u_{i-1}}{h^2} = \frac{b_i}{h} \approx f(ih, t), \quad \text{for } i = 1, ..., J, \tag{150.14}$$

which (as announced) coincides with the previous particle-spring model. Recall that we can imagine

$$u''(ih) \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}, \tag{150.15}$$

and thus (150.14) can be seen as a discrete analog of $\ddot{u} - u'' = f$. Note also that $\frac{b_i(t)}{h} = \frac{1}{h}\int f(x,t)\varphi_i(x)\,dx$ is a local average of $f(x,t)$ around $x = ih$, because the area under a hat equals h,

We see that the matrices M and A are *sparse* in the sense that all elements outside a diagonal band (of width three elements) are zero, as a consequence of the local support of the basis functions. Matrix-vector multiplication is fast for sparse matrices, which means that efficient computational solution of the discrete system $M\ddot{u} + Au = b$ or its lumped analog, is possible.

M and A are both symmetric and positive definite tri-diagonal matrices. The mass matrix M with positive coefficients can be seen as an approximate (scaled) identity matrix.

We can solve the discrete equation (150.14 by time-stepping e.g. as follows:

$$u_i^{n+1} = 2u_i^n - u_i^{n-1} + \frac{k^2}{h^2}(u_{i+1}^n - 2u_i^n + u_{i-1}^n), \quad i = 1, ..., J, \quad n = 1, 2, ..., \tag{150.16}$$

where k is a time step, and u^0 and u^1 are given by initial data.

150.4 Damped Wave: $\ddot{u} + \dot{u} - u'' = f$

A damped elastic string can be modeled by: Find $u(x,t)$ such that

$$\begin{aligned} \rho\ddot{u} + \mu\dot{u} - u'' &= f \quad \text{for } x \in (0,1),\ t > 0, \\ u(0,t) &= u(1,t) = 0 \quad \text{for } t > 0, \\ u(x,0) &= u^0(x), \quad \dot{u}(x,0) = \dot{u}^0(x) \quad \text{for } x \in (0,1), \end{aligned} \tag{150.17}$$

where the new term $\dot{u}$ models viscous damping (damping proportional to velocity) with viscosity coefficient $\mu > 0$, and we also inserted the coefficient $\rho > 0$ representing mass density. The corresponding FEM-model takes the form

$$\rho M\ddot{u}(t) + \mu M\dot{u} + Au(t) = b(t) \quad \text{for } t > 0, \quad u(0) = u^0,\ \dot{u}(0) = \dot{u}^0. \tag{150.18}$$

We note the limit case with $\rho = 0$ and $\mu = 1$:

$$M\dot{u} + Au(t) = b(t) \quad \text{for } t > 0, \quad u(0) = u^0. \tag{150.19}$$

We shall return to this model below with a different interpretation as a discrete heat equation.

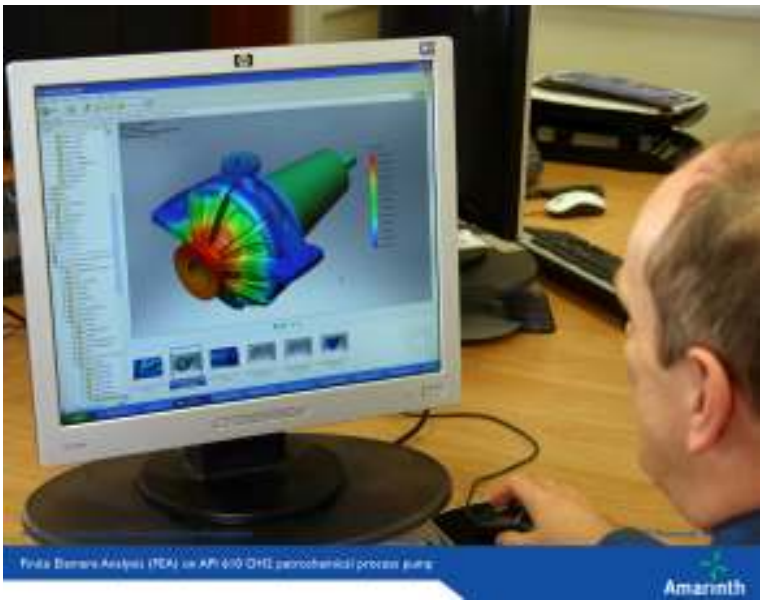


FIGURE 150.5. Exploring FEM engineer inspecting a FEM simulation.

150.5 Stationary Solution

In the stationary case with b independent of time and $\ddot{u} = \dot{u} = 0$, the discrete model is a linear system of equations:

$$Au = b \tag{150.20}$$

which thus models the deflection of an elastic string under a static load, as illustrated in Fig. 151. We now turn to a study of this model.

151
FEM Elasticity or Diffusion: $-u'' = f$

> Mathematics compares the most diverse phenomena and discovers the secret analogies that unite them. (Joseph Fourier)

We consider an stationary elastic string subject to forcing $f(x)$ modeled by the BVP: Find $u : [0,1] \to \mathbb{R}$ such that

$$\begin{aligned} -u''(x) &= f(x) \quad \text{for } x \in (0,1), \\ u(0t) &= u(1) = 0, \end{aligned} \tag{151.1}$$

with corresponding FEM discretization::

$$Au = b, \tag{151.2}$$

where $u = (u_1, ..., u_J)$ and $A = (a_{ij})$ and $b = (b_i)$ with

$$a_{ij} = \int_0^1 \varphi_i'(x)\varphi_j'(x)\,dx, \quad b_i = \int_0^1 f(x)\varphi_i(x)dx, \tag{151.3}$$

and the φ_i are the tent functions introduced above. The matrix A is symmetric and *positive definite* in the sense that

$$Au \cdot u = \int_0^1 (u')^2 dx > 0, \quad \text{if } u \neq 0, \tag{151.4}$$

where we associate with to the vector $u \in \mathbb{R}^J$ the function $u : [0,1] \to \mathbb{R}$

$$u(x) = \sum_{j=1}^{J} u_j \varphi_j(x), \tag{151.5}$$

with the result that as above $u_j = u(jh)$. We thus denote by u both the vector $u = (u_i)$ and the function $u(x)$ just defined.

We see that (151.4) holds by observing that

$$\begin{aligned} Au \cdot u &= \sum_{i,j=1}^{J} a_{ij} u_j u_i = \sum_{i,j=1}^{J} u_j \int_0^1 \varphi_j' \varphi_i' \, dx u_i \\ &= \int_0^1 (\sum_j u_j \varphi_j)' (\sum_i u_i \varphi_i)' \, dx = \int_0^1 u' u' \, dx = \int_0^1 (u')^2 \, dx \end{aligned} \tag{151.6}$$

Since A is symmetric and positive definite, the linear system of equations $Au = b$ has a unique solution u for each given right hand side b.

The BVP (151.1) also describes diffusion of heat in a heat conduction rod, with $u(x)$ representing temperature and $f(x)$ the intensity of a heat source. More generally, (151.1) models diffusion of some substance of concentration $u(x)$ with $f(x)$ a source. Mathematically (151.1) is a the basic example of an *elliptic* differential equation, with the wave equation being *hyperbolic* and the heat equation *parabolic*.

151.1 Read More

- Two-point BVP

FIGURE 151.1. Elastic string loaded by clothes.

152
FEM Error: $-u'' = f$

Read Euler: He is our master in everything. (Laplace)

152.1 The Beauty of FEM

We now seek to estimate the discretization error in (151.2) as as a model of a stationary elastic string (151.1). Let us divide each interval of length h into two intervals of length $\frac{h}{2}$ and let the corresponding FEM solution be denoted by $\bar{u}$. Can we estimate the difference $u - \bar{u}$? We have using the FEM equations defining u and $\bar{u}$

$$\int_0^1 (u - \bar{u})'\varphi_i' \, dx = \int_0^1 f\varphi_i \, dx - \int_0^1 f\varphi_i \, dx, \quad i = 1, ..., J, \tag{152.1}$$

since $\varphi_i(x)$ is piecewise linear also on the finer subdivision. Thus

$$\int_0^1 (u - \bar{u})'(u - \bar{u})' \, dx = \int_0^1 (u - \bar{u})'(\bar{u} - \hat{u})' \, dx, \tag{152.2}$$

where

$$\hat{u}(x) = \sum_{j=1}^{J} \bar{u}(jh)\varphi_j(x). \tag{152.3}$$

can be chosen to take on the values of $\bar{u}$ at $x_j = jh$. In other words, $\hat{u}$ is an *interpolant* of $\bar{u}$ (on a mesh of mesh size h). Using Cauchy's inequality

to bound the right-hand side of (152.2), we obtain

$$\|u' - \bar{u}'\| \leq \|\hat{u}' - \bar{u}'\|, \tag{152.4}$$

where

$$\|w\| = (\int_0^1 w(x)^2 dx)^{\frac{1}{2}}. \tag{152.5}$$

We now seek to bound the difference $\bar{u}' - \hat{u}'$ in terms of the mesh length h. On each subinterval of length h in the mesh underlying $\hat{u}$, the *interpolant* $\hat{u}$ takes on the same value at the endpoints as the finer-mesh $\bar{u}$, which has a "kink" in the middle of the interval equal to

$$\frac{\bar{u}(ih + \frac{h}{2}) - \bar{u}(ih)}{\frac{h}{2}} - \frac{\bar{u}(ih) - \bar{u}(ih - \frac{h}{2})}{\frac{h}{2}}$$
$$= \frac{2}{h}(\bar{u}(ih + \frac{h}{2}) - 2\bar{u}(ih) + \bar{u}(ih - \frac{h}{2})).$$

The difference in slope $\bar{u}' - \hat{u}'$ on the interval is easily seen to be bounded by the kink, which can be expressed as

$$|\bar{u}' - \hat{u}'| \leq \frac{h}{2}|\bar{u}''| \tag{152.6}$$

with

$$\bar{u}_i'' \equiv \frac{\bar{u}(ih + \frac{h}{2}) - 2\bar{u}(ih) + \bar{u}(ih - \frac{h}{2})}{(\frac{h}{2})^2} \approx f(ih + \frac{h}{2}). \tag{152.7}$$

We can thus estimate the difference between u with space step h and $\bar{u}$ with half space step $\frac{h}{2}$ as follows:

$$\|u' - \bar{u}'\| \leq \|\bar{u}' - \hat{u}'\| \leq \frac{h}{2}\|\bar{u}''\| \approx \frac{h}{2}\|f\|. \tag{152.8}$$

Repeating the process with space step $\frac{h}{4}$ as in the proof of the Fundamental Theorem, we are thus led to the following *a priori error estimate* for the difference between a computed u with time step h and a fictional exact solution $\bar{u}$ computed with vanishingly small space step:

Theorem 152.1 *The finite element solution u of the BVP (151.1) on a mesh with mesh size h, satisfies the following error estimate:*

$$\|u' - \bar{u}'\| \leq \|\bar{u}' - \hat{u}'\| \leq h\|\bar{u}''\| \leq h\|f\|. \tag{152.9}$$

where $\hat{u}$ is an h-interpolant of the solution $\bar{u}$ with vanishingly small mesh size.

152.2 A Posteriori and A Priori Error Estimates

The error estimate (152.9) can be viewed in two ways, as:

- An *a posteriori error estimate*, where the error is estimated in terms of the data and more generally also the computed solution u, but not the hypothetical fine-mesh solution $\bar{u}$.
- An *a priori error estimate*, where the error is estimated in terms of a hypothetical fine-mesh solution $\bar{u}$, which is not computed.

The *a posteriori* variant takes the form

$$\|u' - \bar{u}'\| \leq 2h\|f\|, \tag{152.10}$$

and the *a priori* variant:

$$\|u' - \bar{u}'\| \leq 2h\|\bar{u}''\|. \tag{152.11}$$

An *a posteriori* error estimate can be made only *after* u has been computed, because the estimate involves u (in general). An *a priori* estimate states something about the error *before* the computation, but involves a hypothetical (in general unknown) fine-mesh solution.

A *a posteriori* error estimate is directly useful, while the practical value of an *a priori* estimate involving an unknown fine-mesh solution, is unclear.

We shall see that a sharp (accurate) *a posteriori* error estimates can be derived in great generality, and thus are very useful. On the other hand, sharp *a priori* error estimates can be derived only for special problems and thus are less useful.

You will see below that the the technique to derive a posteriori error estimates is based on computational solution of an auxiliary linearized (dual) problem, which reveals the crucial quantitative stability aspects connecting residuals of computed solutions to output errors. In a priori error estimation without computation, the quantitative stability aspects have to be revealed analytically, which is possible only in simple model problems.

The net result is that *a posteriori* error estimation is possible in great generality, because the stability is assessed computationally, and *a priori* error estimation in general is impossible, because the stability cannot be assessed analytically.

152.3 Read More

- Two-point BVP

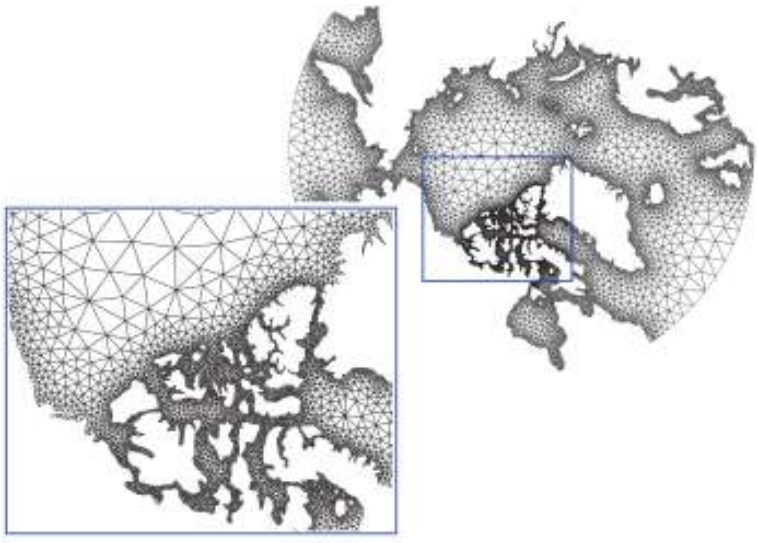


FIGURE 152.1. FEM mesh resolving the narrow straits of the Canadian Arctic Archipelago.

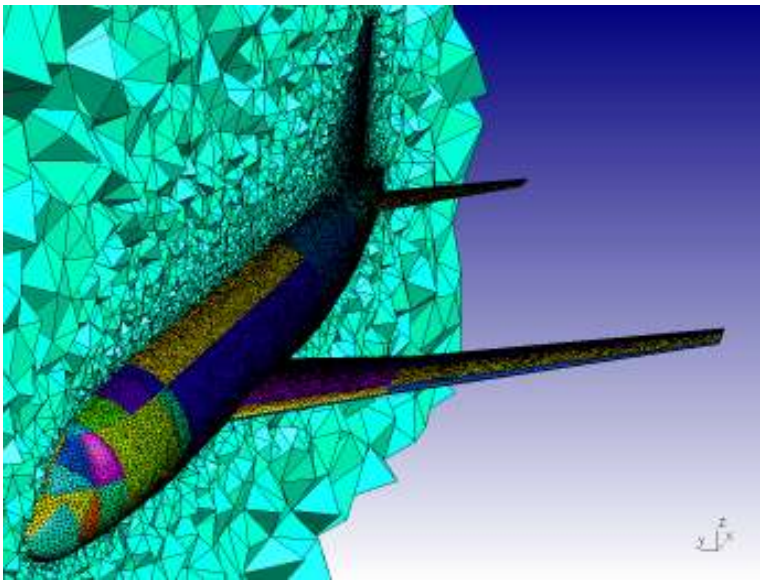


FIGURE 152.2. FEM mesh around an Airbus 319 by Gmsh.

153
FEM as Best Possible

> One cannot inquire into the foundations and nature of mathematics without delving into the question of the operations by which the 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. (Luitzen Egbertus Brouwer)

153.1 A Magical Property

The argument leading to the a priori error estimate (152.9) shows that the FEM solution u on a given mesh makes the error as small as possible in the sense that there is no other function v formed by the same basis functions with a smaller error:

$$\|u' - \bar{u}'\| \leq \|v' - \bar{u}'\| \tag{153.1}$$

You met the same argument in Session Piecewise Linear Interpolation in the proof that the L_2-projection is best-possible.

FEM chooses a best solution on a given mesh in the sense that the deviation to the solution with vanishingly small mesh size, is as small as possible (in the specific sense of the estimate) using functions on the given mesh.

FEM thus has the, at first sight magical, property of choosing a best approximation on a given mesh with finite mesh size to the (exact) solution

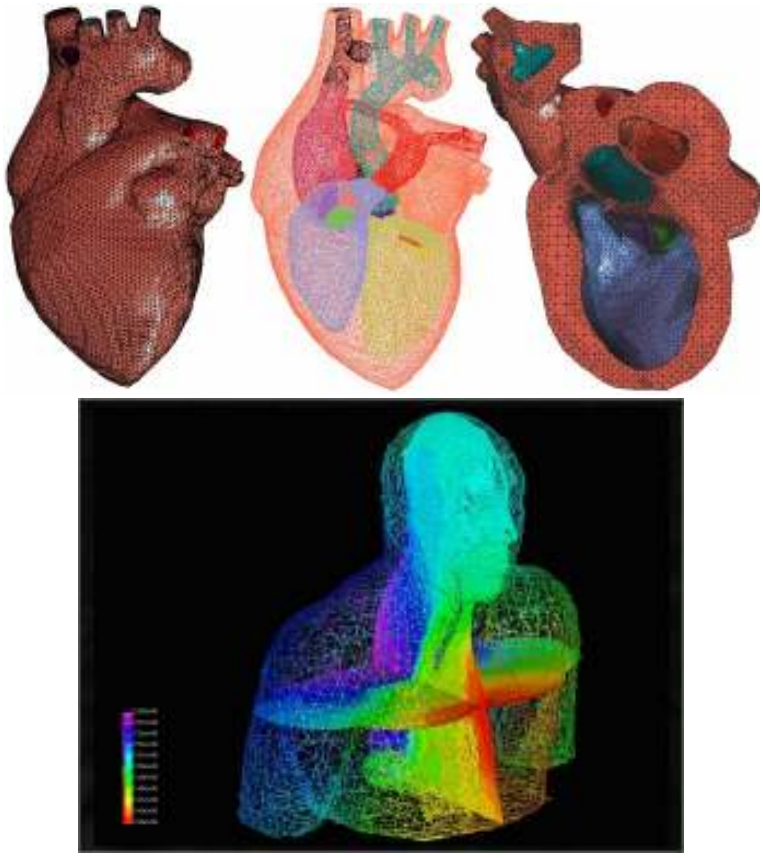


FIGURE 153.1. FEM mesh of human heart and upper body.

with vanishing mesh size, without knowing the exact solution, only the data to the differential equation.

FEM is thus a clever method: It can be expected to do the best possible on a given mesh. Below we shall find the extent and limits of this optimality.

153.2 Learn More

- How good is best possible?
- Two-point BVP

154
FEM Heat: $\dot{u} - u'' = f$

> Heat, like gravity, penetrates every substance of the universe, its rays occupy all parts of space. The object of our work is to set forth the mathematical laws which this element obeys. The theory of heat will hereafter form one of the most important branches of general physics. (Joseph Fourier in Analytical Theory of Heat)

We consider the IBVP: Find $u(x,t)$ such that

$$\begin{aligned} \dot{u} - u'' &= f \quad \text{for } x \in (0,1),\ t > 0, \\ u(0,t) &= u(1,t) = 0 \quad \text{for } t > 0, \\ u(x,0) &= u^0(x), \quad \text{for } x \in (0,1), \end{aligned} \tag{154.1}$$

where $f(x,t)$ is a given forcing function and $u^0(x)$ a given initial value function. This problem describes heat conduction in a rod occupying the interval $[1,0]$ with coefficient of heat conductivity of unit size. Another interpretation is damped elastic string assuming small mass so that the dynamic term $\ddot{u}$ is small and therefore can be neglected, as discussed above.

We repeat the procedure of discretization by FEM introduced above for the wave equation: We seek a solution $u(x,t)$ of the form

$$u(x,t) = \sum_{j=1}^{J} u_j(t)\varphi_j(x) \tag{154.2}$$

with $\varphi_1(x), ..., \varphi_J(x)$, tent functions defined by

$$\varphi_i(jh) = 1 \quad \text{if } j = i, \quad \varphi_i(jh) = 0 \quad \text{else}, \tag{154.3}$$

where $h = \frac{1}{J+1}$ is the step size in x, and $u_1(t), ..., u_J(t)$, are unknown coefficients depending on t.

We seek to determine the coefficient functions $u_1(t), ..., u_J(t)$, by multiplying the differential equation $\dot{u} - u'' = f$ by $\varphi_i(x)$ for $i = 1, ..., J$, and integrate with respect to x to get

$$\sum_{i=1}^{J} \dot{u}_i(t) \int_0^1 \varphi_i(x)\varphi_j(x)\, dx + \sum_{i=1}^{J} u_i(t) \int_0^1 \varphi_i'(x)\varphi_j'(x)\, dx = \int_0^1 f\varphi_j\, dx \tag{154.4}$$

where we used integration by parts to rewrite

$$-\int_0^1 u''\varphi_i\, dx = \int_0^1 u'\varphi_i'\, dx \tag{154.5}$$

moving one derivative from u'' onto φ_i and changing sign, using that $\varphi_i(0) = \varphi_i(1) = 0$.

This is a system of ODEs in the coefficient vector function $u = (u_1, u_2, ..., u_J)$ of the form: Find $u(t)$

$$\begin{aligned} M\dot{u}(t) + Au(t) &= b(t) \quad \text{for } t > 0, \\ u(0) &= u^0, \end{aligned} \tag{154.6}$$

with M and $A = (a_{ij})$ mass and stiffness matrices given by

$$m_{ij} = \int_0^1 \varphi_i(x)\varphi_j(x) \quad a_{ij} = \int_0^1 \varphi_i'(x)\varphi_j'(x)\, dx, \tag{154.7}$$

and the forcing $b(t) = (b_1(t), ..., b_J(t)$ is given by

$$b_i(t) = \int_0^1 f(t, x)\varphi_i(x)dx. \tag{154.8}$$

As before A is symmetric with coefficients

$$a_{ii} = \frac{2}{h},\ a_{i\,i+1} = a_{i\,i-1} = -\frac{1}{h},\ a_{ij} = 0 \text{ else}, \tag{154.9}$$

which corresponds to the approximation

$$u'' = \frac{d^2u}{dx^2} \approx \frac{u^{i+1} - 2u^i + u^{i-1}}{h^2}. \tag{154.10}$$

Also the mass matrix M is the same as above, and can be thought of as an approximate (scaled) identity matrix.

155
FEM Convection: $\dot{u} + u' = 0$

For scholars and laymen alike it is not philosophy but active experience in mathematics itself that can alone answer the question: What is mathematics? (Richard Courant)

155.1 A Basic Model of Convection

We now consider the IBVP: Find $u(x,t)$ such that

$$\begin{aligned} \dot{u} + u' &= 0 \quad \text{for } x \in (0,1),\ t > 0, \\ u(0,t) &= g(t) \quad \text{for } t > 0 \\ u(x,0) &= u^0(x) \quad \text{for } x \in (0,1), \end{aligned} \tag{155.1}$$

where $g(t)$ is a given boundary value and $u^0(x)$ a given initial value. This problem describes convection along the x-axis with velocity 1 to the right. The solution is given by

$$\begin{aligned} u(x,t) &= u^0(x-t) \quad \text{for } x > t, \\ u(x,t) &= g(t-x) \quad \text{for } x < t. \end{aligned} \tag{155.2}$$

The solution $u(x,t)$ is constant on straight lines $x = t + c$ with c a constant, with the value $u(c,0) = u_0(c)$ for $c > 0$ being transported to $u(t+c,t)$, and the value $u(0,t) = g(t)$ being transported to $u(x, x+t)$, signifying that "information is flowing" from left to right (in the positive x-direction) with speed 1. The straight lines $x = t + c$ are called *characteristics*, and the solution values thus are propagated along the characteristics.

155.2 FEM

We seek a solution $u(x,t)$ of the form

$$u(x,t) = \sum_{j=0}^{J} u_j(t)\varphi_j(x) \tag{155.3}$$

where $\varphi_0(x), ..., \varphi_J(x)$, are continuous piecewise linear tent functions defined on $(0,1)$ by

$$\varphi_i(jh) = 1 \quad \text{if } j = i, \quad \varphi_i(jh) = 0 \quad \text{else}, \tag{155.4}$$

where we set the mesh size $h = \frac{1}{J}$. Note that $\varphi_0(x)$ and $\hat{\varphi}_J(x)$ are "half tents", because we restrict to $0 < x < 1$.

We set $u_0(t) = g(t)$ and seek to determine the remaining coefficient functions $u_1(t), ..., u_J(t)$, by multiplying the differential equation $\dot{u} = -u'$ by $\varphi_i(x)$ for $i = 1, ..., J$, and integrating with respect to x to get

$$\sum_{i=0}^{J} \dot{u}_j(t) \int_0^1 \varphi_i(x)\varphi_j(x)\,dx + \sum_{j=0}^{J} u_j(t) \int_0^1 \varphi_i(x)\varphi_j'(x)\,dx = 0, \tag{155.5}$$

resulting in the following ODE: Find $u = (u_0, u_1, u_2, ..., u_J)$ such that

$$\begin{aligned} M\dot{u}(t) + Au(t) = 0, \quad u_0(t) = g(t), \quad \text{for } t > 0, \\ u_j(0) = u^0(jh), \quad j = 1, ..., J, \end{aligned} \tag{155.6}$$

where $M = (m_{ij})$ and $A = (a_{ij})$ with coefficients

$$m_{ij} = \int_0^1 \varphi_i(x)\varphi_j(x)\,dx, \quad a_{ij} = -\int_0^1 \varphi_i(x)\varphi_j'(x)\,dx, \quad i = 1, ..., J, \quad j = 0, ..., J. \tag{155.7}$$

Direct analytical evaluation of the integrals with piecewise polynomial integrands gives

$$\begin{aligned} m_{ii} = \frac{2h}{3},\ m_{i\,i+1} = m_{i\,i-1} = \frac{h}{6},\ m_{ij} = 0 \text{ else},\ i,j = 1, ..., J-1, \\ m_{J,J} = \frac{h}{3},\ m_{J\,J-1} = \frac{h}{6} \\ a_{ii} = 0,\ a_{i\,i+1} = \frac{1}{2},\ a_{i\,i-1} = -\frac{1}{2},\ a_{ij} = 0 \text{ else},\ i = 1, ..., J-1, \\ ,a_{J\,J-1} = -\frac{1}{2},\ a_{J\,J} = \frac{1}{2}. \end{aligned} \tag{155.8}$$

that is the equation for $u_i(t)$ takes the following form for $i = 1, , J-1$,

$$\frac{1}{6}\dot{u}_{i-1} + \frac{2}{3}\dot{u}_i + \frac{1}{6}\dot{u}_{i+1} + \frac{u_{i+1} - u_{i-1}}{2h} = 0, \quad i = 1, ..., J-1, \tag{155.9}$$

where $u_0 = g$ is given, and for $i = J$,

$$\frac{1}{6}\dot{u}_{J-1} + \frac{1}{3}\dot{u}_J + \frac{u_J - u_{J-1}}{2h} = 0. \tag{155.10}$$

If we *lump* the mass matrix moving the off-diagonal coefficients $\frac{1}{6}$ to the diagonal, then we get

$$\dot{u}_i + \frac{u_{i+1} - u_{i-1}}{2h} = 0, \quad \text{for } i = 1, ..., J-1, \ \frac{1}{2}\dot{u}_J + \frac{u_J - u_{J-1}}{2h} = 0, \tag{155.11}$$

where we can imagine

$$u'(ih) \approx \frac{u_{i+1} - u_{i-1}}{2h} \quad \text{for } i = 1, ..., J-1, \quad u'(Jh) \approx \frac{u_J - u_{J-1}}{h}, \tag{155.12}$$

and thus view (155.11) as a discrete analog of $\dot{u} + u' = 0$.

Note that the square stiffness matrix $A = (a_{ij})$ with $i, j = 1, ..., J$ is no longer symmetric, but *anti-symmetric* in the sense that transposition changes the sign: $A^\top = -A$, while for a symmetric matrix transposition doesn't change anything. The sign change under transposition makes convection problem very different from a diffusion problem, motivating to make a distinction between *convection-dominated* and *diffusion-dominated* problems, as indicated in the next section.

We can solve the discrete equations by time-stepping as follows:

$$m_{ii}u_i^{n+1} = -\sum_{j \neq i} m_{ij}u_j^n + dt\sum_{j=1}^{J} a_{ij}u_j^n, \quad i = 1, ..., J, \tag{155.13}$$

where $u_i^n = u_i(nk)$ and k is a time step.

155.3 Central vs Upwind Discrete Derivative

We meet in the discrete problem (155.11) the following two approximations of the derivative $u'(ih)$:

$$\begin{aligned} u'(ih) &\approx \frac{u_{i+1} - u_{i-1}}{2h} \quad \text{(central approximation)}, \\ u'(ih) &\approx \frac{u_i - u_{i-1}}{h} = 0 \quad \text{(upwind approximation)}, \end{aligned} \tag{155.14}$$

where the central approximation uses the two points $(i+1)h$ and $(i-1)h$ centered around ih, while the upwind approximation uses only the point $(i-1)h$ to the left (or upwind since the "wind" is coming from the left with the information propagating from left to right) of ih. We understand that an upwind approximation is more physical, since information is propagated from left to right over the time step. On the other hand the central

approximation is formally of higher accuracy (second order in h) than the upwind approximation(first order in h).

When you compute with the above method you will discover that it gives garbage in certain cases, as a consequence of the (partly unphysical) central approximation of the convection term u'. Below you will see how to modify the standard (basic) Galerkin FEM just presented to work well in all cases.

There two main classes of problems in fluid mechanics: *convection-dominated* problems and *diffusion-dominated* problems, or problems with small viscosity and large viscosity. You will discover that for convection-dominated problems with quickly varying solutions (non-smooth solutions), standard Galerkin does not work well and has to be modified, while for diffusion-dominated problems no modification is necessary. The modification has made FEM into a general method for a wide variety of problems.

155.4 Read More

- Stationary Convection-Diffusion Analysis
- Stationary Convection-Diffusion FEM
- Time-Dependent Convection-Diffusion Analysis
- Time-Dependent Convection-Diffusion FEM

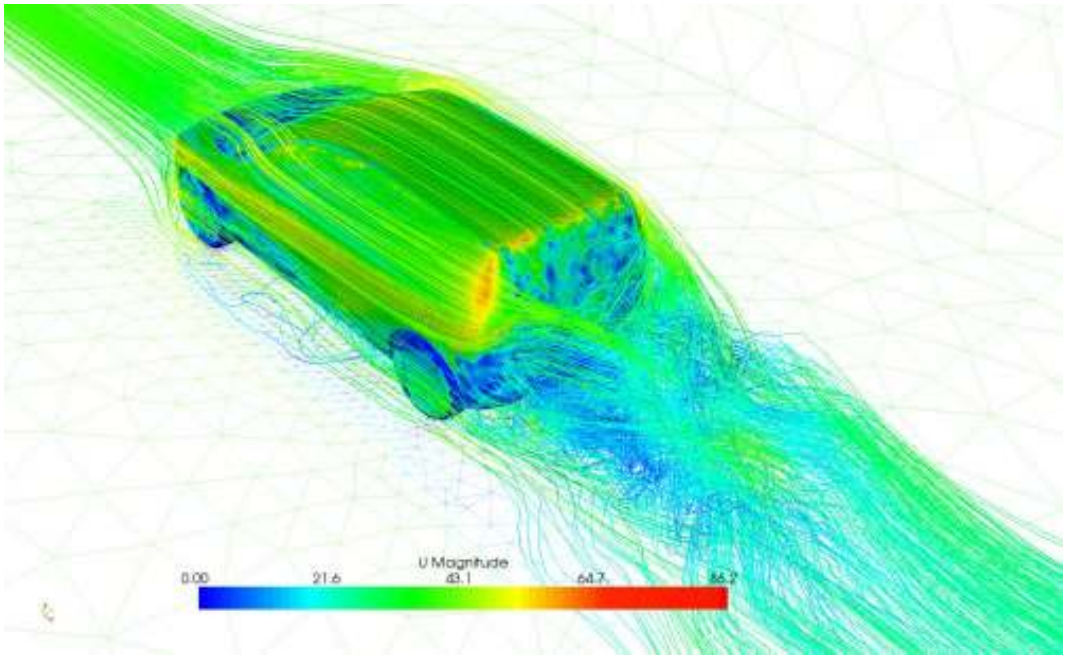


FIGURE 155.1. FEM simulation of convection-dominated airflow around a Volvo.

156
FEM Heat: $\dot{u} - \Delta u = f$

> Either mathematics is too big for the human mind or the human mind is more than a machine. (Kurt Gdel)

We consider the IBVP: Find $u(x,t)$ such that

$$\begin{aligned} \dot{u} - \Delta u &= f \quad \text{for } x \in \Omega,\ t > 0, \\ u(x,t) &= 0 \quad \text{for } x \in \Gamma,\ t > 0, \\ u(x,0) &= u^0(x), \quad \text{for } x \in \Omega, \end{aligned} \tag{156.1}$$

where Ω is a domain in space with boundary Γ, $f(x,t)$ is a given forcing function and $u^0(x)$ a given initial value function. Seek a solution $u(x,t)$ of the form

$$u(x,t) = \sum_{j=1}^{J} u_j(t)\varphi_j(x), \tag{156.2}$$

where $\varphi_1(x), ..., \varphi_J(x)$, are the following continuous piecewise linear *tent functions* defined by

$$\varphi_i(x^j) = 1 \quad \text{if } j = i, \quad \varphi_i(x^j) = 0 \quad \text{else}, \tag{156.3}$$

where now $x^1, ..., x^J$ are the interior nodes of a triangulation of Ω with mesh size h. Since only interiors nodes appear, all basis functions vanish on the boundary Γ.

We seek to determine the coefficient functions $u_1(t), ..., u_J(t)$, by multiplying the differential equation $\dot{u} - \Delta u = f$ by $\varphi_i(x)$ for $i = 1, ..., J$, and

integrating with respect to x to get

$$\sum_{j=1}^{J} \dot{u}_j(t) \int_\Omega \varphi_i \varphi_j \, dx + \sum_{j=1}^{J} u_j u(t) \int_\Omega \nabla\varphi_j \nabla\varphi_i \, dx = \int_\Omega f\varphi_i \, dx, \quad i = 1, ..., J, \tag{156.4}$$

where we used integration by parts to rewrite

$$-\int_\Omega \Delta u \varphi_i \, dx = \int_\Omega \nabla u \nabla \varphi_i \, dx \tag{156.5}$$

moving one derivative from Δu onto φ_i and changing sign, and using that $\varphi_i = 0$ on Γ.

This is a system of ODEs in the coefficient vector function $u = (u_1, u_2, ..., u_J)$ of the form: Find $u(t)$ such that

$$\begin{aligned} M\dot{u}(t) + Au(t) = b(t) \quad &\text{for } t > 0, \\ u(0) &= u^0, \end{aligned} \tag{156.6}$$

with M and $A = (a_{ij})$ mass and stiffness matrices given by

$$m_{ij} = \int_\Omega \varphi_i \varphi_j \, dx, \quad a_{ij} = \int_\Omega \nabla\varphi_i \nabla\varphi_j \, dx, \tag{156.7}$$

and the forcing $b(t)$ has coefficients

$$b_i(t) = \int_\Omega f(t,x)\varphi_i(x) \, dx. \tag{156.8}$$

We can solve this system by time-stepping as above.

156.1 Learn More

- FEM for Heat Equation

157
FEM Poisson: $-\Delta u = f$

> To those who ask what the infinitely small quantity in mathematics is, we answer that it is actually zero. Hence there are not so many mysteries hidden in this concept as they are usually believed to be. (Leonhard Euler)

The stationary version of the previous IBVP is the BVP: Find $u : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} -\Delta u &= f \quad \text{for } x \in \Omega, \\ u(x) &= 0 \quad \text{for } x \in \Gamma, \end{aligned} \tag{157.1}$$

where Ω is a domain in space with boundary Γ, $f(x)$ is a given forcing function. Seek a solution $u(x)$ of the form

$$u(x) = \sum_{j=1}^{J} u_j \varphi_j(x) \tag{157.2}$$

where $\varphi_1(x), ..., \varphi_J(x)$, are tent functions on a triangulation and $u_1, ..., u_J$, are unknown coefficients associated with interior nodes.

We seek to determine the coefficients $u_1, ..., u_J$, by multiplying the differential equation $\dot{u} - \Delta u = f$ by $\varphi_i(x)$ for $i = 1, ..., J$, and integrating with respect to x to get

$$\sum_{j=1}^{J} \dot{u}_j \int_\Omega \varphi_i \varphi_j \, dx + \sum_{j=1}^{J} u_j \int_\Omega \nabla\varphi_i \nabla\varphi_j \, dx = \int_\Omega f\varphi_i \, dx, \quad i = 1, ..., J, \tag{157.3}$$

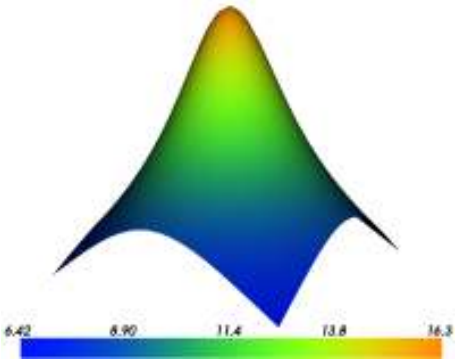


FIGURE 157.1. FEM solution of Poisson's equation.

where we used integration by parts to rewrite

$$-\int_\Omega \Delta u\varphi_i\,dx = \int_\Omega \nabla u\nabla\varphi_i\,dx \tag{157.4}$$

moving one derivative from Δu onto φ_i and changing sign.

This is a linear system in the coefficient vector $u = (u_1, u_2, ..., u_J)$ of the form: Find $u = (u_1, ..., u_J)$ such that

$$Au = b, \tag{157.5}$$

where $A = (a_{ij})$ is a stiffness matrix with coefficients

$$a_{ij} = \int_\Omega \nabla\varphi_i\nabla\varphi_j\,dx \tag{157.6}$$

and the forcing vector b has coefficients

$$b_i = \int_\Omega f(x)\varphi_i(x)\,dx \tag{157.7}$$

The stiffness matrix A is symmetric and positive definite, and the system $Au = b$ can be solved by Gaussian elimination or time-stepping.

157.1 The Finite Element Space V_h

It is useful to introduce the linear space of functions V_h spanned by the basis functions $\varphi_1, , \varphi_J$, which consists of all functions $v(x)$ of the form

$$v(x) = \sum_{j=1}^{J} v_j\varphi_j(x) \tag{157.8}$$

where $v_1, , v_j$, are real coefficients. The finite element solution $u(x)$ is of this form and thus $u \in V_h$, and of course each basis function $\varphi_j \in V_h$. We can now formulate FEM for Possion's equation as follows: Find $u \in V_h$ such that

$$\int_\Omega \nabla u \nabla \varphi_j \, dx = \int_\Omega f \varphi_j \, dx \quad \text{for } j = 1, ..., J, \tag{157.9}$$

which is equivalent to: Find $u \in V_h$ such that

$$\int_\Omega \nabla u \nabla v \, dx = \int_\Omega f v \, dx \quad \text{for all } v \in V_h. \tag{157.10}$$

The basic "best-possible" error estimate can then be expressed as follows: The FEM solution $u \in V_h$ satisfies

$$\|\nabla(u - \bar{u})\| \leq \|\nabla(v - \bar{u})\| \quad \text{for all } v \in V_h, \tag{157.11}$$

where $\bar{u}$ is the (unique) solution with vanishingly small mesh size, where

$$\|\nabla w\| = (\int_\Omega |\nabla w|^2 \, dx)^{\frac{1}{2}}. \tag{157.12}$$

In particular, chosing $v = \hat{u} \in V_h$ as a nodal interpolant, gives the following a priori error estimate

$$\|\nabla(u - \bar{u})\| \leq \|\nabla(\hat{u} - \bar{u})\| \leq Ch\|D^2\bar{u}\|, \tag{157.13}$$

where h is the mesh size, D^2u measures the maximal second derivate of u, and $C \approx 1$.

157.2 Learn More

- FEM for Poisson's Equation

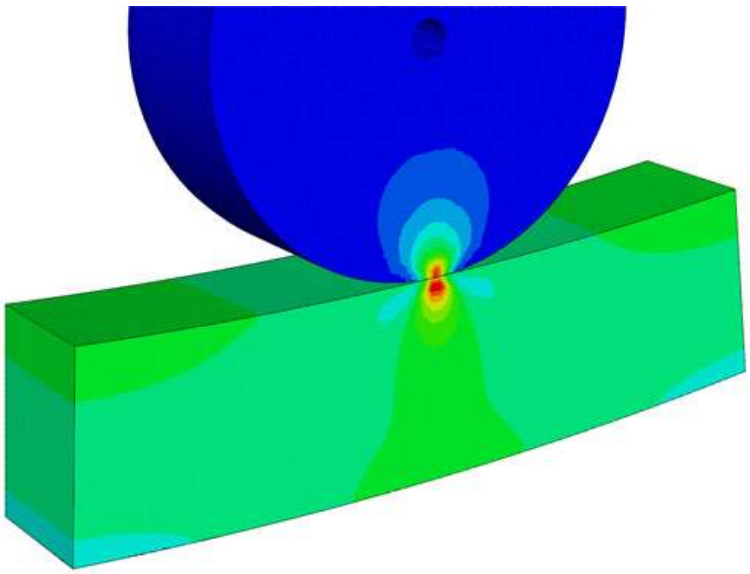


FIGURE 157.2. FEM simulation of stress distribution in wheel on rail.

158
FEM Wave: $\ddot{u} - \Delta u = f$

> The reader will find no figures in this work. The methods which I set forth do not require either constructions or geometrical or mechanical reasonings: but only algebraic operations, subject to a regular and uniform rule of procedure. (Joseph-Louis La-Grange)

FEM for the wave equation leads to the ODE: Find $u = (u_1, ..., u_J)$ such that

$$\begin{aligned} M\ddot{u}(t) + Au(t) = b(t) \quad \text{for } t > 0, \\ u(0) = u^0, \quad \dot{u}(0) = \dot{u}^0, \end{aligned} \tag{158.1}$$

with M, A and $b(t)$ the same as for the heat equation above. Again the ODE an be solved by time stepping as above.

158.1 Learn More

- FEM for Wave Equation

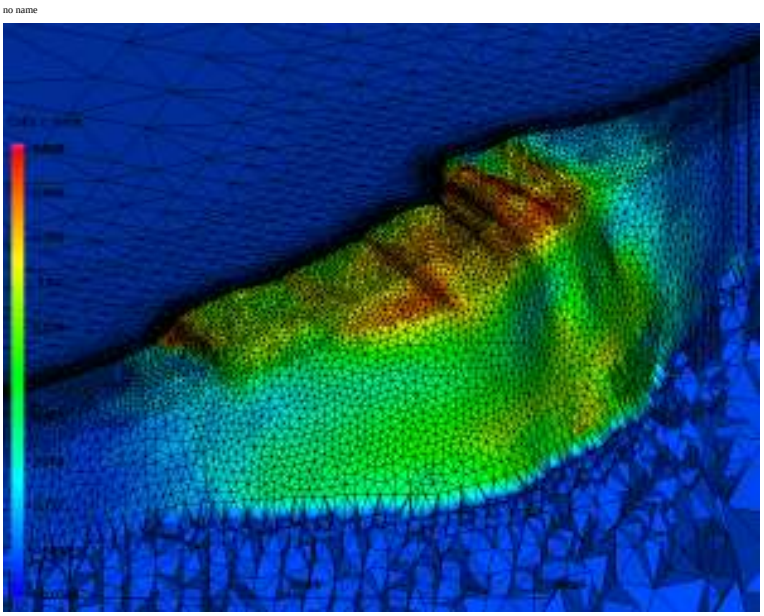


FIGURE 158.1. FEM mesh for simulation of Earth quake in the San Andreas Fault.

FIGURE 158.2. Is this a FEM triangulation of space? (Ask Thomas Saraceno)

159

"Do Nothing" Natural Boundary Conditions

To do nothing at all is the most difficult thing in the world, the most difficult and the most intellectual. (Oscar Wilde)

All that is necessary for the triumph of evil is for good men to do nothing. (Edmund Burke)

159.1 Robin with Neumann and Dirichlet

We now turn to the question of boundary conditions, considered in the World of Differential Equations.

So far, we have here considered Poisson's equation $-\Delta u = f$ in a domain Ω with boundary Γ assuming homogeneous Dirichlet boundary conditions $u = 0$ on Γ. We now consider *Robin boundary conditions* of the form

$$\kappa u + \frac{\partial u}{\partial n} = g \quad \text{on } \Gamma, \tag{159.1}$$

where κ and ν are non-negative coefficients and g is given, and $\frac{\partial u}{\partial n} = n \cdot \nabla u$ with n the outward unit normal to Γ. In the context of the heat equation, the boundary condition expresses with $\kappa > 0$ proportionality between the heat flux across the boundary $n \cdot \nabla u$ and the temperature difference $(u_+ - u)$ if we choose $g = \kappa u_+$ through the coefficient κ, where u_+ is a given temperature just outside and u the unkown temperature inside the boundary.

Chosing $\kappa = 0$ gives the *Neumann boundary condition*

$$\frac{\partial u}{\partial n} = g \quad \text{on } \Gamma, \tag{159.2}$$

and letting κ become large, we effectively obtain the non-homogeneous Dirichlet condition

$$u = u_+ \tag{159.3}$$

if we choose $g = \kappa u_+$ as above.

We can thus view the Robin condition to effectively contain also (non-homogeneous) Neumann and Dirichlet boundary conditions. The advantage with a Robin condition is that it is a *natural boundary condition* or in other words a "do nothing" condition, in the sense that the finite element functions are not required to satisfy any boundary conditions at all. The Robin boundary condition is instead implicitely contained in the Galerkin variational formulation by changing the stiffness matrix $A = (a_{ij})$ to

$$a_{ij} = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \, dx + \int_\Gamma \kappa\varphi_i\varphi_j \, dx, \tag{159.4}$$

and the load vector to

$$b_i = \int_\Omega f\varphi_i \, dx + \int_\Gamma g\varphi_i \, dx. \tag{159.5}$$

Galerkin's method thus implements the Robin boundary condition in *variational form*; as compared to explicite enforcement of the boundary condition in *strong form*, as when requiring both trial and test functions to satisfy a homogeneous Dirichlet condition $u = 0$.

A Robin condition in variational form is also a called a *natural boundary condition*, while a Dirichlet condition is strong form i callled an *essential boundary condition*.

Does this work? Yes! But why is here "do nothing" OK? If you want to know, take a look at FEM in 2d and 3d or FEm for Poisson.

Recall that implementing non-homogeneous Dirichlet conditions in strong form requires care to ask trial functions to satisfy the given boundary conditions, while letting the test functions satisfy a homogeneous condition. Further, trying to satisfy Neumann conditions in strong form leads to great difficulties. All of this can be avoided by simply using a Robin condition implemented in variational form as a "do nothing" condition. Sometimes, do nothing is the best you can do...in a Leibnizian best of worlds.

160

Linearization and Stability of Initial Value Problems

The logos of somewome to that base anything, when most characteristically mantissa minus, comes to nullum in the endth: orso, here is nowet badder than the sin of Aha with his cosin Lil, verswaysed on coversvised, and all that's consecants and cotangincies... (Finnegans Wake, James Joyce)

160.1 Introduction

We now address the basic problem of *stability* of solutions to differential equations as a measure of the *sensitivity of solutions to perturbations in given data*.

We consider our basic IVP: Find $u : [0, T] \to \mathbb{R}^d$ such that

$$\dot{u}(t) = f(u(t)) \quad \text{for } 0 < t \leq T,\ u(0) = u^0, \tag{160.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$ is a given bounded Lipschitz continuous function, $u^0 \in \mathbb{R}^d$ is a given initial value and $[0, T]$ as a given time interval.

To study the stability of a given solution $u(t)$ to small perturbations in given data, e.g. in the given initial data u^0, we will consider an associated *linearized problem* that arises upon linearizing the function $v \to f(v)$ at the solution $u(t)$.

160.2 Stationary Solutions

We consider first the simplest case of a *stationary solution* $u(t) = \bar{u}$ for $0 \le t \le T$, that is a solution $u(t)$ of (219.1) that is independent of time t. Since $\dot{u}(t) = 0$ if $u(t)$ is independent of time, $u(t) = \bar{u}$ is a stationary solution if $f(\bar{u}) = 0$ and $u^0 = \bar{u}$, where $\bar{u} = (\bar{u}_1, ..., \bar{u}_d) \in \mathbb{R}^d$.

The equation $f(\bar{u}) = 0$ corresponds to a system of d equations $f_i(\bar{u}_1, ..., \bar{u}_d) = 0$, $i = 1, ..., d$, in d unknowns $\bar{u}_1, ..., \bar{u}_d$, where the f_i are the components of f. We studied computational solution of such systems in Newton's Method and Fixed Point Iteration.

Here, we assume the existence of a stationary solution $u(t) = \bar{u}$ so that $\bar{u} \in \mathbb{R}^d$ satisfies the equation $f(\bar{u}) = 0$. In general, there may be several roots $\bar{u}$ of the equation $f(v) = 0$ and thus there may be several stationary solutions. We also refer to a stationary solution $u(t) = \bar{u}$ as an *equilibrium solution*.

160.3 Linearization at a Stationary Solution

We shall now study perturbations of a given stationary solution under small perturbations of initial data. We thus assume $f(\bar{u}) = 0$ and denote the corresponding equilibrium solution by $\bar{u}(t)$ for $t > 0$, that is $\bar{u}(t) = \bar{u}$ for $t > 0$. We consider the initial value problem (219.1) with $u^0 = \bar{u} + \varphi^0$, where $\varphi^0 \in \mathbb{R}^d$ is a given small perturbation of the initial data $\bar{u}$. We denote the corresponding solution by $u(t)$ and focus attention on the corresponding perturbation in the solution, that is $\psi(t) = u(t) - \bar{u}(t) = u(t) - \bar{u}$. We want to derive a differential equation for the perturbation $\psi(t)$, and to this end we linearize f at $\bar{u}$ and write

$$f(u(t)) = f(\bar{u} + \psi(t)) = f(\bar{u}) + f'(\bar{u})\psi(t) + e(t),$$

where $f'(\bar{u})$ is the Jacobian of $f : \mathbb{R}^d \to \mathbb{R}^d$ at $\bar{u}$ and the error term $e(t)$ is quadratic in $\psi(t)$ (and thus is very small if $\psi(t)$ is small). Since $f(\bar{u}) = 0$ and $u(t)$ satisfies (219.1), we have

$$\dot{\psi}(t) = \frac{d}{dt}(\bar{u} + \psi(t))) = f(u(t)) = f'(\bar{u})\psi(t) + e(t).$$

Neglecting the quadratic term $e(t)$, we are led to a linear initial value problem,

$$\dot{\varphi}(t) = f'(\bar{u})\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0, \tag{160.2}$$

or

$$\dot{\varphi}(t) = A\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0, \tag{160.3}$$

where $A = f'(\bar{u})$ is a constant $d \times d$ matrix and $\varphi(t)$ is an approximation of the perturbation $\psi(t) = u(t) - \bar{u}$ up to a second order term.

If the matrix A is diagonalizable, so that $A = B\Lambda B^{-1}$ where B is a non-singular $d \times d$ matrix and Λ is a diagonal matrix with the eigenvalues $\lambda_1, ..., \lambda_d$ of A on the diagonal, then we can write the solution to (160.3) as

$$\varphi(t) = B\exp(t\Lambda)B^{-1}\varphi^0 \quad \text{for } t \geq 0. \tag{160.4}$$

where $\exp(t\Lambda)$ is a diagonal matrix with diagonal elements $\exp(t\lambda_1), ..., \exp(t\lambda_d)$. We see that each component of $\varphi(t)$ is a linear combination of $\exp(t\lambda_1)$,..., $\exp(t\lambda_d)$ and the sign of the real part Re λ_i of λ_i determines if the corresponding term grows or decays exponentially. If some Re $\lambda_i > 0$, then we have exponential growth of certain perturbations, which indicates that the corresponding stationary solution $\bar{u}$ is *unstable*. On the other hand, if all Re $\lambda_i \leq 0$, then we would expect $\bar{u}$ to be *stable*.

These considerations are qualitative in nature, and to be more precise we should base judgements of stability or instability on quantitative estimates of perturbation growth. In the diagonalizable case, (219.5) implies in the Euclidean vector and matrix norms that

$$\|\varphi(t)\| \leq \|B\|\|B^{-1}\| \max_{i=1,...,d} \exp(t\lambda_i)\|\varphi^0\|. \tag{160.5}$$

We see that the maximal perturbation growth is governed by the maximal exponential factors $\exp(t\lambda_i)$ as well as the factors $\|B\|$ and $\|B^{-1}\|$ related to the transformation matrix B. If the transformation matrix B is orthogonal, then $\|B\| = \|B^{-1}\| = 1$, and the perturbation growth is governed solely by the exponential factors $\exp(t\lambda_i)$. We give this case special attention:

160.4 Stability Analysis when $f'(\bar{u})$ Is Symmetric

If $A = f'(\bar{u})$ is symmetric so that $A = Q\Lambda Q^{-1}$ with Q orthogonal and Λ a diagonal matrix with real diagonal elements λ_i, then

$$\|\varphi(t)\| \leq \max_{i=1,...,d} \exp(t\lambda_i)\|\varphi^0\|. \tag{160.6}$$

In particular, if all eigenvalues $\lambda_i \leq 0$ then perturbations $\varphi(t)$ cannot grow with time, and we say that the solution $\bar{u}$ is stable. On the other hand, if some eigenvalue $\lambda_i > 0$ and the corresponding eigenvector is g_i then $\varphi(t) = \exp(t\lambda_i)g_i$ solves the linearized initial value problem (219.3) with $\varphi^0 = g_i$, and evidently the particular perturbation $\varphi(t)$ grows exponentially. We then say that the solution $\bar{u}$ is *unstable*. Of course, the size of the positive eigenvalues influence the perturbation growth, so that if $\lambda_i > 0$ is small, then then growth is slow and the instability is mild. Likewise, if λ_i is small negative, then the exponential decay is slow.

160.5 Stability Factors

We may express the stability features of a particular perturbation φ^0 through a *stability factor* $S_d(T,\varphi_0)$ defined as follows:

$$S_d(T,\varphi_0) = \max_{0\le t\le T} \frac{\|\varphi(t)\|}{\|\varphi^0\|},$$

where $\varphi(t)$ solves the linearized problem (219.3) with initial data φ^0. The stability factor $S_d(T,\varphi_0)$ measures the maximal growth of the norm of $\varphi(t)$ over the time interval $[0,T]$ versus the norm of the initial value φ_0. with the subscript d representing data. In computational solution we meet another stability factor denoted by S_c with c for computation.

We can now seek to capture the overall stability features of a stationary solution $\bar{u}$ by maximization over all different perturbations:

$$S_d(T) = \max_{\varphi^0\neq 0} S(T,\varphi_0).$$

If the stability factor $S(T)$ is large, then some perturbations grow very much over the time interval $[0,T]$, which indicates a strong sensitivity to perturbations or *instability*. On the other hand, if $S(T)$ is of moderate size then the perturbation growth is moderate, which signifies *stability*. Using the Euclidean matrix norm, we can also express $S(T)$ as

$$S_d(T) = \max_{0\le t\le T} \|\exp(tA)\|.$$

EXAMPLE 160.1. If $A = f'(\bar{u})$ is symmetric with eigenvalues $\lambda_1, ..., \lambda_d$, then

$$S_d(T) = \max_{i=1,\ldots,d} \max_{0\le t\le T} \exp(t\lambda_i).$$

In particular, if all $\lambda_i \le 0$, then $S(T) = 1$.

EXAMPLE 160.2.

The initial value problem for a pendulum takes the form

$$\begin{aligned} \dot{u}_1 = u_2, \quad \dot{u}_2 = -\sin(u_1) \quad \text{for } t > 0,\\ u_1(0) = u_{01},\ u_2(0) = u_{02}, \end{aligned}$$

corresponding to $f(u) = (u_2, -\sin(u_1)$ and the equilibrium solutions are $\bar{u} = (0,0)$ and $\bar{u} = (\pi,0)$. We have

$$f'(\bar{u}) = \begin{pmatrix} 0 & 1 \\ -\cos(\bar{u}_1) & 0 \end{pmatrix},$$

and the linearized problem at $\bar{u} = (0,0)$ thus takes the form

$$\dot{\varphi}(t) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \varphi(t) \equiv A_0\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0,$$

with solution

$$\varphi_1(t) = \varphi_1^0\cos(t) + \varphi_2^0\sin(t), \quad \varphi_2(t) = -\varphi_1^0\sin(t) + \varphi_2^0\cos(t).$$

It follows by a direct computation (or using that $\begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}$ is an orthogonal matrix), that for $t > 0$

$$\|\varphi(t)\|^2 = \|\varphi_0\|^2,$$

and thus the norm $\|\varphi(t)\|$ of a solution $\varphi(t)$ of the linearized equations is constant in time, which means that the stability factor $S(T) = 1$ for all $T > 0$. We conclude that if the norm of a perturbation is small initially, it will stay small for all time. This means that the equilibrium solution $\bar{u} = (0, 0)$ is *stable*. More precisely, if the pendulum is perturbed initially a little from its bottom position, the pendulum will oscillate back and forth around the bottom position with constant amplitude. This fits our direct experimental experience of course.

Note that the linearized operator A_0 is non-symmetric; the eigenvalues of A_0 are purely imaginary $\pm i$, which says that $\|\varphi(t)\| = \|\varphi_0\|$, that is a perturbation neither grows nor decays. Another way to derive this fact is to use the fact that A_0 is *antisymmetric*, that is $A_0^\top = -A_0$, which shows that $(A_0\varphi, \varphi) = (\varphi, A_0^\top\varphi) = -(\varphi, A_0\varphi) = -(A_0\varphi, \varphi)$, and thus $(A_0\varphi, \varphi) = 0$, where $(\cdot, \cdot)$ is the $\mathbb{R}^2$ scalar product. It follows from the equation $\dot{\varphi} = A_0\varphi$ upon multiplication by φ that $0 = (\dot{\varphi}, \varphi) = \frac{1}{2}\frac{d}{dt}(\varphi, \varphi) = \frac{1}{2}\frac{d}{dt}\|\varphi\|^2$, which proves that $\|\varphi(t)\|^2 = \|\varphi_0\|^2$.
The linearized problem at $\bar{u} = (\pi, 0)$ reads

$$\dot{\varphi}(t) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\varphi(t) \equiv A_\pi\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0,$$

with symmetric matrix A_π with eigenvalues ± 1. Since one eigenvalue is positive, the stationary solution $\bar{u} = (\pi, 0)$ is unstable. More precisely, the solution is given by

$$\varphi_1 = \frac{\varphi_1^0}{2}(e^t + e^{-t}) + \frac{\varphi_2^0}{2}(e^t - e^{-t}), \quad \varphi_2 = \frac{\varphi_1^0}{2}(e^t - e^{-t}) + \frac{\varphi_2^0}{2}(e^t + e^{-t}),$$

and due to the exponential factor e^t, perturbations will grow exponentially in time, and thus an initially small perturbation will become large as soon as $t \geq 10$ say. Physically, this means that if the pendulum is perturbed initially a little from its top position, the pendulum will eventually move away from the top position, even if the initial perturbation is very small. This fact of course has direct experimental evidence: to balance a pendulum with the weight in the top position is tricky business. Small perturbations quickly grow to large perturbations and the equilibrium solution $(\pi, 0)$ of the pendulum is unstable.

160.6 Stability of Time-Dependent Solutions

We now seek to extend the scope to linearization and linearized stability for a time-dependent solution $\bar{u}(t)$ of (219.1). We want to study solutions of the form $u(t) = \bar{u}(t) + \psi(t)$, where $\psi(t)$ is a perturbation. Using $\frac{d}{dt}\bar{u} = f(\bar{u})$ and linearizing f at $\bar{u}(t)$, we obtain

$$\frac{d}{dt}(\bar{u} + \psi)(t) = f(\bar{u}(t)) + f'(\bar{u}(t))\psi(t) + e(t),$$

with $e(t)$ quadratic in $\psi(t)$. This leads to the linearized equation

$$\dot{\varphi}(t) = A(t)\varphi(t) \quad \text{for } t > 0, \varphi(0) = \varphi_0, \tag{160.7}$$

where $A(t) = f'(\bar{u}(t))$ is an $d \times d$ matrix that now depends on t if $\bar{u}(t)$ depends on t. We have no analytical solution formula to this general problem and thus although the stability properties of the given solution $\bar{u}(t)$ are expressed through the solutions $\varphi(t)$ of the linearized problem (219.10), it may be difficult to analytically assess these properties. We may define stability factors $S(T, \varphi_0)$ and $S(T)$ just as above, and we may say that a solution $\bar{u}(t)$ is stable if $S(T)$ is moderately large, and unstable if $S(T)$ is large. To determine $S(T)$ in general, we have to use numerical methods and solve (219.10) with different initial data φ^0. We return to the computation of stability factors in the next chapter on adaptive solvers for initial value problems.

160.7 Sum Up

The question of stability of solutions to initial value problems is of fundamental importance. We can give an affirmative answer in the case of a stationary solution with corresponding symmetric Jacobian. In this case a positive eigenvalue signifies instability, with the instability increasing with increasing eigenvalue, and all eigenvalues non-positive means stability. The case of an anti-symmetric Jacobian also signifies stability with the norm of perturbations being constant in time. If the Jacobian is non-normal we have to watch out and remember that just looking at the sign of the real part of eigenvalues may be misleading: in the non-normal case algebraic growth may in fact dominate slow exponential decay for finite time. In these cases and also for time-dependent solutions, an analytical stability analysis may be out of reach and the desired information about stability may be obtained by numerical solution of the associated linearized problem.

Chapter 160 Problems

160.1. Determine the stationary solutions to the system

$$\begin{aligned}\dot{u}_1 &= u_2(1-u_1^2),\\ \dot{u}_2 &= 2-u_1u_2,\end{aligned}$$

and study the stability of these solutions.

160.2. Determine the stationary solutions to the following system (Minea's equation) for different values of $\delta > 0$ and γ,

$$\begin{aligned}\dot{u}_1 &= -u_1-\delta(u_2^2+u_3^2)+\gamma,\\ \dot{u}_2 &= -u_2-\delta u_1u_2,\\ \dot{u}_3 &= -u_3-\delta u_1u_3,\end{aligned}$$

and study the stability of these solutions.

160.3. Determine the stationary solutions of the system (219.1) with (a) $f(u) = (u_1(1-u_2), u_2(1-u_1))$, (b) $f(u) = (-2(u_1-10)+u_2\exp(u_1), -2u_2-u_2\exp(u_1))$, (c) $f(u) = (u_1+u_1u_2^2+u_1u_3^2, -u_1+u_2-u_2u_3+u_1u_2u_3, u_2+u_3-u_1^2)$, and study the stability of these solutions.

160.4. Determine the stationary solutions of the system (219.1) with (219.1) with (a) $f(u) = (-1001u_1+999u_2, 999u_1-1001u_2)$, (b) $f(u) = (-u_1+3u_2+5u_3, -4u_2+6u_3, u_3)$, (c) $f(u) = (u_2, -u_1-4u_2)$, and study the stability of these solutions.

160.5. Analyze the stability of the following variant of the linearized problem (219.8) with $\epsilon > 0$ small,

$$\dot{\varphi}(t) = \begin{pmatrix}-\nu & \kappa\\ \epsilon & -\nu\end{pmatrix}\varphi(t) \equiv A_{\nu,\kappa,\epsilon}\varphi(t) \quad \text{for } t>0, \quad \varphi(0)=\varphi^0, \tag{160.8}$$

by diagonalizing the matrix $\equiv A_{\nu,\kappa,\epsilon}$. Note that the diagonalization degenerates as ϵ tends to zero (that is, the two eigenvectors become parallel). Check if $A_{\nu,\kappa,\epsilon}$ is a normal or non-normal matrix.

161
Time Discretization by FEM

> On two occasions I have been asked (by members of Parliament), "Pray, Mr Babbage, if you put into the machine wrong figures, will the right answer come out?". I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question. (Babbage (1792-1871))

161.1 Introduction

FEM is used not only for discretization in space, but also for discretization in time, or time-stepping. Altogether, FEM is a methodology for discretization in space-time.

FEM for time discretization comes in two forms:

- *Discontinuous Galerkin* with polynomials of degree q: dG(q), $q = 0, 1, 2, ...,$
- *Continuous Galerkin* with polynomials of degree q: cG(q), $q = 1, 2, ...$

For an IVP of the basic form

$$\begin{aligned} &\dot{u}(t) = f(u(t)), \quad \text{for } t \in (0, T], \\ &u(0) = u^0, \end{aligned} \tag{161.1}$$

the basic methods dG(0) and cG(1) take the form dG(0) or *Backward*

Euler:

$$u^{n+1} = u^n + f(u^{n+1}, ndt)dt \tag{161.2}$$

where $u(t) = u^{n+1}$ is constant on $(ndt, (n+1)dt]$, and piecewise constant discontinuos on $[0, T]$.

Midpoint method cG(1):

$$u^{n+1} = u^n + \int_{ndt}^{(n+1)dt} f(u(t), t)dt \approx u^n + \frac{1}{2}(f(u^n) + f(u^{n+1}))dt \tag{161.3}$$

where $u(t)$ is linear on $[ndt, (n+1)dt]$, and piecewise linear continuous on $[0, tT]$.

Both methods are *implicit* requiring the solution of a system of equations in u^{n+1}, since u^{n+1} appears on the right hand side.

We compare with the basic *explicit Forward Euler* method:

$$u^{n+1} = u^n + f(u^n, ndt)dt \tag{161.4}$$

where u^{n+1} is directly updated by evaluating $f(u^n)$.

161.2 Read More

- Scalar IVP
- System IVP

161.3 Adaptive Error Control

For simplicity we have so far assumed the time step to be constant, but this is not economical.

In this chapter, we discuss the important issue of *adaptive error control* for numerical methods for initial value problems. This is the subject of automated choice of the time step with the purpose of controlling the numerical error to within a given tolerance level. The basic idea is to combine *feed-back* information from the computation concerning the *residual* of the computed solution and the results of auxiliary computations of *stability factors*. We focus first on on the cG(1) method and then comment on the backward Euler method, also referred to as dG(0), the discontinuous Galerkin method with piecewise constants.

We also discuss the application of cG(1) and dG(0) to a class of so-called *stiff* IVPs typically arising in chemical reaction modeling.

161.4 The cG(1) Method

We recall that cG(1), the continuous Galerkin method with polynomials of order 1, for the initial value problem $\dot{u}(t) = f(u(t))$ for $t > 0$, $u(0) = u^0$, with $f : \mathbb{R}^d \to \mathbb{R}^d$, takes the form

$$U(t_n) = U(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(U(t))\,dt, \quad n = 1, 2, ..., \tag{161.5}$$

where $U(t)$ is continuous piecewise linear with nodal values $U(t_n) \in \mathbb{R}^d$ at an increasing sequence of discrete time levels $0 = t_0 < t_1 < ...$, and $U(0) = u^0$. If we evaluate the integral in (220.1) with the midpoint quadrature rule, we obtain the Midpoint method:

$$U(t_n) = U(t_{n-1}) + k_n f\left(\frac{U(t_n) + U(t_{n-1})}{2}\right), \quad n = 1, 2, ..., \tag{161.6}$$

where $k_n = t_n - t_{n-1}$ is the time step. The cG(1)-method is the first in a family of cG(q)-methods with $q = 1, 2, ...,$, where the solution is approximated by continuous piecewise polynomials of order q. The Galerkin "orthogonality" of cG(1) is expressed by the fact that the method can be formulated

$$\int_{t_{n-1}}^{t_n} (\dot{U}(t) - f(U(t))) \cdot v\,dt = 0, \quad n = 1, 2, ..., \tag{161.7}$$

for all $v \in \mathbb{R}^d$. This says that the *residual*

$$R(U(t) = \dot{U}(t) - f(U(t)), \quad t \in [0, T], \tag{161.8}$$

of the continuous piecewise linear approximate solution $U(t)$ is *orthogonal* to the constant functions $v(t) = v \in \mathbb{R}^d$ on each subinterval (t_{n-1}, t_n). The residual $\dot{u}(t) - f(u(t))$ of the exact solution is zero since $\dot{u}(t) = f(u(t))$, while the residual of $R(U(t))$ of the approximate solution $U(t)$ is non-zero in general. Similarly, in cG(q) the residual is orthogonal on (t_{n-1}, t_n) to polynomials of degree $q - 1$. Note that (220.1) is a vector equation that reads

$$U_i(t_n) = U_i(t_{n-1}) + \int_{t_{n-1}}^{t_n} f_i(U(t))\,dt, \quad n = 1, 2, ..., \; i = 1, ..., d,$$

as can be seen from (220.3) upon setting $v = e_i$, $i = 1, ..., d$.

We will now study the problem of *automatic step-size control* with the purpose of keeping the error

$$\|u(T) - U(T)\| \leq TOL,$$

where $T = t_N$ is a final time and TOL is a given tolerance, while using as few time steps as possible. The objective is the same as that of computing an integral over an interval $[0, T]$ using numerical quadrature to a certain tolerance using as few quadrature points as possible. This is exactly the problem we meet in the case of a scalar initial value problem $\dot{u}(t) = f(u(t), t)$ with $f(u(t), t) = f(t)$.

We shall derive an *a posteriori error estimate* in which the final error $\|u(T)-U(T)\|$ is estimated in terms of the residual $R(U(t) = \dot{U}(t)-f(U(t))$ and certain *stability factors* that measure the *accumulation* of the numerical errors introduced in each time step.

The a posteriori error estimate takes the form

$$\|u(T) - U(T)\| \leq S_c(T) \max_{0\leq t\leq T} \|k(t)R(U(t)\|, \tag{161.9}$$

where $k(t) = k_n = t_n - t_{n-1}$ for $t \in [t_{n-1}, t_n)$ and where the stability factor $S_c(T)$ is defined as follows. We consider the linearized problem

$$-\dot{\varphi}(t) = A^\top(t)\varphi(t) \quad \text{for } 0 < t < T,\ \varphi(T) = \varphi^0, \tag{161.10}$$

where

$$A(t) = \int_0^1 f'(su(t) + (1 - s)U(t))\, ds.$$

We note that replacing $u(t)$ by $U(t)$ gives the following approximate formula for $A(t)$,

$$A(t) \approx f'(U(t)),$$

assuming $U(t)$ is close to $u(t)$. We conclude that $A(t)$ is close to the Jacobian $f'(u(t))$ of $f(v)$ at $v = u(t)$ if $U(t)$ is a reasonable approximation of $u(t)$ Note that the dual $A^\top(t)$ of $A(t)$ occurs in (220.6). Note further that the linearized dual problem (220.6) runs *backward* in time since the initial value $\varphi(T) = \varphi^0$ is specified at time $t = T$. We are now ready to introduce the following stability factors:

$$\begin{aligned} S_d(T) &= \max_{\varphi^0\in\mathbb{R}^d} \frac{\|\varphi(0)\|}{\|\varphi^0\|}, \\ S_c(T) &= \max_{\varphi^0\in\mathbb{R}^d} \frac{\int_0^T \|\dot{\varphi}(s)\|\, ds}{\|\varphi^0\|}, \end{aligned} \tag{161.11}$$

where φ solves (220.6). We note that the stability factors measure different features of the the dual solution φ. The stability factor $S_d(t)$ measures the maximal perturbation growth over the time interval $[0, T]$. We met this factor in the previous chapter. We shall see that this factor is tailored to measure the effect of an error in the initial data u^0 and the "d" in S_d refers to "data". The stability factor $S_c(t)$ measures the integral of $\|\dot{\varphi}\|$ over $[0, T]$ and is geared to evaluate the error in cG(1) and the "c" in S_c refers to "computation".

We shall give the proof of (220.5) below, first in a very simple case with $n = 1$ and $f(u(t)) = au(t)$ with a a constant and then in the general case. The proofs are very similar. Before plunging into the proofs, we shall try to digest the a posteriori error estimate, and see how it can be used to design an adaptive algorithm aiming at controlling the final error $\|u(T) - U(T)\|$ on a given tolerance level with as few time steps as possible.

The stability factors $S_c(T)$ and $S_d(T)$ can be computed by numerically solving the linearized dual problem (220.6) with $\varphi^0 = e_i$ for $i = 1, ..., d$. If d is large, then we may reduce the variation of the initial data by limiting the error control to certain components only, or by trying to choose φ^0 parallel to $u(T) - U(T)$, which we approximate as $U_h(T) - U_H(T)$ with $U_h(T)$ and $U_H(T)$ being approximations computed with two different tolerances.

161.5 Adaptive Time Step Control for cG(1)

We recall the basic error estimate (220.5):

$$\|u(T) - U(T)\| \leq S_c(T) \max_{0 \leq t \leq T} \|k(t)R(t)\|, \tag{161.12}$$

where $R(t) = \dot{U}(t) - f(U(t))$ and we assume that the stability factor $S(T)$ has been computed or estimated. We will return to this issue below. To achieve $\|u(T) - U(T)\| \leq TOL$, we use (220.5) to choose the time steps $k_n = t_n - t_{n-1}$ so that

$$k(t) = k_n \approx \frac{TOL}{S_c(T)R_n} \quad \text{for } t \in [t_{n-1}, t_n), \tag{161.13}$$

where

$$R_n = \max_{t_{n-1} \leq t \leq t_n} \|\dot{U}(t) - f(U(t))\|$$

is the residual on the time interval $[t_{n-1}, t_n)$. Note that the residual R_n is computable from the computed solution $U(t)$ and if $S_c(T)$ is known, timestepalg gives an equation for the time step $k_n = t_n - t_{n-1}$, where t_{n-1} already known. As with adaptive numerical quadrature, (220.9) yields a nonlinear equation for the time step $k_n = t_n - t_{n-1}$ that we can seek to solve using some form of trial-and-error strategy or by prediction, e.g. replacing R_n by R_{n-1}.

161.6 Analysis of cG(1) for a Linear Scalar IVP

We shall now prove an a posteriori error estimate for cG(1) for a a linear scalar IVP of the form

$$\dot{u}(t) = au(t) + f(t) \quad \text{for } t > 0, u(0) = u^0, \tag{161.14}$$

where a is a constant and $f(t)$ is a given function. The analysis is based on representing the error in terms of the solution $\varphi(t)$ of the following dual problem:

$$\begin{cases} -\dot{\varphi} = a\varphi & \text{for } T > t \geq 0, \\ \varphi(T) = e(T), \end{cases} \tag{161.15}$$

where $e = u - U$. Note again that (220.11) runs "backwards" in time starting at time t_N and that the time derivative term $\dot{\varphi}$ has a minus sign. We start from the identity

$$|e(T)|^2 = |e(T)|^2 + \int_0^T e\,(-\dot{\varphi} - a\varphi)\,dt,$$

and integrate by parts to get the following representation of $|e(T)|^2$,

$$|e(T)|^2 = \int_0^T (\dot{e} - ae)\varphi\,dt + e(0)\varphi(0),$$

where we allow $U(0)$ to be different from $u(0)$, corresponding to an error in the initial value $u(0)$. Since u solves the differential equation (220.10), that is $\dot{u} + au = f$, we have

$$\dot{e} - au = \dot{u} - au - \dot{U} + aU = f - \dot{U} + aU,$$

and thus we obtain the following representation of the error $|e(T)|^2$ in terms of the residual $R(U) = \dot{U} - aU - f$ and the dual solution φ ,

$$|e(T)|^2 = \int_0^T (f + aU - \dot{U})\varphi\,dt + e(0)\varphi(0) = -\int_0^{t_N} R(U)\varphi\,dt + e(0)\varphi(0). \tag{161.16}$$

Next, we use the Galerkin orthogonality of cG(1),

$$\int_{t_{n-1}}^{t_n} R(U)\,dt = 0 \quad \text{for } n = 1, 2, ...,$$

to rewrite (220.12) as

$$e(T)^2 = -\int_0^T R(U)(\varphi - \bar{\varphi})\,dt + e(0)\varphi(0), \tag{161.17}$$

where $\bar{\varphi}$ is the mean-value of φ over each time interval, that is

$$\bar{\varphi}(t) = \frac{1}{k_n}\int_{t_{n-1}}^{t_n} \varphi(s)\,ds \quad \text{for } t \in [t_{n-1}, t_n).$$

We shall now use

$$\int_{I_n} |\varphi - \bar{\varphi}|\,dt \leq k_n \int_{I_n} |\dot{\varphi}|\,dt,$$

which follows by integration from the facts that

$$\varphi(t) - \bar{\varphi}(t) = \frac{1}{k_n} \int_{t_{n-1}}^{t_n} (\varphi(t) - \varphi(s))\, ds,$$

and

$$|\varphi(t) - \varphi(s)| \le \int_s^t |\dot{\varphi}(\sigma)|\, d\sigma \le \int_{t_{n-1}}^{t_n} |\dot{\varphi}(\sigma)|\, d\sigma \quad \text{for } s, t \in [t_{n-1}, t_n].$$

Thus, (220.13) implies

$$\begin{aligned} |e(T)|^2 &\le \sum_{n=1}^N R_n \int_{I_n} |\varphi - \bar{\varphi}| dt + |e(0)||\varphi(0)| \\ &\le \sum_{n=1}^N k_n\, R_n \int_{I_n} |\dot{\varphi}| dt + |e(0)||\varphi(0)|, \end{aligned} \tag{161.18}$$

where

$$R_n = \max_{t_{n-1} \le t \le t_n} |R(U(t))|.$$

Bringing out the max of $k_n R_n$ over n, we get

$$|e(T)|^2 \le \max_{1 \le n \le N} k_n R_n \int_0^{t_N} |\dot{\varphi}|\, dt + |e(0)||\varphi(0)|.$$

Recalling that $\varphi(T) = e(T)$ and using the definitions of $S_c(t_N)$ and $S_d(t_N)$, we get the following final estimate,

$$|e(T)| \le S_c(T) \max_{0 \le t \le T} |k(t) R(U(t))| + S_d(T)|e(0)|.$$

The stability factors $S_c(T)$ and $S_d(T)$ measure the effects of the accumulation of error in the approximation. To give the analysis a quantitative meaning, we have to give a quantitative bound of this factor. The following lemma gives an estimate for $S_c(T)$ and $S_d(T)$ in the cases $a \le 0$ and the case $a \ge 0$ with possibly vastly different stability factors. We notice that the solution $\varphi(t)$ of (220.11) is given by the explicit formula

$$\varphi(t) = e(T) \exp(a(T - t)).$$

We see that if $a \le 0$, then the solution $\varphi(t)$ decays as t decreases from T, and the case $a \le 0$ is thus the "stable case". If $a > 0$ then the exponential factor $\exp(aT)$ enters, and depending on the size of a this case is "unstable". More precisely, we conclude directly from the explicit solution formula that

Lemma 161.1 *The stability factors $S_c(T)$ and $S_d(T)$ satisfy if $a > 0$,*

$$S_d(T) \le \exp(aT), \quad S_c(T) \le \exp(aT), \tag{161.19}$$

and if $a \le 0$, then

$$S_d(T) \le 1, \quad S_c(T) \le 1. \tag{161.20}$$

161.7 Analysis of cG(1) for a General IVP

The extension of the a posteriori error analysis to a general IVP $\dot{u} = f(u)$ with $f : \mathbb{R}^d \to \mathbb{R}^d$ goes as follows. We recall that the linearized dual problem takes the form

$$-\dot{\varphi}(t) = A^\top(t)\varphi(t) \quad \text{for } 0 < t < T,\ \varphi(T) = e(T), \tag{161.21}$$

with

$$A(t) = \int_0^1 f'(su(t) + (1-s)U(t))\,ds,$$

where $u(t)$ is the exact solution and $U(t)$ the approximate solution. We now use the fact that

$$\begin{aligned} A(t)e(t) &= \int_0^1 f'(su(t) + (1-s)U(t))e(t)\,ds \\ &= \int_0^1 \frac{d}{ds} f(su(t) + (1-s)U(t))\,ds = f(u(t)) - f(U(t)), \end{aligned} \tag{161.22}$$

where we used the Chain rule and the Fundamental Theorem of Calculus. We start from the identity

$$\|e(T)\|^2 = \|e(T)\|^2 + \int_0^T e \cdot (-\dot{\varphi} - A^\top\varphi)\,dt,$$

and integrate by parts to get the error representation,

$$\|e(T)\|^2 = \int_0^T (\dot{e} - Ae) \cdot \varphi\,dt + e(0) \cdot \varphi(0),$$

where we allow $U(0)$ to be different from $u(0)$, corresponding to an error in the initial value $u(0)$. Since u solves the differential equation $\dot{u} - f(u) = 0$, (220.18) implies

$$\dot{e} - Ae = \dot{u} - f(u) - \dot{U} + f(U) = -\dot{U} + f(U),$$

and thus we obtain the following representation of the error $\|e(T)\|^2$ in terms of the residual $R(U) = \dot{U} - f(U)$ and the dual solution φ,

$$\|e(T)\|^2 = -\int_0^{t_N} R(U)\varphi\,dt + e(0)\varphi(0). \tag{161.23}$$

From this point, the proof proceeds just as in the scalar case considered above and we thus obtain the following a posteriori error estimate

$$\|e(T)\| \le S_c(T) \max_{0\le t\le T} \|k(t)R(U(t))\| + S_d(T)\|e(0)\|,$$

which can be used a basis for adaptive time step control as described above. The stability factors $S_c(T)$ and $S_d(T)$ may be estimated by solving the dual problem with suitable initial data. The proof of the a posteriori error estimate shows that the stability factors may be defined by

$$\begin{aligned} S_d(T) &= \frac{\|\varphi(0)\|}{\|e(T)\|}, \\ S_c(T) &= \frac{\int_0^T \|\dot{\varphi}(s)\|\,ds}{\|e(T)\|}, \end{aligned} \tag{161.24}$$

where φ solves the linearized dual problem with initial data $\varphi(T) = e(T)$. As indicated, to compute the stability factors $S_d(T)$ and $S_c(T)$, we may solve the dual problem with some estimation of $e(T)$ obtained by solving the initial value problem with two tolerances and approximating $e(T)$ by the difference of the corresponding approximate solutions. Alternatively, choosing $\varphi(T) = e_i$, we obtain a posteriori error control for error component $e_i(T)$. If d is not large, we may this way control all components of the error, and if d is large, we may choose a couple different i at random.

The size of the stability factors indicate the degree of stability of the solution $u(t)$ being computed. If the stability factors are large, the residuals $R(U(t)$ and $e(0)$ have to be made correspondingly smaller by choosing smaller time steps and the computational problem is more demanding.

161.8 Analysis of Backward Euler for a General IVP

We now derive an a posteriori error estimate for the backward Euler method for the IVP (219.1):

$$U(t_n) = U(t_{n-1}) + k_n f(U(t_n)), \quad n = 1, 2, ..., N, \quad U(0) = u^0.$$

We associate a function $U(t)$ defined on $[0, T]$ to the function values $U(t_n)$, $n = 0, 1, ..., N$, as follows:

$$U(t) = U(t_n) \quad \text{for } t \in (t_{n-1}, t_n].$$

In other words, $U(t)$ is piecewise constant on $[0, T]$ and takes the value $U(t_n)$ on $(t_{n-1}, t_n]$, and thus takes a jump from the value $U(t_{n-1})$ to the value $U(t_n)$ at the time level t_{n-1}.

We can now write the backward Euler method as,

$$U(t_n) = U(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(U(t))\,dt,$$

or equivalently

$$U(t_n)\cdot v = U(t_{n-1})\cdot v + \int_{t_{n-1}}^{t_n} f(U(t))\cdot v\,dt, \tag{161.25}$$

for all $v\in\mathbb{R}^d$. This method os also referred to as dG(0), that is the *discontinuous Galerkin method of order zero, corresponding to approximating the exact solution by a piecewise constant function $U(t)$ satisfying the orthogonality condition* (220.21).

We are now ready to derive an a posteriori error estimate following the same strategy as for the cG(1) method. We start from the identity

$$\|e(T)\|^2 = \|e(T)\|^2 + \sum_{n=1}^{N}\int_{t_{n-1}}^{t_n} e\cdot(-\dot\varphi - A^\top\varphi)\,dt,$$

and integrate by parts on each subinterval (t_{n-1},t_n) to get the following error representation,

$$\begin{aligned}\|e(T)\|^2 = &\sum_{n=1}^{N}\int_{t_{n-1}}^{t_n}(\dot e - Ae)\cdot\varphi\,dt\\ &-\sum_{n=2}^{N-1}(U(t_n)-U(t_{n-1}))\varphi(t_{n-1}),\end{aligned}$$

where the last term results from the jumps of $U(t)$ at the the nodes $t=t_{n-1}$ and we assume $U(0)=u(0)$ for simplicity. Since u solves the differential equation $\dot u - f(u)=0$, (220.18) and the fact that $\dot U$ on (t_{n-1},t_n) imply

$$\dot e - Ae = \dot u - f(u) - \dot U + f(U) = -\dot U + f(U) = f(U)\quad\text{on } (t_{n-1},t_n),$$

and thus we obtain

$$\|e(T)\|^2 = -\sum_{n=2}^{N-1}(U(t_n)-U(t_{n-1}))\varphi(t_{n-1}) + \int_0^{t_N} f(U)\varphi\,dt.$$

Using (220.21) with $v=\bar\varphi$, the mean value of φ as above, we get

$$\begin{aligned}\|e(T)\|^2 = -\sum_{n=2}^{N-1}(U(t_n)-U(t_{n-1}))\cdot(\varphi(t_{n-1})-\bar\varphi(t_{n-1}))\\ +\sum_{n=1}^{n}\int_{t_{n-1}}^{t_n} f(U)(\varphi-\bar\varphi)\,dt.\end{aligned}$$

We note that

$$\int_{t_{n-1}}^{t_n} f(U)(\varphi-\bar\varphi)\,dt = 0,$$

since $f(U(t))$ is constant on $(t_{n-1}, t_n]$, and $\bar{\varphi}$ is the mean value of φ, and thus the error representation takes the final form

$$\|e(T)\|^2 = -\sum_{n=2}^{N-1} (U(t_n) - U(t_{n-1})) \cdot (\varphi(t_{n-1}) - \bar{\varphi}(t_{n-1})).$$

Using

$$\|\varphi(t_{n-1}) - \bar{\varphi}(t_{n-1})\| \leq \int_{t_{n-1}}^{t_n} \|\dot{\varphi}(t)\|\, dt,$$

we obtain the following a posteriori error estimate for the backward Euler method ,

$$\|e(T)\| \leq S_c(T) \max_{1\leq n\leq N} \|U(t_n)) - U(t_{n-1}))\|. \tag{161.26}$$

Note the very simple form of this estimate involving the jumps $\|U(t_n)) - U(t_{n-1}))\|$ playing the role the residual. The a posteriori error estimate (220.22) can be used as a basis for an algorithm for adaptive time step control of the following form: for $n = 1, 2, ...$, choose k_n so that

$$\|U(t_n)) - U(t_{n-1}))\| \approx \frac{TOL}{S_c(T)}.$$

161.9 Stiff Initial Value Problems

A *stiff* initial value problem $\dot{u} = f(u)$ may be characterized by the fact that the stability factors $S_d(T)$ and $S_c(T)$ are of moderate size even for large T, while the norm of the linearized operator $f'(u(t))$ is large, that is the Lipschitz constant L_f is very large. Such initial value problems are common for example in models of chemical reaction with reactions on a range of time scales from slow to fast. Typical solutions include so-called *transients* where the fast reactions make the solution change quickly over a short (initial) time interval, after which the fast reactions are "burned out" and the slow reactions make the solution change on a longer time scale.

The prototype of a stiff initial value problem has the form

$$\dot{u} = f(u) \equiv -Au \quad \text{for } t > 0,\ u(t) = u^0 = (u_i^0), \tag{161.27}$$

where A is a constant symmetric positive semidefinite $d \times d$ matrix with non-negative eigenvalues λ_i ranging from zero to large positive values. Accordingly, the norm of the matrix A is large and L_f is large. By diagonalization, we may reduce to the case when A is a diagonal matrix with non-negative diagonal elements λ_i, in which case the solution $u(t) = (u_i(t))$ is given by

$$u_i(t) = \exp(-\lambda_i\, t)u_i^0 \quad \text{for } t > 0, \tag{161.28}$$

with $u^0 = (u_i^0)$. This explicit solution formula shows that a component $u_i(t)$ corresponding to a large positive eigenvalue λ_i decays very quickly to zero, while a component with a small eigenvalue stays almost constant for a long time and eventually decays to zero. The sign of the eigenvalues is evidently crucial: if some λ_i was negative, then the corresponding solution component would explode exponentially more or less quickly depending on the size of λ_i. In particular, (220.24) with the λ_i non-negative implies

$$\|u(t)\| \leq \|u^0\| \quad \text{for } t > 0, \tag{161.29}$$

which indicates a form of stability with stability factor equal to 1 in the sense that the norm of the solution does not increase in time.

The dual problem corresponding to (220.23) takes the form

$$-\dot{\varphi} + A\varphi = 0 \quad \text{for } T > t > 0,\ \varphi(T) = \psi,$$

with ψ given data at time $t = T$. As a counterpart of (220.25), we conclude that $S_d(T) \leq 1$. We can similarly show that $S_c(T)$ grows very slowly with increasing T. We sum up: (220.23) represents a stiff problem; stability factors are of moderate size even for large T while the norm of the (linearized) operator A is large.

From numerical point of view, stiff problems may seem particularly friendly since the stability factors grow very slowly with time, but there is one hook that has attracted a lot of attention in the literature on numerical methods for initial value problems, namely the failure of an explicit method like the forward Euler method. We write this method for the equation $\dot{u} = -Au$ in the form

$$U^n = U^{n-1} - k_n A U^{n-1}$$

with U^n an approximation of $u(t_n)$ and $0 = t_0 < t_1 < ...$ an increasing sequence of time levels, and $k_n = t_n - t_{n-1}$. If A is diagonal with diagonal elements $\lambda_i \geq 0$, then

$$U_i^n = (1 - k_n\lambda_i)U_i^{n-1}$$

and if λ_i is large positive, then $|1 - k_n\lambda_i|$ may be much larger than 1 unless the time step k_n is sufficiently small ($k_n \leq 2/|\lambda_i|$ for all i) and the the numerical solution will then quickly explode to infinity, while the corresponding exact solution quickly decays to zero. The explicit Euler method will thus give completely wrong results unless sufficiently small time steps are used. This may lead to very inefficient time-stepping since after the transients have died out, the solution may vary only slowly and large time steps would be desirable. We note that the time step limit $k_n \leq 2/|\lambda_i|$ for all i, is set by the largest eigenvalue $\max \lambda_i$, while the time long-time scale is set by the smallest eigenvalue $\min \lambda_i$, so that if the quotient $\max \lambda_i / \min \lambda_i$ is large (which signifies a stiff problem), then explicit Euler would be inefficient outside transients.

On the other hand, the dG(0), or implicit Euler method,

$$U^n + k_n A U^n = U^{n-1}$$

with

$$U_i^n = (1 + k_n \lambda_i)^{-1} U_i^{n-1}$$

will be stable and work very well without step size limitation because $1 + k_n \lambda_i \geq 1$ for all $\lambda_i \leq 0$.

For the cG(1)-method, we will have

$$U_i^n = \frac{1 - k_n \lambda_i}{1 + k_n \lambda_i} U_i^{n-1}$$

and stability prevails because

$$|\frac{1 - k_n \lambda_i}{1 + k_n \lambda_i}| \leq 1$$

for all $\lambda_i \geq 0$.

We conclude that both dG(0) and cG(1) may be used for stiff problems, but both these methods are implicit and require the solution of system of equations at each time step. More precisely, dG(0) for a problem of the form $\dot{u} = f(u)$ takes the form

$$U^n - k_n f(U^n) = U^{n-1}.$$

At each time step we have to solve an equation of the form $v - k_n f(v) = U^{n-1}$ with U^{n-1} given. To this end we may try a damped fixed point iteration in the form

$$v^{(m)} = v^{(m-1)} - \alpha(v^{(m-1)} - k_n f(v^{(m-1)}) - U^{n-1}),$$

with α some suitable matrix (or constant in the simplest case). Choosing $\alpha = I$, and iterating once with $v^0 = 0$ corresponds to the explicit Euler method. Convergence of the fixed point iteration requires that

$$\|I + k_n \alpha f'(v)\| < 1$$

for relevant values of v, which could force α to be small (e.g. in the stiff case with $f'(v)$ having large negative eigenvalues) and result in slow convergence. A first try could be to choose α to be a diagonal matrix with $\alpha_i = (f'_{ii})(v^{m-1}))^{-1}$ (corresponding to *diagonal scaling*) and hope that the number of iterations would not be too large. In some cases more efficient iterative solvers would have to be used.

161.10 On Explicit Time-Stepping for Stiff Problems

We just learned that explicit time-stepping for stiff problems require small time steps outside transients and thus may be inefficient. We shall now indicate a way to get around this limitation through a process of stabilization, where a large time step is accompanied by a couple of small time steps. The resulting method has similarities with the control system of a modern (unstable) jet fighter like the Swedish JAS Gripen, the flight of which is controlled by quick small flaps of a pair of small extra wings ahead of the main wings, or balancing a stick vertically on the finger tips if we want a more domestic application.

We shall now explain the basic (simple) idea of the stabilization and present some examples, as illustrations of fundamental aspects of adaptive IVP-solvers and stiff problems. Thus to start with, suppose we apply explicit Euler to the scalar problem

$$\begin{aligned} \dot{u}(t) + \lambda u(t) &= 0 \quad \text{for } t > 0, \\ u(0) &= u^0, \end{aligned} \tag{161.30}$$

with $\lambda > 0$ taking first a large time step K satisfying $K\lambda > 2$ and then m small time steps k satisfying $k\lambda < 2$, to get the method

$$U^n = (1 - k\lambda)^m (1 - K\lambda) U^{n-1}, \tag{161.31}$$

altogether corresponding to a time step of size $k_n = K + mk$. Here K gives a large unstable time step with $|1 - K\lambda| > 1$ and k is a small time step with $|1 - k\lambda| < 1$. Defining the polynomial function $p(x) = (1 - \theta x)^m (1 - x)$, where $\theta = \frac{k}{K}$, we can write the method (220.27) in the form

$$U^n = p(K\lambda) U^{n-1}.$$

For stability we need

$$|p(K\lambda)| \le 1, \quad \text{that is } |1 - k\lambda|^m (K\lambda - 1) \le 1,$$

or

$$m \ge \frac{\log(K\lambda - 1)}{-\log|1 - k\lambda|} \approx 2\log(K\lambda), \tag{161.32}$$

with $c = k\lambda \approx 1/2$ for definiteness.

We conclude that m may be quite small even if $K\lambda$ is large, since the logarithm grows so slowly, and then only a small fraction of the total time would be spent on stabilizing time-stepping with the small time steps k.

To measure the efficiency gain we introduce

$$\alpha = \frac{1 + m}{K + km} \in (1/K, 1/k),$$

which is the number of time steps per unit interval with stabilized explicit Euler method, and by (220.28)) we have

$$\alpha \approx \frac{1+2\log(K\lambda)}{K+\log(K\lambda)/\lambda} \approx 2\lambda\frac{\log(K\lambda)}{K\lambda} \ll 2\lambda, \tag{161.33}$$

for $K\lambda \gg 1$. On the other hand, the number of time steps per unit interval for the usual explicit Euler is

$$\alpha_0 = 1/k = \lambda/2, \tag{161.34}$$

choosing a maximum time step $k = 2/\lambda$.

The cost reduction factor using the stabilized explicit Euler method would thus be

$$\frac{\alpha}{\alpha_0} \approx \frac{4\log(K\lambda)}{K\lambda}$$

which can be quite significant for large values of $K\lambda$.

We now present some examples using an adaptive cg(1) IVP-solver in stabilized explicit form with just a few iterations in each time step, which allows large time steps. In all problems we note the initial transient, where the solution components change quickly, and the oscillating nature of the time step sequence outside the transient with large time steps followed by some small stabilizing time steps.

EXAMPLE 161.1. We apply the indicated method to the scalar problem equation (220.26) with $u^0 = 1$ and $\lambda = 1000$, and display the result in Figure 220.1. The cost reduction factor with comparison to a standard explicit method is large: $\alpha/\alpha_0 \approx 1/310$.

EXAMPLE 161.2. We now consider the 2×2 diagonal system

$$\begin{aligned} \dot{u}(t) + \begin{pmatrix} 100 & 0 \\ 0 & 1000 \end{pmatrix} u(t) = 0 \quad \text{for } t > 0, \\ u(0) = u^0, \end{aligned} \tag{161.35}$$

with $u^0 = (1, 1)$. There are now two eigenmodes with large eigenvalues that have to be stabilized. The cost reduction is $\alpha/\alpha_0 \approx 1/104$.

EXAMPLE 161.3. The is the so-called HIRES problem ("High Irradiance RESponse") from plant physiology which consists of the following

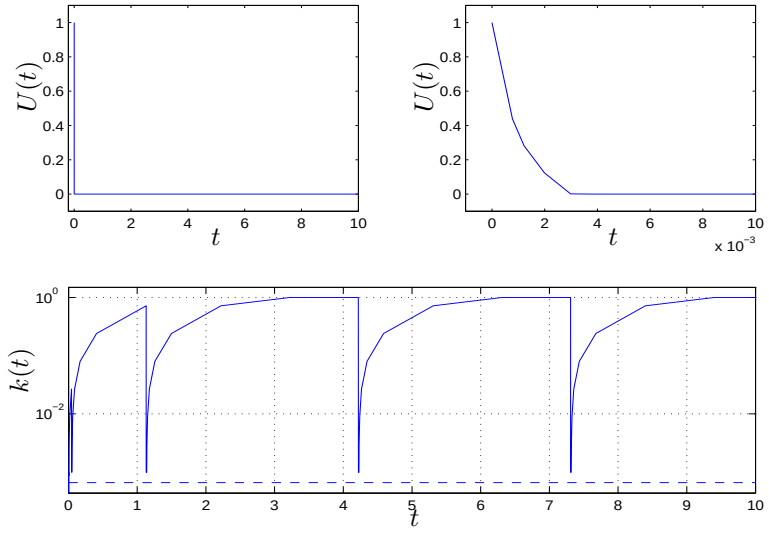


FIGURE 161.1. Solution and time step sequence for eq. (220.26), $\alpha/\alpha_0 \approx 1/310$.

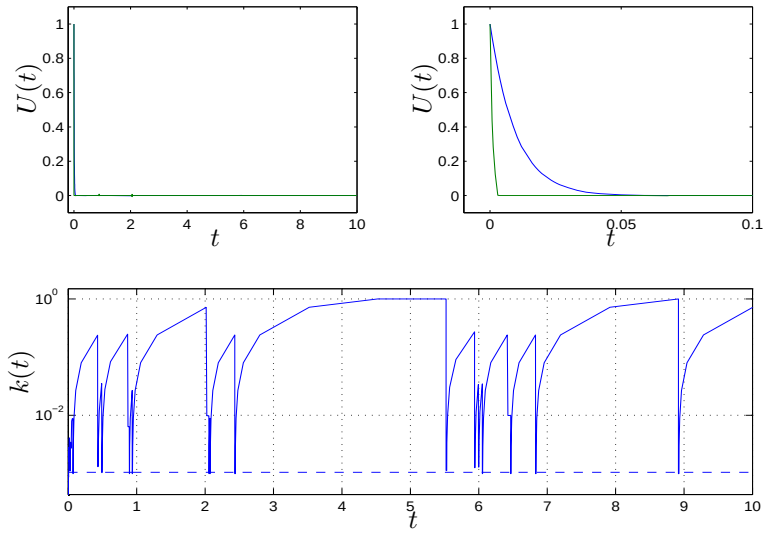


FIGURE 161.2. Solution and time step sequence for eq. (220.31), $\alpha/\alpha_0 \approx 1/104$.

eight equations:

$$\begin{cases} \dot{u}_1 &= -1.71u_1 + 0.43u_2 + 8.32u_3 + 0.0007, \\ \dot{u}_2 &= 1.71u_1 - 8.75u_2, \\ \dot{u}_3 &= -10.03u_3 + 0.43u_4 + 0.035u_5, \\ \dot{u}_4 &= 8.32u_2 + 1.71u_3 - 1.12u_4, \\ \dot{u}_5 &= -1.745u_5 + 0.43u_6 + 0.43u_7, \\ \dot{u}_6 &= -280.0u_6u_8 + 0.69u_4 + 1.71u_5 - 0.43u_6 + 0.69u_7, \\ \dot{u}_7 &= 280.0u_6u_8 - 1.81u_7, \\ \dot{u}_8 &= -280.0u_6u_8 + 1.81u_7, \end{cases} \tag{161.36}$$

together with the initial condition $u^0 = (1.0, 0, 0, 0, 0, 0, 0, 0.0057)$. We present the solution and the time step sequence in Figure 220.3. The cost is now $\alpha \approx 8$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/33$.

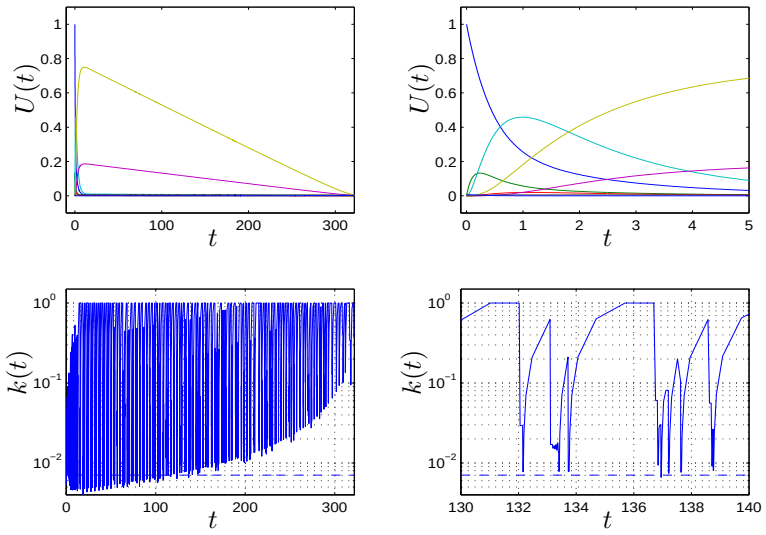


FIGURE 161.3. Solution and time step sequence for eq. (220.32), $\alpha/\alpha_0 \approx 1/33$.

EXAMPLE 161.4. The "Chemical Akzo-Nobel" problem consists of the following six equations:

$$\begin{cases} \dot{u}_1 &= -2r_1 + r_2 - r_3 - r_4, \\ \dot{u}_2 &= -0.5r_1 - r_4 - 0.5r_5 + F, \\ \dot{u}_3 &= r_1 - r_2 + r_3, \\ \dot{u}_4 &= -r_2 + r_3 - 2r_4, \\ \dot{u}_5 &= r_2 - r_3 + r_5, \\ \dot{u}_6 &= -r_5, \end{cases} \tag{161.37}$$

where $F = 3.3 \cdot (0.9/737 - u_2)$ and the reaction rates are given by $r_1 = 18.7 \cdot u_1^4 \sqrt{u_2}$, $r_2 = 0.58 \cdot u_3 u_4$, $r_3 = 0.58/34.4 \cdot u_1 u_5$, $r_4 = 0.09 \cdot u_1 u_4^2$ and $r_5 = 0.42 \cdot u_6^2 \sqrt{u_2}$. We integrate over the interval $[0, 180]$ with initial condition $u^0 = (0.437, 0.00123, 0, 0, 0, 0.367)$. Allowing a maximum time step of $k_{\max} = 1$ (chosen arbitrarily), the cost is only $\alpha \approx 2$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/9$. The actual gain in a specific situation is determined by the quotient between the large time steps and the small damping time steps, as well as the number of small damping steps that are needed. In this case the number of small damping steps is small, but the large time steps are not very large compared to the small damping steps. The gain is thus determined both by the stiff nature of the problem and the tolerance (or the size of the maximum allowed time step).

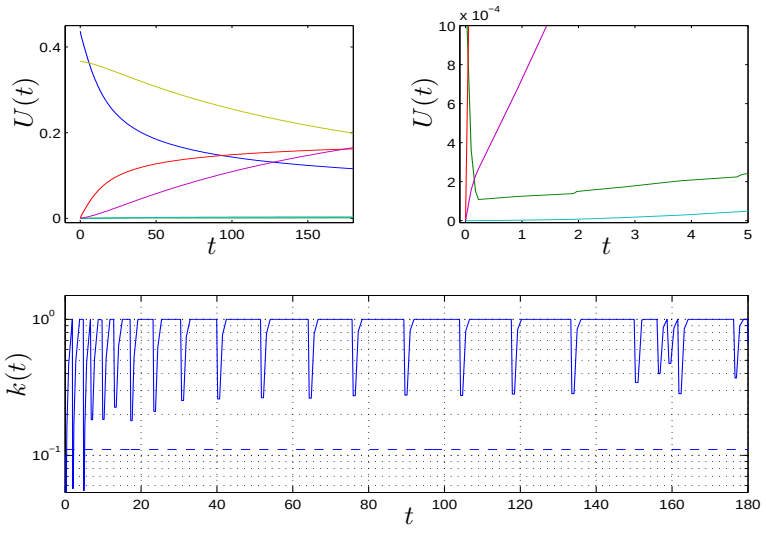


FIGURE 161.4. Solution and time step sequence for eq. (220.33), $\alpha/\alpha_0 \approx 1/9$.

EXAMPLE 161.5. We consider now Van der Pol's equation:

$$\ddot{u} + \mu(u^2 - 1)\dot{u} + u = 0,$$

which we write as

$$\begin{cases} \dot{u}_1 &= u_2, \\ \dot{u}_2 &= -\mu(u_1^2 - 1)u_2 - u_1. \end{cases} \tag{161.38}$$

We take $\mu = 1000$ and solve on the interval $[0, 10]$ with initial condition $u^0 = (2, 0)$. The time step sequence behaves as desired with only a small portion of the time spent on taking small damping steps. The cost is now $\alpha \approx 140$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/75$.

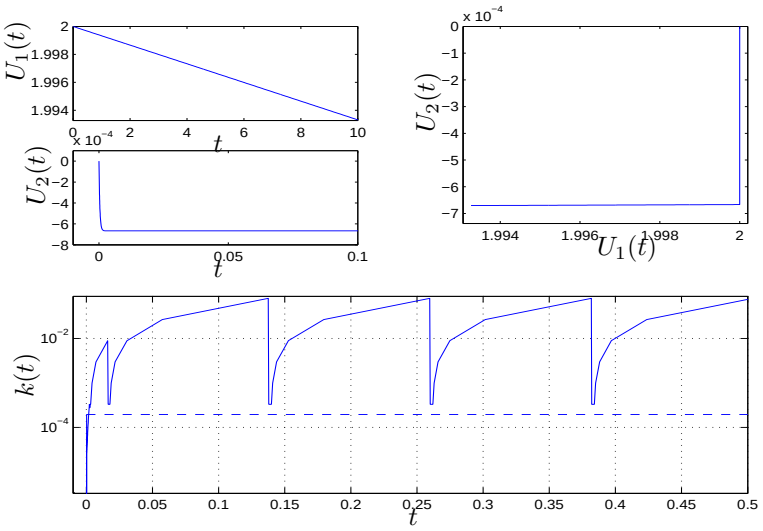


FIGURE 161.5. Solution and time step sequence for eq. (220.34), $\alpha/\alpha_0 \approx 1/75$.

Chapter 161 Problems

161.1. Compute the stability factors $S_d(T)$ and $S_c(T)$ for the linear scalar IVP $\dot{u}(t) = -\lambda(t)u(t)$ for $t > 0$, $u(0) = u^0$, where $\lambda(t)$ depends on time t and (a) $\lambda(t) \geq 0$, (b) $\lambda(t) < 0$.

161.2. Compute $S_d(T)$ and $S_c(T)$ for the linear 2×2 system $\dot{u}_1 = u_2, \quad \dot{u}_2 = -u_1$ for $t > 0$, $u(0) = u^0$.

161.3. Implement adaptive IVP-solvers based on dG(0) and cG(1) and apply the solvers to different problems.

161.4. Show that the a posteriori error estimate for cG(1) may be written on the form $\|e(T)\| \leq S_c(T)\max_{0\leq t\leq T}\|k(t)(f(U(t)) - \bar{f}(U(t)))\| + S_d(T)\|e(0)\|$, where $\bar{f}(U(t))$ is the mean-value of $f(U(t))$ over each time interval.

161.5. Show that choosing in the dual problem $\varphi(T) = e_i$ gives control of error component $e_i(T)$.

161.6. Develop explicit versions of dG(0) and cG(1) based on fixed point iteration at each time step. Show that with diagonal scaling such an explicit method may work very well for some stiff problems.

162
General Galerkin G2

> All government, indeed every human benefit and enjoyment, every virtue, and every prudent act, is founded on compromise and barter. (Edmund Burke)
>
> The art of doing mathematics consists in finding that special case which contains all the germs of generality. (David Hilbert)
>
> Mathematics as an expression of the human mind reflects the active will, the contemplative reason, and the desire for aesthetic perfection. Its basic elements are logic and intuition, analysis and construction, generality and individuality. (Richard Courant)

We have seen that FEM as Galerkin's method with piecewise polynomials requires a modification to work well for convection-dominated problems arising in fluid mechanics. Galerkin's method for a differential equation on a domain Ω written as

$$A(u) = f \tag{162.1}$$

takes the form: Find $u \in V_h$ such that

$$(A(u) - f, v) = 0 \quad \text{for all } v \in V_h, \tag{162.2}$$

where

$$(v, w) = \int_\Omega vw \, ds,$$

FIGURE 162.1. General Galerkin.

and we may think of V_h as the linear space of continuous piecewise linear functions on a triangulation of Ω.

We may compare with a *Least-Squares Method* of the form: Find $u \in V_h$ such that

$$\|A(u) - f\|^2 \leq \|A(v) - f\|^2 \quad \text{for all } v \in V_h, \tag{162.3}$$

where

$$\|v\|^2 = (v, v),$$

that is, $u \in V_h$ minimizes the residual $A(u) - f$ in the norm $\|\cdot\|$. If $A(u)$ is linear in u, then Least Squares minimization is expressed as

$$(A(u) - f, A(v)) = 0 \quad \text{for all } v \in V_h. \tag{162.4}$$

We can say that Galerkin's method seeks to make the residual $R(u) \equiv A(u) - f$ small in a *weak sense* of (162.2) asking the residual $R(u)$ to be orthogonal to V_h, while the Least Squares Method (247.3) seeks to make $R(u)$ in a strong sense of the norm $\|\cdot\|$.

It turns out that Galerkin's method is too weak for some problems and that Least Squares is too strong in general. Luckily a certain weighted combination turns out to be just right. If $A(u)$ is linear then this fortunate combination, which we refer to as *General Galerkin* or in short *G2*, takes the form: Find $u \in V_h$ such that

$$(A(u) - f, v) + \delta(A(u) - f, A(v)) = 0 \quad \text{for all } v \in V_h, \tag{162.5}$$

where $\delta \approx h$ appears as a weight on the least squares term.

For the basic convection equation $u' = f$, choosing $\delta = \frac{h}{2}$ turns the central difference quotient approximation of u' of Galerkin's method, into a G2 upwind difference quotient:

$$\frac{u_{i+1} - u_{i-1}}{2h} - \frac{h}{2}\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = \frac{u_i - u_{i-1}}{h} \tag{162.6}$$

corresponding to applying Galerkin's method to the modified equation

$$u' - \frac{h}{2}u'' = f - \frac{h}{2}f' \tag{162.7}$$

with a stabilizing diffusion term $-\frac{h}{2}u''$ on the left hand side, compensated by the corresponding force $-\frac{h}{2}f'$.

In short, G2 is compromise beween the too weak Scylla of Galerkin and the too strong Carybdis of Least Squares. For diffussion-dominated problems, Galerkin works fine (G2 with $\delta = 0$), while G2 with $\delta \approx h$ works fine for convection-dominated problems.

To automate mathematical modeling based on differential equations, we need an efficient method for discretizing the differential equations arising in applications. G2 fills this need, and serves as a basis of FEniCS.

FIGURE 162.2. Pooh contemplating variational formulation of FEM.

163
Error Control by Duality

Light and matter are both single entities, and the apparent duality arises in the limitations of our language. (Heisenberg) (Convincing?)

163.1 Galerkin Method

Consider a Galerkin method of the form: Find $u \in V_h$ such that

$$(Au, v) = (f, v) \quad \text{for all } v \in V_h, \tag{163.1}$$

where A is a differential operator, f a given forcing and $(\cdot, \cdot)$ a relevant scalar product.

163.2 Output Error

Suppose we want to estimate the error in the output (u, ψ), where ψ is a given weight function, as compared the to the output $(\bar{u}, \psi)$ of a fictitious fine-grid solution
baru with zero residual $R(\bar{u}) = A\bar{u} - f = 0$, in terms of the computable residual $R(u) = Au - f$.

163.3 Error Representation by Duality

Let φ satisfy $A^\top\varphi = \psi$ and consider the following identity:

$$(u-\bar{u},\psi) = (u-\bar{u},A^\top\varphi) = (Au-A\bar{u},\varphi) = (Au-f,\varphi) = (R(u),\varphi), \tag{163.2}$$

which proves the following error representation

$$(u-\bar{u},\psi) = (R(u),\varphi), \tag{163.3}$$

expressing the output error $(u-\bar{u},\psi)$ in terms of the residual $R(u)$ and the dual solution φ.

163.4 Galerkin Orthogonality

The Galerkin equation (163.1) expresses the *Galerkin orthogonality*

$$(R(u),v) = 0 \quad \text{for all } v \in V_h, \tag{163.4}$$

which we now use to rewrite (163.3) as a refined error representation in the form

$$(u-\bar{u},\psi) = (R(u),\varphi-\varphi_h), \tag{163.5}$$

where φ_h is an interpolant of φ. Assuming that the interpolation error $\varphi-\varphi_h$ can be estimated as

$$\|\varphi-\varphi_h\| \le C_i\|hD\varphi\|, \tag{163.6}$$

where $\|v\| = \sqrt{(v,v)}$ and $D\varphi$ measures first-order derivatives of φ and C_i is an interpolation constant, we obtain using Cauchy's inequality

$$|(u-\bar{u},\psi)| \le C_i\|D\varphi\|\|hR(u)\|. \tag{163.7}$$

If we now assume that

$$\|D\varphi\| \le S, \tag{163.8}$$

where S is referred to as a *stability constant*, then the a posteriori error estimate takes the concrete form

$$|(u-\bar{u},\psi)| \le S\|hR(u)\|, \tag{163.9}$$

where we included the interpolation constant in S.

Finally, by computing S by solving the dual problem (typically on the same mesh as that underlying V_h) and directly evaluating $\|D\varphi\|$ for the corresponding solution φ, we obtain a concrete estimate of the output error in terms of $S\|hR(u)\|$, where the presence of the factor h multiplying the residual $R(u)$ is the payoff of using a Galerkin method.

Part X

Simulators

164
Tools and Perspective

> I want to describe an unrealistic reality in a realistic way. (Fernando Botero)

> A painted landscape is always more beautiful than a real one, because there's more there. Everything is more sensual, and one takes refuge in its beauty. And man needs spiritual expression and nourishing. It's why even in the prehistoric era, people would scrawl pictures of bison on the walls of caves. Man needs music, literature, and painting-all those oases of perfection that make up art-to compensate for the rudeness and materialism of life. (Fernando Botero)

In World of Games you have constructed simulations and simulators mainly based on particle-spring models. We now expand the scope to include differential equation models using the following basic tools:

- mathematical modeling of physical phenomemna by differential equations,
- G2 for computational solution of differential equations,
- FEnICS/Unicorn as realization of G2.

Ultimately, all models of physical reality can be seen as different forms of particle-mass models, because physics ultimately consists of (some form of) particles interacting by (some form of) forces.

In particular, a differential equation model becomes a discrete particle-mass model after discretization by FEM. The advantage of using this

methodology is that the assignment of particle masses and spring constants (related to a given mesh), is automatized by G2 from a differential equation model with few material parameters.

In direct particle-spring modeling, on the other hand, spring constants and particle masses have to be specified manually, which can be very time-consuming.

Differential equation models automatically discretized by G2 thus offer a very efficient tool for simulation, because differential equations express basic laws understandable to humans often with few material parameters, yet allow a very rich output. This is nothing but Leibniz':

- *Best of all Possible Worlds = most complex world governed by most simple laws.*

We shall now illustrate some of the capabilities with focus on phenomena of fluid and stucture/solid dynamics including fluid-structure interaction. As illustration we shall consider the following activities:

1. flying (fluid, fluid-structure)
2. sailing (fluid, fluid-structure)
3. jumping (structure)
4. shooting (fluid)
5. speaking (fluid-structure acoustics)
6. predicting weather and climate.

With this inspiration you will be able to construct simulators (and related games if you want) for a very large variety of applications.

FIGURE 164.1. Combined digital and mechanical simulator.

165
Flying

165.1 To Read

- Why It Is Possible to Fly
- Why Paragliding Is Possible
- Why Birds Can Fly
- Why Bumblebees Can Fly
- Why Wingsuit Flying Is Possible

165.2 To Browse

- The Mathematical Secret of Flight
- Mathematical Theory of Flight

165.3 Watch

- Early attempts
- Wingsuit flying

- First take-off Airbus 380
- Simulated Airbus 380 landing
- Airplane spin
- Crash of JAS Gripen.

165.4 Model: Incompressible NS with Slip

Flight in air at subsonic speeds (up to say 300 km/s) is described by the incompressible Navier-Stokes equations. G2 for incompressible NS is presented in Navier-Stokes: Quick and Easy and implemented in Unicorn. It is natural to start by computing the flow of air around a fixed 3d wing to find the distribution of forces on the wing surface at different *angles of attack*. The results of the computations can be condensed into curves or tables of lift, drag and *pitching moment* as functions of the angle of attack and flight velocity.

The Reynolds number wings of airplanes (and bigger birds) is so large ($> 10^5$) that a slip boundary condition can be used as simple model for a turbulent boundary layer. Lift, drag and pitching moment can then be computed on meshes with about 100.000 mesh-points, using e.g. FEniCS/Unicorn.

165.5 Simulator Based on Lift/Drag Curves

A simple flight simulator can be designed from the lift/drag/moment curves of a single wing.

165.6 Direct Simulation

A more advanced simulator for extreme operations like take-off and landing at maximal angle of attack, and dynamics of spin et cet, requires real-time solution of the NS-equations. For this purpose you can use FEniCS/Unicorn.

165.7 Flapping Wings

Extend to flapping wing using fluid-structure modeling.

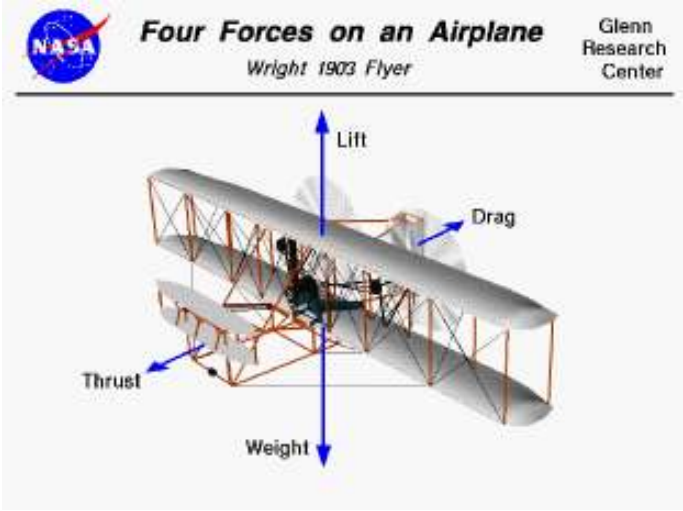


FIGURE 165.1. Forces on an airplane

FIGURE 165.2. In the cockpit.

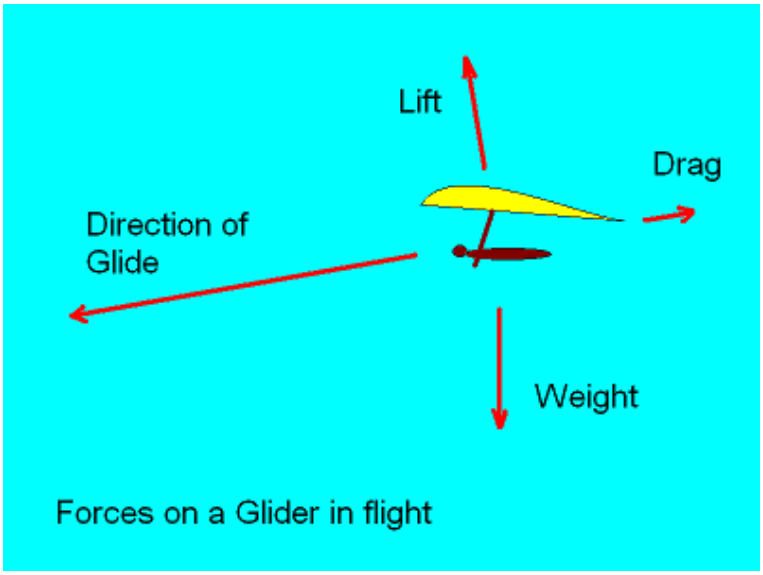


FIGURE 165.3. Forces on a glider.

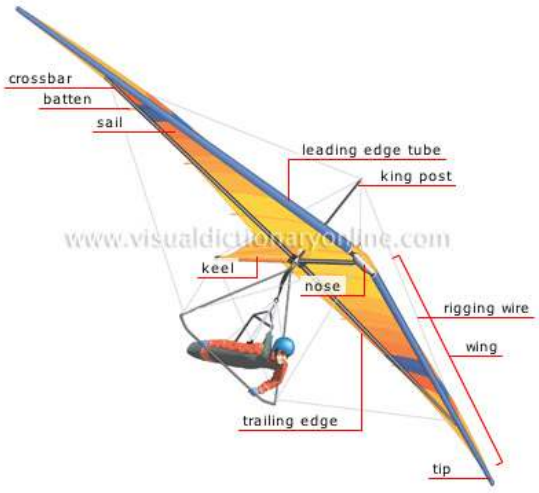


FIGURE 165.4. Hang glider.

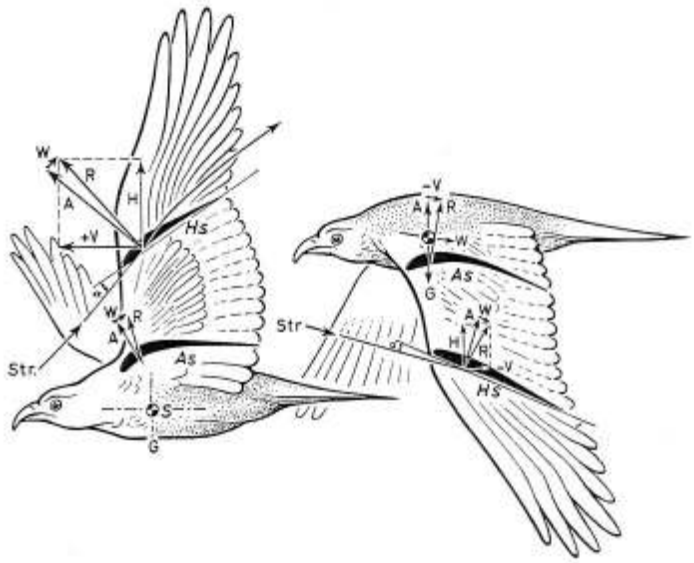


FIGURE 165.5. Forces on bird in flight.

165.8 Bird/Insect Flight

Watch:

- Why Birds Can Fly
- Bird Flight Slow Motion
- Blender Bird Wing Test

Construct a simulator for gliding and flapping flight of small and large birds and insects, using e.g. FEniCS/Unicorn for fluid-structure interaction. Determine if a bumble bee can fly.

166
Sailing

166.1 To Read

- Why It Is Possible to Sail

166.2 To Browse

- Sailing 1
- sailing 2

166.3 Watch

- Americas Cup 1930 vs 2010
- Alinghi vs Oracle 2010
- Americas Cup start tactics at Turning Torso
- Volvo Ocean Race

FIGURE 166.1. Vasa tipping over. Why?

166.4 Fluid Dynamics of Sailing

The sail and keel of a sailboat both act as wings creating lift which powers a sailboat to overcome drag from sail, keel and hull, and in particular makes it possible to sail upwind. Compare Sailing Game.

166.5 Model: Incompressible NS with Slip

Both the flow of air and water can be modeled by the incompressible Navier-Stokes equations with slip boundary condition on the sail and hull combined. Air and water can be modeled as a variable density fluid with the free water surface represented by a transition zone from high to low density (water).

166.6 Stability of Floating Bodies

The motion of the boat through the water is determined by bouyancy forces connecting to Archimedes Principle and the forward thrust from the sail.

The stability of floating bodies (in 2d say) connects to the relative horisontal motion of the of centers of gravity for the body and its submerged part under tilting from horisontal forces (e.g. wind).

The forces (and moments) acting on a floating body may be computed by direct summation (integration) of gravity forces and fluid forces acting on little pieces of the body, or from centers of gravity and bouyancy as just indicated.

166.7 Simulator

Construct a Sailing Simulator using experience from designing a flight simulator. Start using given lift/drag curves for sail and keel, and take the heeling into account using Archimedes principle. Follow up with real time direct fluid-structure simulation e.g. using FEniCS/Unicorn.

166.8 Investigations

- Determine the performance of different designs of sail and hull.
- Why did Vasa tip over? Who was responsible for the design? Was the instability a surprise? What was the verdict of the judiciary process following the catastrophy? Was the ship "well built but badly designed"?

166.9 BMW ORACLE Americas Cup 2010

The 33rd Americas Cup 2010 was won by the US challenger BMW ORACLE with rigid wing-sail, see BMW ORCACLE vs ALINGHI. Can you explain the outcome?

FIGURE 166.2. BMW ORACLE wins Americas Cup 2010.

167
Jumping and Falling

> If there be light, then there is darkness; if cold, heat; if height, depth; if solid, fluid; if hard, soft; if rough, smooth; if calm, tempest; if prosperity, adversity; if life, death. (Pythagoras)

167.1 To Watch

- Passive Fall of Cow
- Hexapod Simulator
- Walking Robot
- Walking Humanoid
- Humanoid Assistant
- High Jump Simulation

167.2 Passive Structure Dynamics

Model the passive dynamics of elastic bodies subject to gravity and contact forces by using a Navier/Lagrange model. Passive dynamics means that no interior extra forces, like muscle forces, are added.

FIGURE 167.1. Gillian Murphy overcoming gravity in *grand jete*.

167.3 Active Structure Dynamics

Model active dynamics by inserting suitable interior (muscle) forces in the above model.

167.4 Simulators

Construct a simulator for passive/active structure dynamics, e.g. using FEniCS/Unicorn.

167.5 Investigation

- simulate ballet jump
- simulate high jump.

168
Shooting

168.1 Spinning Balls

A spinning ball in flight will be subject to a lift force transversal to the flight trajectory, which will add to the gravity force and give additional curvature to the trajectory.

The flight of a spinning ball can be modeled by the (incompressible) Navier-Stokes equations with a mixture of no-slip and slip boundary conditions resulting in unsymmetric separation and lift.

Spinning balls are important in soccer, baseball, tennis, table-tennis...See Why A Topspin Tennis Ball Curves Down.

168.2 Bow and Arrow

Model the action of bow and arrrow, with an elastic model of the bow and a fluid model for the flight of the arrow.

168.3 Read More

- String Theory

169
Racing

169.1 Watch

- Panda ODE Car

169.2 Simulator

Model the dynamics of a (stock-car) vehicle using a Navier/Lagrange elasto-plastic model.

FIGURE 169.1. Cross road racing.

170
Roulette and Chaos

170.1 To Watch

- Roulette wheel simulation

170.2 Simulator

A roulette is a mechanical system with outcome which is so difficult to predict that it can be used as a random number generator or *chaos machine* in a game of chance. This is an effect of *sensitive dependence of initial conditions* which means that very small changes in the way the ball is launched in each play will change the final position of the ball. Over very many plays we can expect that each of the 37 numbers will come up approximately with a frequency of 1/37.

Construct a chaos machine e.g. as an elastic ball interacting with a jagged solid surface.

170.3 Investigation

Is the motion chaotic?

FIGURE 170.1. Roulette: Pointwise unpredictable, mean-value predicatble system.

171
Predicting Weather and Climate

171.1 To Read

- Global Circulation Models
- On Climate Sensitivity 1, 2, 3

171.2 To Watch

- Desktop climate simulation
- Supercomputer climate simulation
- Thermohaline circulation
- Global Circulation
- Atmospheric winds
- Coriolis forces.

171.3 Simulator

Construct a combined ocean-atmosphere simulator based on the Navier-Stokes equations with the ocean incompressible and the atmosphere com-

pressible subject to incoming radiative forcing and outgoing blackbody radiation.

171.4 Investigation

Seek to determine climate sensitivity as the increase of global temperature upon a doubling of CO2 in the atmosphere.

Part XI

Technology With Simulation

172
Reality of the Virtual

172.1 To Think About

- Reality of the Virtual vs Virtual Reality

173

Incompressible Navier-Stokes: Quick and Easy

> My attention was drawn to various mechanical phenomena, for the explanation of which I discovered that a knowledge of mathematics was essential. (Reynolds)
>
> By this research it is shown that there is one, and only one, conceivable purely mechanical system capable of accounting for all the physical evidence, as we know it in the Universe. (Reynolds)

173.1 Introduction

The Navier-Stokes equations is the basic model for fluid flow and describe a variety of phenomena in hydro and aero-dynamics, processing industry, biology, oceanography, geophysics, meteorology and astrophysics. Fluid flow in all these applications usually contains features of both *turbulent* and *laminar* flow, with turbulent flow being irregular with rapid fluctuations in space and time and laminar flow being more organized. The basic question of *Computational Fluid Dynamics* CFD is how to efficiently and reliably solve the Navier-Stokes equations numerically for both laminar and turbulent flow.

The Navier-Stokes equations is a system of nonlinear differential equations coupling the phenomena of convection and diffusion. Traditionally, the study of the Navier-Stokes equations is separated into *incompressible* and *compressible* flow, using different dependent variables: *primitive variables* (velocity, pressure, temperature) for incompressible flow and *conservation variables* (density, momentum, energy) for compressible flow. We focus in

this chapter on the incompressible Navier-Stokes equations in the case of constant density, viscosity and temperature, with the velocity and pressure as variables. We present the cG(1)dG(0) finite element method with cG(1) in space and dG(0) in time, and follow up with the corresponding cG(1)dG(1) and cG(1)cG(1) methods. In Fig. 173.2 and Fig. 173.3 below we show results from computations of two time-dependent bench-marks: flow around a bluff body and flow in a channel with a back-ward facing step.

173.2 The Incompressible Navier-Stokes Equations

The Navier-Stokes equations for an incompressible Newtonian fluid with constant kinematic viscosity $\nu > 0$, unit density and constant temperature enclosed in a volume Ω in $\mathbb{R}^3$ with boundary Γ, take the form: find the velocity/pressure (u, p) such that

$$\begin{aligned} \tfrac{\partial u}{\partial t} + (u \cdot \nabla)u - \nu\Delta u + \nabla p &= f && \text{in } \Omega \times I, \\ \nabla \cdot u &= 0 && \text{in } \Omega \times I, \\ u &= w && \text{on } \Gamma \times I, \\ u(\cdot, 0) &= u^0 && \text{in } \Omega, \end{aligned} \tag{173.1}$$

where $u = (u_1, u_2, u_3)$ is the velocity and p the pressure of the fluid and f, w, u^0, $I = (0, T)$, is a given driving force, boundary data, initial data and time interval, respectively. Recall that

$$\frac{\partial v}{\partial t} + (u \cdot \nabla)v = \frac{\partial v}{\partial t} + \sum_{i=1}^{3} u_i \frac{\partial v}{\partial x_i} \tag{173.2}$$

is the *particle derivative* of a quantity $v(x, t)$ measuring the rate of change of $v(x(t), t)$ with respect to time, that is the rate of change of v along a trajectory $x(t)$ of a fluid particle with velocity $u(x, t)$, satisfying $\frac{dx}{dt} = u(x(t), t)$. In particular, $\frac{\partial u}{\partial t} + (u \cdot \nabla)u$ is the acceleration (rate of change of velocity) of a fluid particle. The expression $\nu\Delta u - \nabla p$ represents the total force on a fluid particle resulting from of viscous shear force and an isotropic pressure. The first equation of (174.1), which is a vector equation

$$\frac{\partial u_i}{\partial t} + (u \cdot \nabla)u_i - \nu\Delta u_i + \frac{\partial p}{\partial x_i} = f_i, \qquad i = 1, 2, 3,$$

is the *momentum equation* expressing Newton's second law stating that the acceleration is proportional to the force, and the second equation expresses the incompressibility condition. We consider here the case of Dirichlet boundary conditions with the velocity u being prescribed on the boundary Γ. Below we consider Neumann and Robin boundary conditions. Below we will often write for short $(u \cdot \nabla)u = u \cdot \nabla u$.

The linear *Stokes equations* are obtained omitting the nonlinear term $u\cdot\nabla u$, which is possible if the velocity u is small, corresponding to *creeping flow*.

The *Reynolds number* Re is defined by $Re = \frac{uL}{\nu}$, where u represents a velocity and and L a length scale characteristic of the flow. The size of the Reynolds number is decisive. If $Re \sim 1$, then the flow is very viscous, a situation met in e.g. polymer flow or forming processes. In most applications in areo/hydro-dynamics, Re is much larger than 1, often very large up to 10^6 or even larger. In these cases with small viscosity, the flow may be very complex or turbulent.

There is a stationary analog of (174.1) assuming the solution to be independent of time along with the driving force and boundary data. A stationary solution normally arises as a limit of a time-dependent solution as time tends to infinity, and this is often reflected in the computation of a stationary solution through some kind of time-stepping until convergence. For larger Reynolds numbers, stable stationary solutions in general do not not exist.

173.3 The Basic Energy Estimate for Navier-Stokes

We now derive a basic stability estimate of energy type for the velocity u of a (u,p) of Navier-Stokes equation (174.1) assuming for simplicity that $f = 0$ and $w = 0$. Scalar multiplication of the momentum equation by u and integration with respect to x gives

$$\frac{1}{2}\frac{d}{dt}\int_\Omega |u|^2\,dx + \nu\sum_{i=1}^3\int_\Omega |\nabla u_i|^2\,dx = 0,$$

because by partial integration (with boundary terms vanishing),

$$\int_\Omega \nabla p\cdot u\,dx = -\int_\Omega p\nabla\cdot u\,dx = 0$$

and

$$\int_\Omega (u\cdot\nabla)u\cdot u\,dx = -\int_\Omega (u\cdot\nabla)u\cdot u\,dx - \int_\Omega \nabla\cdot u|u|^2\,dx$$

so that

$$\int_\Omega (u\cdot\nabla)u\cdot u\,dx = 0.$$

Integrating next with respect to time, we thus obtain the following basic stability estimate for any time $T > 0$:

$$\|u(\cdot,T)\|^2 + 2\nu\sum_{i=1}^3\int_0^T \|\nabla u_i\|^2\,dt = \|u^0\|^2, \tag{173.3}$$

where $\|\cdot\|$ denotes the $L_2(\Omega)$-norm. This estimate gives a bound on the velocity with the second term on the left representing the dissipation from the viscosity of the fluid. We see that the growth of this term over time corresponds to a decrease of the velocity (momentum) of the flow.

The case of large Reynold's number corresponding to small ν, with a normalization of velocity and typical length scale to unit size, is of particular interest with typically turbulent flows occuring. In laminar flow with small viscosity the dissipation is small because velocity gradients are not large, while in turbulent flow the dissipation is significant because the velocity gradients are large corresponding to a decay of velocities in the case of no driving forces.

FIGURE 173.1. Jacques-Louis Lions (1928-2001), founder of the French School of Numerical Analysis:"...optimal control problems for distributed parameter systems modeled by partial differential equations obviously connect to fundamental aspects of Body & Soul..."

173.4 Lions and his School

Jacques-Louis Lions (1929-2001), see Fig. 173.1, carried the strong French mathematical tradition coupled to physics and mechanics through the second half of the 20th century with important contributions to the theory and practice of partial differential equations using tools from Functional Analysis in the spirit of Sobolev. He created the French School of Numerical Analysis, which boomed with the development of the finite element method starting in the 1960s. Among many other things, Lions proved existence and uniqueness of solutions to the Navier-Stokes equations with a regularizing viscosity modification as indicated below.

173.5 Turbulence: Lipschitz with Exponent 1/3?

The mathematical modeling and simulation of turbulent flow represents one of the open problems of classical mechanics and physics, where today computational methods open new possibilities in the form of *Large Eddy Simulation* LES with *subgrid modeling*. Turbulent flow has features (vortices) on a range of scale from largest macroscopic of diameter of order one to smallest of order $\nu^{3/4}$, with ν the viscosity, assuming normalization to characteristic macroscopic velocity and length scale of order one, so that the macroscopic Reynolds number Re equals $1/\nu$. I typical applications Re may be of size 10^8 in which case the smallest length scale may be roughly of order 10^{-6} requiring of the order of 10^{18} degrees of freedom in a *Direct Numerical Simulation* DNS with resolution of all scales. This is way beyond the capacity of any computer within sight, with the present limit being set for DNS with a smallest scale of size 10^{-3} corresponding to Reynolds number roughly of order 10^4. To simulate flows with larger Reynolds number we may seek a *subgrid model* with the objective of modeling the effect on resolvables scales of unresolved scales. This may be possible using features of *scale similarity* of turbulent flow reflecting a certain repetition of flow features in a cascade from coarser to finer scales down to the smallest vortices where significant dissipation occurs. In Fig. 173.4 we show a jet undergoing transition from laminar to turbulent flow on a $128 \times 32 \times 32$ mesh.

Let us give an argument indicating a feature of scale similarity first presented by the Russian mathematician Kolmogorov 1941: Let then h be the smallest scale, that is the diameter of the smallest vorticity, and let $\bar{u}$ be the corresponding velocity of the smallest vorticity. We may then argue that we should have $\bar{u}h \sim \nu$, since the break up of larger vortices into to smaller should continue until the local Reynolds number becomes small enough (of size 50-100). Further, turbulent dissipation on the smallest scale of order one would mean that $\nu(\frac{\bar{u}}{h})^2 \sim 1$. From these two relations, we conclude that $h \sim \nu^{3/4}$ as anticipated and also that $\bar{u} \sim \nu^{1/4}$. We conclude that

$$|u(x) - u(y)| \sim |x - y|^{1/3}$$

for $y = x + h$, and by scale similarity we may expect this relation to hold for general x and y, that is, that the turbulent velocity should be Lipschitz (Hölder) continuous with exponent 1/3.

Does the above derivation have any to do with reality? Yes, both physical experiments and DNS indicate that turbulent flow indeed has features of scale similarity with Lipschitz (Hölder) continuity with exponent 1/3. This gives hope that subgrid modeling may be feasible for turbulent flow and thus that computational simulation of turbulent flow would be possible, and more and more so as the computational power increases.

Summing up, it thus appears that computational simulation of turbulent flow may be possible, and this would in a way settle most questions from a practical point of view: we would be able to simulate and predict turbulent flow. However, we would still lack a mathematical model of turbulence more tractable than simply the Navier-Stokes equations in DNS. So, as human beings we may not be able to "understand turbulence" in the same way as we can understand e.g. the fundamental solution of the Laplacian ($\frac{1}{4\pi|x|}$), but we would be able to computationally simulate turbulent flow. Maybe this is the most we can ask for?

173.6 Existence and Uniqueness of Solutions

The question of existence and uniqueness of solutions to the Navier-Stokes equations is one of the unsolved problems of mathematics. If we change the viscosity from a Newtonian constant viscosity ν to a non-Newtonian solution dependent viscosity $\hat{\nu} = \nu + Ch^2|\nabla u|$, where h is a parameter corresponding to a smallest scale, then, existence and uniqueness is possible to prove using standard methods as shown by Lions. Since with h small the modification will be small, except where ∇u is very large, the modification may be viewed as a regularization eliminating certain extreme situations with very large velocity gradients, where at any rate the Newtonian property of constant viscosity may be questioned. This directly couples to subgrid modeling of turbulent flow, where $\hat{\nu}$ corresponds to a so called *turbulent viscosity*, with the constant C to be modeled computationally.

173.7 Numerical Methods

Trying to solve the incompressible Navier-Stokes equations numerically, we meet the following difficulties:

- instabilities from discretization of convection terms,
- pressure instabilities in equal order interpolation of velocity and pressure.

The simplest cure to convection instability is to increase the viscosity ν in the computation so that $\nu \geq uh$, where u is the local fluid velocity and h is the local mesh size. The simplest stabilization of the pressure p, is to modify the incompressibility equation $\nabla \cdot u = 0$ to $-\nabla \cdot (\delta \nabla p) + \nabla \cdot u = 0$, with $\delta \approx h^2$ with $h(x)$ the local mesh size.

In Galerkin methods the stabilization can be achieved in higher-order consistent form by adding least-squares control of residuals. We present this approach below in the context of the cG(1)dG(0) method with cG(1)

in space and dG(0) in time. We also present corresponding cG(1)cG(1) and cG(1)dG(1) methods.

173.8 The Stabilized cG(1)dG(0) Method

We now present the cG(1)dG(0) method for (174.1) starting with the case of homogeneous Dirichlet boundary conditions. Let $0 = t_0 < t_1 < ... < t_N = T$ be a sequence of discrete time levels with associated time steps $k_n = t_n - t_{n-1}$. Let W_h be the usual finite element space of continuous piecewise linear functions on a triangulation $\mathcal{T}_h = \{K\}$ of Ω with mesh function $h(x)$. Let W_h^0 be the space of functions in W_h vanishing on Γ. We shall seek an approximate velocity $U(x,t)$ such that $U(x,t)$ is continuous and piecewise linear in x for each t, and $U(x,t)$ is piecewise constant in t for each x. Similarly, we shall seek an approximate pressure $P(x,t)$ which is continuous piecewise linear in x and piecewise constant in t. More precisely, we shall seek $U^n \in V_h^0$ with $V_h^0 = W_h^0 \times W_h^0 \times W_h^0$ and $P^n \in W_h$ for $n = 1, ..., N$, and we shall set

$$\begin{aligned} U(x,t) &= U^n(x) \quad x \in \Omega, \quad t \in (t_{n-1}, t_n], \\ P(x,t) &= P^n(x) \quad x \in \Omega, \quad t \in (t_{n-1}, t_n]. \end{aligned} \tag{173.4}$$

Further we write for velocities $v = (v_i)$ and $w = (w_i)$

$$(v,w) = \int_\Omega v \cdot w \, dx, \qquad (\nabla v, \nabla w) = \int_\Omega \sum_i^3 \nabla v_i \cdot \nabla w_i \, dx,$$

and similarly for scalar functions p and q defined on Ω:

$$(p,q) = \int_\Omega pq \, dx.$$

We now formulate the cG(1)dG(0) method without stabilization as follows: For $n = 1, ..., N$, find $(U^n, P^n) \in V_h^0 \times W_h$ such that

$$\begin{aligned} (\frac{U^n - U^{n-1}}{k_n}, v) + (U^n \cdot \nabla U^n + \nabla P^n, v) + (\nu \nabla U^n, \nabla v) &= (f^n, v) \quad \forall v \in V_h^0, \\ (\nabla \cdot U^n, q) &= 0 \quad \forall q \in W_h, \end{aligned} \tag{173.5}$$

where $U^0 = u^0$, and we set $f^n(x) = f(x, t_n)$. We see that the discrete equations result from multiplication of the momentum equation with $v \in V_h^0$ and the incompressibility equation by $q \in W_h$, followed by integration over Ω including integration by parts in the term $(-\nu\Delta U, v)$.

We can write the cG(1)dG(0) method without stabilization alternatively as follows: For $n = 1, ..., N$, find $(U^n, P^n) \in V_h^0 \times W_h$ such that

$$\begin{aligned}(\frac{U^n - U^{n-1}}{k_n}, v) + (U^n \cdot \nabla U^n + \nabla P^n, v) + (\nabla \cdot U^n, q)\\ + (\nu \nabla U^n, \nabla v) = (f^n, v) \quad \forall (v, q) \in V_h^0 \times W_h,\end{aligned} \tag{173.6}$$

where we simply added the equations in 173.5.

The cG(1)dG(0) method with stabilization takes the form: For $n = 1, ..., N$, find $(U^n, P^n) \in V_h^0 \times W_h$ such that

$$\begin{aligned}(\frac{U^n - U^{n-1}}{k_n}, v) + (U^n \cdot \nabla U^n + \nabla P^n, v + \delta(U^n \cdot \nabla v + \nabla q)) + (\nabla \cdot U^n, q)\\ + (\nu \nabla U^n, \nabla v) = (f^n, v + \delta(U^n \cdot \nabla v + \nabla q)) \quad \forall (v, q) \in V_h^0 \times W_h,\end{aligned} \tag{173.7}$$

where δ is a stabilization parameter defined as follows: $\delta(x) = h^2(x)$ in the case of *diffusion-dominated* flow with $\nu \geq Uh$, and

$$\delta = (\frac{1}{k} + \frac{U}{h})^{-1} \tag{173.8}$$

in the case of *convection dominated* flow with $\nu < Uh$. Note that if $k \approx \frac{h}{U}$, which is a natural choice of time step in the convection-dominated case, then $\delta \approx \frac{1}{2}\frac{h}{U}$. Note further that the stabilized form (173.7) of the cG(1)dG(0) method is obtained by replacing v by $v + \delta(U^n \cdot \nabla v + \nabla q)$ in the terms $(U^n \cdot \nabla U^n + \nabla P^n, v)$ and (f^n, v). In principle, we should make the replacement throughout, but in the present case of the cG(1)dG(0), only the indicated terms get involved because of the low order of the approximations. The perturbation in the stabilized method is of size δ, and thus the stabilized method has the same order as the original method (first order in h if $k \sim h$).

Letting v vary in (173.7) while choosing $q = 0$, we get the following equation (the discrete momentum equation):

$$\begin{aligned}(\frac{U^n - U^{n-1}}{k_n}, v) + (U^n \cdot \nabla U^n + \nabla P^n, v + \delta U^n \cdot \nabla v)\\ + (\nu \nabla U^n, \nabla v) = (f^n, v + \delta U^n \cdot \nabla v) \quad \forall v \in V_h^0,\end{aligned} \tag{173.9}$$

and letting q vary while setting $v = 0$, we get the following discrete pressure equation:

$$(\delta \nabla P^n, \nabla q) = -(\delta U^n \cdot \nabla U^n, \nabla q) - (\nabla \cdot U^n, q) + (\delta f^n, \nabla q) \quad \forall q \in W_h. \tag{173.10}$$

We normally seek to solve the system (173.7) iteratively alternatively solving the velocity equation (173.9) for U^n with P^n given, and the pressure equation (173.10) for P^n with U^n given.

173.9 The cG(1)cG(1) Method

We present the following cG(1)cG(1) variant of the cG(1)dG(0) method with cG(1) in time instead of dG(0): For $n = 1, ..., N$, find $(U^n, P^n) \in V_h^0 \times W_h$ such that

$$(\frac{U^n - U^{n-1}}{k_n}, v) + (\hat{U}^n \cdot \nabla \hat{U}^n + \nabla P^n, v + \delta(\hat{U}^n \cdot \nabla v + \nabla q)) + (\nabla \cdot \hat{U}^n, q) + (\nu \nabla \hat{U}^n, \nabla v) = (f^n, v + \delta(\hat{U}^n \cdot \nabla v + \nabla q)) \quad \forall (v, q) \in V_h^0 \times W_h, \tag{173.11}$$

where $\hat{U}^n = \frac{1}{2}(U^n + U^{n-1})$. Evidently, we obtained the cG(1) version by changing from U^n to $\hat{U}^n$ in all terms but the first in the cG(1)dG(0) method.

173.10 The cG(1)dG(1) Method

We shall now formulate the cG(1)dG(1) method obtained by replacing dG(0) by dG(1) in the cG(1)dG(0) method. In this method the discrete velocity $U(x,t)$ is piecewise linear linear in time on each time interval I_n, with possibly discontinuities at the discrete time levels t_n. More precisely, we make the Ansatz:

$$U^n(x,t) = \frac{t_n - t}{k_n} U_+^{n-1}(x) + \frac{t - t_{n-1}}{k_n} U_-^n(x), \qquad \text{for } t_{n-1} < t < t_n, \tag{173.12}$$

where U_+^{n-1} and U_-^n belong to V_h^0. We note that

$$U_\pm^n(x) = \lim_{s \to 0^+} U(x, t_n \pm s)$$

is the limit of $U(x,t)$ as t approaches t_n from below $(-)$, or above $(+)$. The cG(1)dG(1) method takes the form: For $n = 1, ..., N$, find U^n of the form (173.12) and $P^n \in W_h$, such that for all $v(x,t) = w_1(x,t) + (t - t_{n-1})w_2(x,t)$ with $w_1, w_2 \in V_h^0$ and $q \in W_h$,

$$\begin{aligned}(U_+^{n-1} - U_-^{n-1}, v) + \int_{t_{n-1}}^{t_n} ((\dot{U}^n + U^n \cdot \nabla U^n \\ + \nabla P^n, v + \delta(\dot{U}^n + U^n \cdot \nabla v + \nabla q)) + (\nabla \cdot U^n, q))\, dt \\ + \int_{t_{n-1}}^{t_n} (\nu \nabla U^n, \nabla v)\, dt = \int_{t_{n-1}}^{t_n} (f^n, v + \delta(\dot{U} + U^n \cdot \nabla v + \nabla q)).\end{aligned} \tag{173.13}$$

We may similarly let P be piecewise linear discontinuous in time.

173.11 Neumann Boundary Conditions

To properly model Neumann boundary conditions, we first need to recall that the components σ_{ij} of the *total stress tensor* $\sigma = (\sigma_{ij})$ acting on a fluid element, are given by

$$\sigma_{ij} = \bar{\sigma}_{ij} - p\delta_{ij}, \quad i,j = 1,2,3,$$

where the *stress deviatoric* $\bar{\sigma} = (\bar{\sigma}_{ij})$ is coupled to the *strain tensor* $\epsilon(u) = (\epsilon_{ij}(u))$ with components

$$\epsilon_{ij}(u) = (\partial u_i/\partial x_j + \partial u_j/\partial x_i)/2, \quad i,j = 1,2,3,$$

through the constitutive relation of a *Newtonian fluid*:

$$\bar{\sigma}_{ij} = 2\nu\epsilon_{ij}(u), \quad i,j = 1,2,3,$$

where ν is the constant viscosity, and $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. We observe that the trace of the stress deviatoric is zero, that is,

$$\sum_{i=1}^{3} \bar{\sigma}_{ii} = 2\nu \sum_{i=1}^{3} \epsilon_{ii}(u) = 2\nu\nabla\cdot u = 0,$$

and thus the total stress σ is decomposed into a stress deviatoric $\bar{\sigma}$ with zero trace and an isotropic pressure p. Further, a direct computation shows that

$$\nu\Delta u - \nabla p = \nabla\cdot\sigma, \tag{173.14}$$

where $\nabla\cdot\sigma$ is a vector with components $(\nabla\cdot\sigma)_i$ given by

$$(\nabla\cdot\sigma)_i = \sum_{j=1}^{3} \frac{\partial\sigma_{ij}}{\partial x_j}.$$

Multiplying 173.14 by $v = (v_i)$ with $v = 0$ on Γ and integrating by parts, we find that

$$\nu(\nabla u, \nabla v) + (\nabla p, v) = 2\nu(\epsilon(u), \epsilon(v)) + (\nabla p, v),$$

where

$$(\epsilon(u), \epsilon(v)) = \sum_{i,j=1}^{3} \int_\Omega \epsilon_{ij}(u)\epsilon_{ij}(v)\, dx.$$

We are thus led to replace the term $(\nu\nabla u, \nabla v)$ by the term $(2\nu\epsilon(u), \epsilon(v))$ in variational formulations of the Navier-Stokes equations. In the case of Dirichlet boundary conditions for the velocity the two expressions are equal, since the test velocity v vanishes on Γ, but in the case of Neumann type

boundary conditions the replacement opens the possibility of enforcing in variational form a Neumann boundary condition of the form

$$\sum_{j=1}^{3}\sigma_{ij}n_j=\sum_{j=1}^{3}\bar{\sigma}_{ij}n_j-pn_i=\sum_{j=1}^{3}2\nu\epsilon_{ij}(u)n_j-pn_i=g_i \text{ on } \Gamma_2,\ i=1,2,3, \tag{173.15}$$

which expresses that the total force on the boundary part Γ_2 is equal to the given force $g=(g_i)$. For example, if $g=0$, then this condition expresses that the total force is zero on Γ_2, which we may use as an outflow boundary condition simulating that the fluid freely flows out into a large reservoir. More precisely, the presence of the terms

$$-(p,\nabla\cdot v)+(2\nu\epsilon(u),\epsilon(v))$$

in a variational formulation with v varying freely on Γ_2, will enforce a homogeneous Neumann boundary condition 173.15 upon integration by parts.

We now consider a typical situation with the boundary Γ decomposed into two parts Γ_1 an Γ_2 with the velocity being equal to a given velocity w on Γ_1 and imposing the homogeneous Neumann condition 173.15 on Γ_2. For simplicity, we assume that w is independent of time, the extension to time dependence of w being evident. Typically, w will be zero on a part of Γ_1 and will be directed into Ω on the remaining part corresponding to a given inflow.

We let V_h be the space of continuous piecewise linear velocities v on a triangulation $\mathcal{T}_h=\{K\}$ of Ω with mesh function $h(x)$, satisfying the boundary condition $v=w$ on Γ_1, and let V_h^0 be the corresponding test space of functions with $v=0$ on Γ_1. Let W_h be the space of continuous piecewise linear pressures p on $\mathcal{T}_h=\{K\}$, and W_h^0 the corresponding test space of pressures q such that $q=0$ on Γ_2.

The stabilized cG(1)dG(0) method can be formulated as follows: For $n=1,...,N$ seek $U^n\in V_h$ and $P^n\in W_h$ such that

$$\begin{aligned}&(\frac{U^n-U^{n-1}}{k_n},v)+(U^n\cdot\nabla U^n,v+\delta U^n\cdot\nabla v)-(P^n,\nabla\cdot v)\\&+(2\nu\epsilon(U^n),\epsilon(v))=(f^n,v+\delta U^n\cdot\nabla v)\quad\forall v\in V_h^0,\end{aligned} \tag{173.16}$$

$$(\delta\nabla P^n,\nabla q)=-(\delta U^n\cdot\nabla U^n,\nabla q)-(\nabla\cdot U^n,q)+(\delta f^n,\nabla q)\quad\forall q\in W_h^0, \tag{173.17}$$

where we choose P^n on Γ_2 according to 173.15 with $g=0$ and u replaced by U. Again we seek to solve the system iteratively alternatively solving the velocity equation (173.16) for U^n with P^n given, and the pressure equation (173.17) for P^n with U^n given.

173.12 Computational Examples

We now present some computational examples of 3d time dependent flows, using the stabilized cG(1)cG(1) method on a mesh with meshsize $h = 1/32$.

In Figure Fig. 173.2 we present the solution of a bluff body problem: a flow in a channel with 1x1 square cross section and length 4, with a square obstacle with side length 0.25 centered at $(0.5, 0.5, 0.5)$. We have used zero Dirichlet boundary condition for the velocity on the side walls and Neumann outflow boundary conditions on the outflow boundary. On the inflow a parabolic velocity is prescribed.

In Figure Fig. 173.3 we present the solution of a step down problem in a similar channel with a step down of height and length 0.5.

Finally in Figure Fig. 173.4 we present computations of transition to turbulence in a circular jet flow with streamwise velocity 1 in the jet and zero outside the jet, where we apply a small random perturbation. Here we have used periodic boundary conditions in all directions.

Chapter 173 Problems

173.1. Prove for that a solution (u, p) of (174.1) with $f = 0 and w = 0$ satisfies the following energy estimate for $t > 0$:

$$\int_\Omega |u(x,t)|^2 + 2\nu \int_0^t \int_\Omega |\nabla u(x,s)|^2 dx ds = \int_\Omega |u^0(x)|^2 dx.$$

Hint: Multiply the momentum equation by u and use that if $\nabla \cdot u = 0$, then

$$\int_\Omega (u \cdot \nabla) u \cdot u \, dx = 0,$$

which follows by integration by parts.

173.2. Prove a basic stability estimate for (173.7) by choosing $(v, q) = (U, P)$.

> Thus the methods of Lagrange and Hamilton are undoubtedly *useful* in helping us to carry out the primary task of dynamics - namely, to find out how systems move. But it would be wrong to think that this is the sole purpose of these general methods or even their main purpose. They do much more. In fact, they teach us what dynamics *really is* : It is the study of certain types of differential equations. (Synge and Griffiths, Principles of Mechanics, 1959)

> I sing the body electric,
> The armies of those I love engirth me and I engirth them,
> They will not let me off till I go with them, respond to them,
> And discorrupt them, and charge them full
> with the charge of the soul.

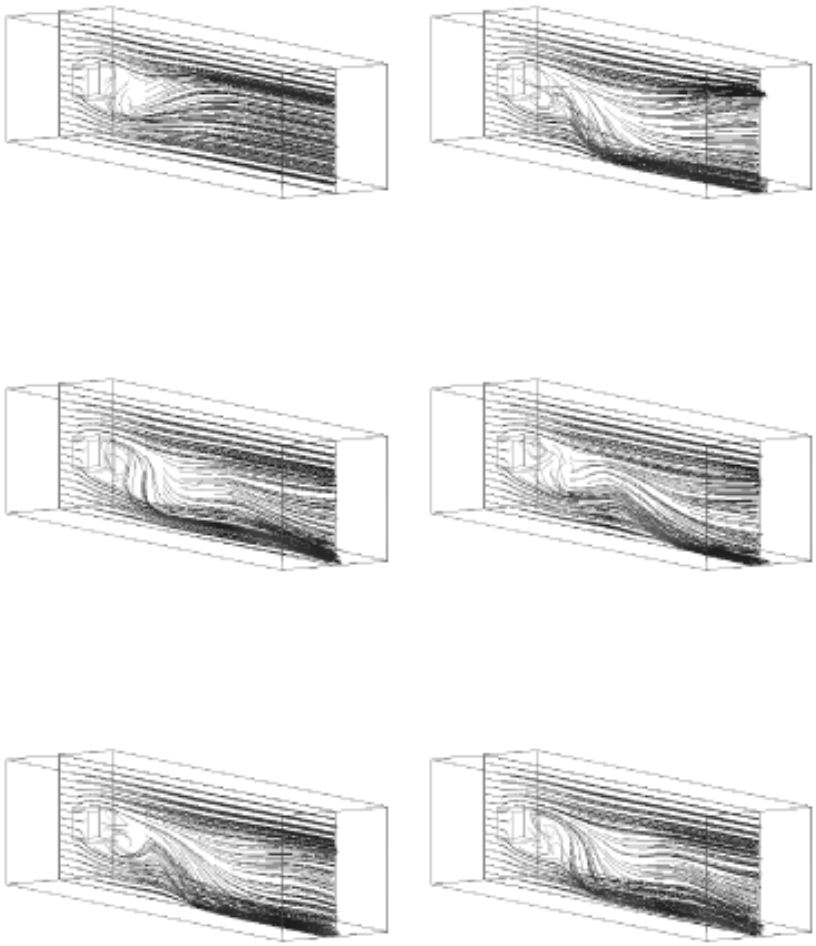


FIGURE 173.2. Bluff body flow computations for $t = 2, 4, 6, 8, 10, 12$.

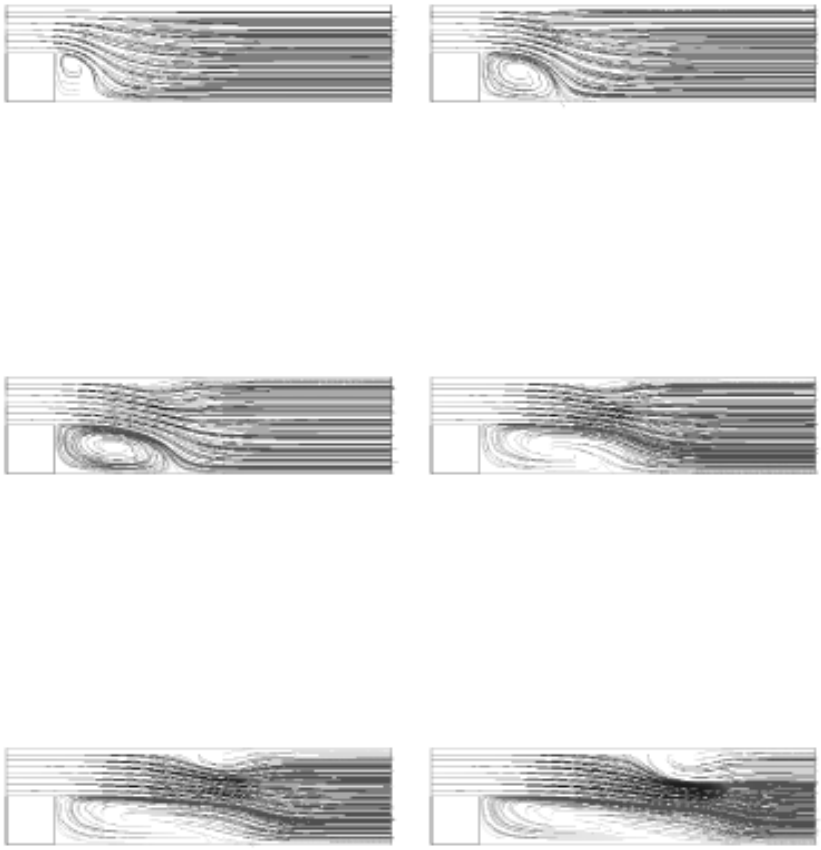


FIGURE 173.3. Step down flow computations for $t = 1, 2, 3, 4, 5, 6$.

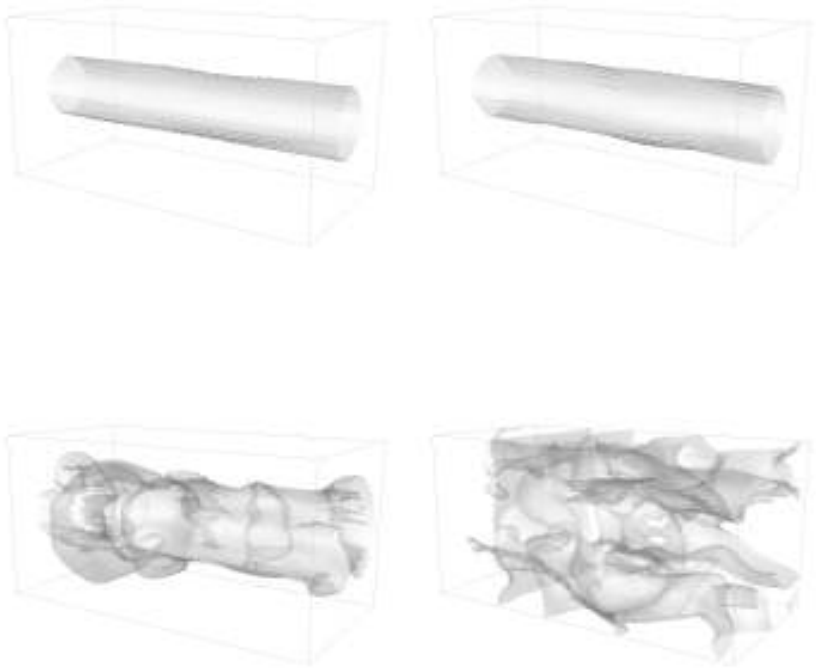


FIGURE 173.4. Streamwise velocity isosurfaces for $|u_1| = 0.02$ in a jet in transition from laminar to turbulent flow, for $t = 5, 7, 10, 15$

Was it doubted that those who corrupt
their own bodies conceal themselves?
And if those who defile the living are as bad as
they who defile the dead?
And if the body does not do fully as much as the soul?
And if the body were not the soul, what is the soul?
(Walt Whitman).

174
The Mystery of Flight

When you lean back for take-off in a jumbojet, maybe the following question flashes through your mind: How is it possible that the 400 squaremeter wings can carry 400 tons at a wingload of 1 ton per squaremeter in sustained flight in the air? Or maybe you are satisfied with some of the explanations offered in popular science, like higher velocity and lower pressure on the upper surface of the wing because it is curved and air there has a longer path to travel than below? Or maybe you are an aeroplane engineer or pilot and know very well why an airplane can fly?

In either case, you should get a bit worried by reading that the authority NASA on its website [43] dismissses all popular science theories for lift, including your favorite one, as being incorrect, but then refrains from presenting any theory claimed to be correct! NASA surprisingly ends with an empty out of reach: *To truly understand the details of the generation of lift, one has to have a good working knowledge of the Euler Equations.* The Plane&Pilot Magazine [44] has the same message and New York Times [8] informs us:

- *To those who fear flying, it is probably disconcerting that physicists and aeronautical engineers still passionately debate the fundamental issue underlying this endeavor: what keeps planes in the air?*

174.1 Overview

In this arcticle we present a new mathematical and physical explanation of the generation of *lift* L and *drag* D of a wing based on new discoveries of the dynamics of *turbulent* airflow around a wing, obtained by computational solution of the basic mathematical model of fluid dynamics: the *Navier-Stokes/Euler equations*. When flying in the air, the downward gravitational force is balanced by upward wing lift L, while backward wing drag D is balanced by forward thrust from engine, and wing-beat for birds, or descent in gliding flight without forward thrust.

We show that a wing creates lift as a reaction force from redirecting air downwards, referred to as *downwash*, with less than $1/3$ coming from the lower wing surface pushing air down and the major remaining part from the upper surface sucking air down, with a resulting *lift/drag quotient* $\frac{L}{D}$ of size $10-20$.

The enigma of flight is why the air flow separates from the upper wing surface at the *trailing edge*, and not before, with the flow after separation being redirected downwards according to the tilting of the wing or *angle of attack*. We will reveal the secret to be an effect of a fortunate combination of features of *slightly viscous incompressible flow* including a crucial *instability mechanism at separation* analogous to that seen in the swirling flow down a bathtub drain, generating both suction on the upper wing surface and drag.

We show that this mechanism of lift and drag is operational for angles of attack smaller than a critical value of about $16-20$ degrees depending on the shape of the wing, for which the flow separates from the upper wing surface well before the trailing edge with a sudden increase of drag and decrease of lift referred to as *stall*.

It is absolutely crucial that $\frac{L}{D}$ is large, of size 10 or bigger, since otherwise the muscle power of a bird would not suffice, and the fuel consumption of an airplane would be prohibitive. Flying on a tilted barn door at 45 degrees angle of attack with $\frac{L}{D}\approx 1$, is not an option.

An outline of the article is as follows: We first recall classical theories for lift and drag and then in pictures describe the new theory. We support the new theory by computational solutions of the Navier-Stokes equations, also showing that the classical theories are incorrect. We then present basic aspects of the mathematics of *turbulent solutions of the Navier-Stokes equations* underlying the new theory.

174.2 Newton, d'Alembert and Kutta-Zhukovsky

The problem of explaining *why* it is possible to fly in the air using wings has haunted scientists since the birth of mathematical sciences. The mystery is *how* a sufficiently large ratio $\frac{L}{D}$ can be created.

In the *gliding flight* of birds and airplanes with fixed wings at subsonic speeds, $\frac{L}{D}$ is typically between 10 and 20, which means that a good glider can glide up to 20 meters upon loosing 1 meter in altitude, or that Charles Lindberg could cross the Atlantic in 1927 at a speed of 50 m/s in his 2000 kg *Spirit of St Louis* at an effective engine thrust of 150 kp (with $\frac{L}{D} = 2000/150 \approx 13$) from 100 horse powers.

By Newton's 3rd law, lift must be accompanied by downwash with the wing redirecting air downwards. The enigma of flight is the mechanism of a wing generating substantial downwash at small drag, which is also the enigma of sailing against the wind with both sail and keel acting like wings creating substantial lift [30].

Classical mathematical mechanics could not give an answer to the mystery of gliding flight: Newton computed by elementary mechanics the lift of a tilted flat plate redirecting a horisontal stream of fluid particles, but obtained a disappointingly small value proportional to the square of the angle of attack. To Newton the flight of birds was inexplicable, and human flight certainly impossible.

D'Alembert followed up in 1752 by formulating his paradox about zero lift/drag of *inviscid incompressible irrotational steady flow* referred to as *potential flow,* which seemed to describe the airflow around a wing since the viscosity of air is very small so that it can be viewed as being inviscid (with zero viscosity). Mathematically, potential flow is given as the gradient of a *harmonic funtion* satisfying *Laplace's equation.*

At speeds less than say 300 km/h air flow is almost incompressible, and since a wing moves into still air the flow it could be be expected to be irrotational without swirling rotating vortices. D'Alembert's mathematical potential flow thus seemed to capture physics, but nevertheless had neither lift nor drag, against all physical experience. The wonderful mathematics of potential flow and harmonic functions thus showed to be without physical relevance: This is *D'Alembert's paradox* which came to discredit mathematical fluid mechanics from start [30, 48, 20].

To explain flight d'Alembert's paradox had to be resolved, but nobody could figure out how and it was still an open problem when Orwille and Wilbur Wright in 1903 showed that heavier-than-air human flight in fact was possible in practice, even if mathematically it was impossible.

Mathematical fluid mechanics was then saved from complete collapse by the young mathematicians Kutta and Zhukovsky, called the father of Russian aviation, who explained lift as a result of perturbing potential flow by a large-scale circulating flow or *circulation* around the two-dimensional section of a wing, and by the young physicist Prandtl, called the father of

modern fluid dynamics, who explained drag as a result of a *viscous boundary layer* [45, 46, 47, 21].

This is the basis of state-of-the-art [16, 37, 14, 2, 19, 49, 50], which essentially is a simplistic theory for lift without drag at small angles of attack in inviscid flow and for drag without lift in viscous flow. However, state-of-the-art does not supply a theory for lift-and-drag covering the real case of *3d slightly viscous turbulent* flow of air around a 3d wing of a jumbojet at the critical phase of take-off at large angle of attack (12 degrees) and subsonic speed (270 km/hour), as evidenced in e.g. [1, 3, 4, 6, 8, 10, 34, 36, 41]. The simplistic theory allows an aeroplane engineer to roughly compute the lift of a wing a crusing speed at a small angle of attack, but not the drag, and not lift-and-drag at the critical phase of take-off [42, 13]. The lack of mathematics has to be compensated by experiment and experience. The first take off of the new Airbus 380 must have been a thrilling experience for the design engineers.

174.3 From Old to New Theory of Flight

A couple of years ago we stumbled upon a resolution of d'Alembert's paradox [4, 30], when computing turbulent solutions of the basic mathematical model of fluid mechanics, the Navier-Stokes equations. The resolution naturally led us to a new theory of flight, which we will explain below. You will find that it is quite easy to grasp, because it can be explained using different levels of mathematics. We start out easy with the basic principle in concept form and then indicate some of the mathematics with references to more details. Supporting information is given in the Google knols [32] and [33].

Before proceeding to work we recall both folklore and state-of-the-art mathematics explantions of flight as being either correct but trivial, or nontrivial but incorrect, as follows:

- Downwash generates lift: trivial without explanation of reason for downwash from suction on upper wing surface.
- Low pressure on upper surface: trivial without explanation why.
- Low pressure on curved upper surface because of higher velocity (by Bernouilli's law), because of longer distance: incorrect.
- Coanda effect: The flow sticks to the upper surface by viscosity: incorrect.
- Kutta-Zhukovsky: Lift comes from circulation: incorrect.
- Prandtl: Drag comes mainly from viscous boundary layer: incorrect.

174.4 The Principle of Flying

We will find that the secret of flight is revealed in Fig. 174.1: To the left we see potential flow around a portion of a long wing with zones of high (H) and low (L) pressure giving no net lift, because the pressure is high on top of the wing at the trailing edge and low below. This makes the flow leave the wing in the same direction as it approaches, thus without downwash and lift.

Potential flow is a mathematical solution without lift/downwash of the Navier-Stokes equations (with vanishing viscosity), which however is fundamenatlly different from the flow observed in reality with lift/downwash. Potential flow is a fictional mathematical solution without physical relevance, and the reason hides the secret of both d'Alembert's paradox and flight: Potential flow is very sensitive to a specific form of perturbation and thus is unstable and non-physical.

Potential flow is similar to an inverted pendulum in upright equilibrium or a pen balancing on its tip, which is a mathematical solution of the equations of motion, but an unstable non-physical solution which under a small perturbation away from the fully upright position will change into a different swinging motion. Potential flow without lift/downwash changes under a specific form of perturbation into a different more stable physical flow with lift/downwash, with a turbulent fluctuating layer including the perturbation attaching to the trailing edge, as we will see in computational simulations below with movies on [31].

FIGURE 174.1. Correct explanation of lift by perturbation of potential flow (left) at separation from physical low-pressure turbulent counter-rotating rolls (middle) changing the pressure and velocity at the trailing edge into a flow with downwash and lift (right).

The specific form of perturbation is illustrated in the middle picture of Fig.174.1 showing a layer of counter-rotating rolls of swirling flow attaching to the trailing edge, with each roll similar to the swirling flow in a bathtub drain. The layer of rolls is distributed all along the trailing edge and is not related to the wing tip vortex, which often is seen at landing in moist air, since we assume the wing to be long. The perturbation switches the pressure distribution of potential flow at the trailing edge since the pressure

inside the rolls is low, into the flow depicted to the right which has both lift, downwash and drag.

The specific perturbation thus hides the secret of flight as a flow with both lift, downwash and drag. By understanding mathematically the origin and nature of the instability mechanism generating the counter-rotating rolls at the separation of potential flow, which we do in more detail below, we will be able to reveal the mathematical secret of flight. In short, the counterrotating rolls develop when the opposing flows from above and below meet on top of the wing before separation and first are retarded and then accellerated and stretched in the flow direction, as shown in detail in [4, 30, 26, 23]. We understand that inside the rolls of swirling flow the pressure must be low to keep the roll together, and it is this low pressure that annihilates the high pressure on top to allow the flow to leave the wing in the direction of the upper surface tangent with substantial downwash as illustrated in the figure.

We see that the fundamental instability mechanism changes the flow at the trailing edge to give lift, but does not change the flow at the leading edge where the flow gives positive lift. Real flow thus shares a very important property with potential flow, namely to not separate at the crest of the flow above the leading edge. If it did, downwash and lift would be lost: This is what happens when a wing stalls at a too large angle of attack.

Summing up we have that lift comes from the instability mechanism at separation consisting of counter-rotating low-pressure rolls of swirling flow, which also creates drag by suction from the low pressure. Thus lift comes along with drag: No lift without drag. Lift without drag is an illusion, although still a common dream.

174.5 Comparison with Kutta-Zhukovsky

We compare with the classical explanation presented by Kutta-Zhukovsky illustrated in Fig.174.2, which you find in most books claiming to explain flight: We see again potential flow, now around a section of the wing, but combined with a different perturbation consisting of large scale circulating flow around the wing. This perturbation also changes the pressure distribution to give lift/downwash as illustrated in the picture to the right. However, as we will see below, the circulating flow around the wing does not arise in reality: Kutta-Zhukovsky's circulating flow is purely fictional and generates lift/downwash by a non-physical mechanism which does not occur in reality.

Nevertheless, with no alternative in sight, Kutta-Zhukovsky's trick to generate lift/downwash is generally viewed as a mathematically sophisticated way of explaining flight, beyond comprehension for most people. We shall find that the true reason it cannot be understood, is that it does not

make sense, simply because there is no physical mechanism to generate the large scale circulation around the wing, nor the associated so-called *starting vortex* behind the wing supposedly balancing the circulation indicated in the right picture of Fig.174.2.

We observe that Kutta-Zhulovsky flow is two-dimensional, since both potential flow and circulation is constant in the wing direction and thus can be depicted in a plane figure, while the true flow is fully three-dimensional with the specific perturbation bringing in a variation in the wing direction. Kutta-Zhukovsky flow is like potential flow a non-physical two-dimensional stationary flow, while the real flow around a wing is a three-dimensional partially fluctuating turbulent flow.

FIGURE 174.2. Incorrect Kutta-Zhukovsky explanation of lift by perturbation of potential flow (left) by unphysical circulation around the section (middle) resulting in flow with downwash/lift and starting vortex (right).

174.6 Effects of Small Viscosity

We conclude that flying is possible because of a fortunate combination of the following properties of real slightly viscous incompressible flow:

- non-separation at the crest of a wing because the flow is there similar to potential flow,
- the instability mechanism of potential flow at separation changes the pressure distribution at the trailing edge to give lift, and drag.

Slightly viscous flow has small *skin friction* along the boundary, which makes it similar to potential flow with zero skin friction satisfying a *slip* boundary condition at a solid boundary modeling that fluid particles can slide along the boundary without friction. Small skin friction can thus be modeled by zero skin friction requiring the normal velocity to vanish at the boundray, but imposing no restriction on the tangential velocity.

For a more viscous fluid like syrup with larger skin friction, instead a *no-slip* boundary condition is used requiring that both normal and tangential flow velocities vanish on the boundary modeling that fluid particles close to the boundary have small speed and connect to the interior flow by a

boundary layer where the flow speed changes from zero to the free stream speed. The effect of a no-slip boundary condition causing a boundary layer, is that the flow separates at the crest with loss of lift as compared to slightly viscous flow. This is because in a viscous boundary layer the pressure gradient normal to the boundary vanishes and thus cannot contribute to the normal acceleration required to keep fluid particles following the curvature of the boundary after the crest, as shown in detail in [27]. It is thus the *slip boundary condition modeling a turbulent boundary layer* in slightly viscous flow, which forces the flow to suck to the upper surface and create downwash. Gliding flight in viscous flow is thus not possible, which explains why small insects do not practice gliding flight because to them air appears to be viscous.

174.7 Wellposedness vs Clay Millennium Problem

In order to judge the physical relevance of a mathematical solution, stability must be assessed. Only *wellposed* solutions which are suitably stable in the sense that small perturbations have small effects when properly measured, have physical significance as observable pheonomena, as made clear by in particular the mathematician J. Hadamard in 1902 [15]. However, the completely crucial and fundamental question whether solutions of the Navier-Stokes equations are wellposed, has not been studied because of lack mathematical techniques for quantitative analysis, as evidenced in the formulation of the Clay Millennium Prize Problem on the Navier-Stokes equations excluding wellposedness [28, 26]. G. Birkhoff was heavily criticized for posing this question in [20], and refrained from further studies. The first step towards resolution of d'Alembert's paradox and the mathematical secret of flight is thus to pose the question if potential flow is wellposed, and then to realize that it is not. It took 256 years to take these steps.

174.8 Computed Lift and Drag

We now a take a closer look at solutions of the Navier-Stokes equations, computed by the General Galerkin finite element method G2 [4]. These solutions should tell us the truth because the Navier-Stokes equations express the basic laws of physics of conservation of mass and momentum (Newton's 2nd law), which cannot be doubted. We focus on the case of slightly viscous incompressible flow of relevance for airplanes at subsonic speeds and larger birds. The fact that the fluid has small viscosity is of crucial importance both for physics and computation: First, the flow is then turbulent with a turbulent boundary layer allowing the flow to suck to the upper surface

of the wing and cause downwash and lift. Second, a turbulent boundary layer can be modeled by a slip or small friction boundary condition which makes it possible to simulate the flow without computationally resolving thin boundary layers, which is impossible with any forseeable computer [42].

We have indicated that the basic mechanism for the generation of lift of a wing consists of counter-rotating rolls of low-pressure *streamwise vorticity* (swirling flow) generated by instability at separation, which reduce the high pressure on top of the wing before the trailing edge of potential flow and thus allow downwash, but which also generate drag. At closer examination of the quantitative distributions of lift and drag forces around the wing, we discover large lift at the expense of small drag resulting from leading edge suction, which answers the opening question of of how a wing can generate a lift/drag ratio larger than 10.

The secret of flight is in concise form uncovered in Fig. 174.3 showing G2 computed lift and and drag coefficients of a Naca 0012 3d wing as functions of the angle of attack α, as well as the circulation around the wing. We see that the lift and drag increase roughly linearly up to 16 degrees, with a lift/drag ratio of about 13 for $\alpha > 3$ degrees, and that lift peaks at stall at $\alpha = 20$ after a quick increase of drag. and flow separation at the leading edge.

We see that the circulation remains small for α less than 10 degrees without connection to lift, and conclude that the theory of lift of by Kutta-Zhukovsky is fictional without physical correspondence: There is lift but no circulation. Lift does not originate from circulation. The incorrect explanation by Kutta-Zhukovsky is illustrated in Fig. 174.2 which is found in books on flight aerodynamics.

Inspecting Figs. 174.4-174.6 showing velocity, pressure and vorticity and Fig. 174.7 showing lift and drag distributions over the upper and lower surfaces of the wing (allowing also pitching moment to be computed), we can now, with experience from the above preparatory analysis, identify the basic mechanisms for the generation of lift and drag in incompressible slightly viscous flow around a wing at different angles of attack α: We find two regimes before stall at $\alpha = 20$ with different, more or less linear growth in α of both lift and drag, a main phase $0 \leq \alpha < 16$ with the slope of the lift (coefficient) curve equal to 0.09 and of the drag curve equal to 0.008 with $L/D \approx 14$, and a final phase $16 \leq \alpha < 20$ with increased slope of both lift and drag. The main phase can be divided into an initial phase $0 \leq \alpha < 4 - 6$ and an intermediate phase $4 - 6 \leq \alpha < 16$, with somewhat smaller slope of drag in the initial phase. We now present details of this general picture.

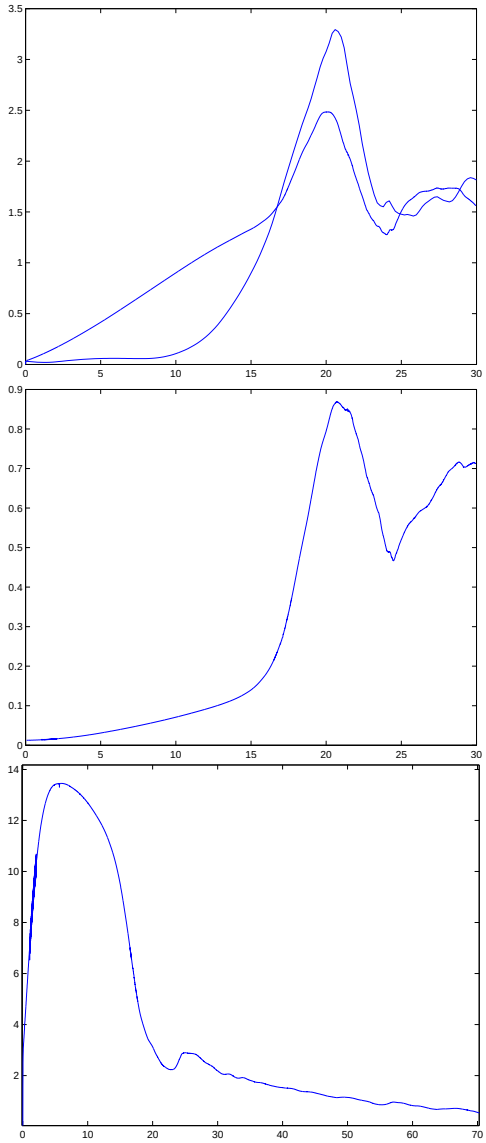


FIGURE 174.3. G2 lift coefficient and circulation as functions of the angle of attack (top), drag coefficient (middle) and lift/drag ratio (bottom) as functions of the angle of attack.

174.9 Phase 1: $0 \leq \alpha \leq 4-6$

At zero angle of attack with zero lift there is high pressure at the leading edge and equal low pressures on the upper and lower crests of the wing because the flow is essentially potential and thus satisfies Bernouilli's law of high/low pressure where velocity is low/high. The drag is about 0.01 and results from rolls of low-pressure streamwise vorticity attaching to the trailing edge. As α increases the low pressure below gets depleted as the incoming flow becomes parallel to the lower surface at the trailing edge for $\alpha = 6$, while the low pressure above intenisfies and moves towards the leading edge. The streamwise vortices at the trailing edge essentially stay constant in strength but gradually shift attachement towards the upper surface. The high pressure at the leading edge moves somewhat down, but contributes little to lift. Drag increases only slowly because of negative drag at the leading edge.

174.10 Phase 2: $4-6 \leq \alpha \leq 16$

The low pressure on top of the leading edge intensifies to create a normal gradient preventing separation, and thus creates lift by suction peaking on top of the leading edge. The slip boundary condition prevents separation and downwash is created with the help of the low-pressure wake of streamwise vorticity at rear separation. The high pressure at the leading edge moves further down and the pressure below increases slowly, contributing to the main lift coming from suction above. The net drag from the upper surface is close to zero because of the negative drag at the leading edge, known as *leading edge suction*, while the drag from the lower surface increases (linearly) with the angle of the incoming flow, with somewhat increased but still small drag slope. This explains why the line to a flying kite can be almost vertical even in strong wind, and that a thick wing can have less drag than a thin.

174.11 Phase 3: $16 \leq \alpha \leq 20$

This is the phase creating maximal lift just before stall in which the wing partly acts as a bluff body with a turbulent low-pressure wake attaching at the rear upper surface, which contributes extra drag and lift, doubling the slope of the lift curve to give maximal lift ≈ 2.5 at $\alpha = 20$ with rapid loss of lift after stall.

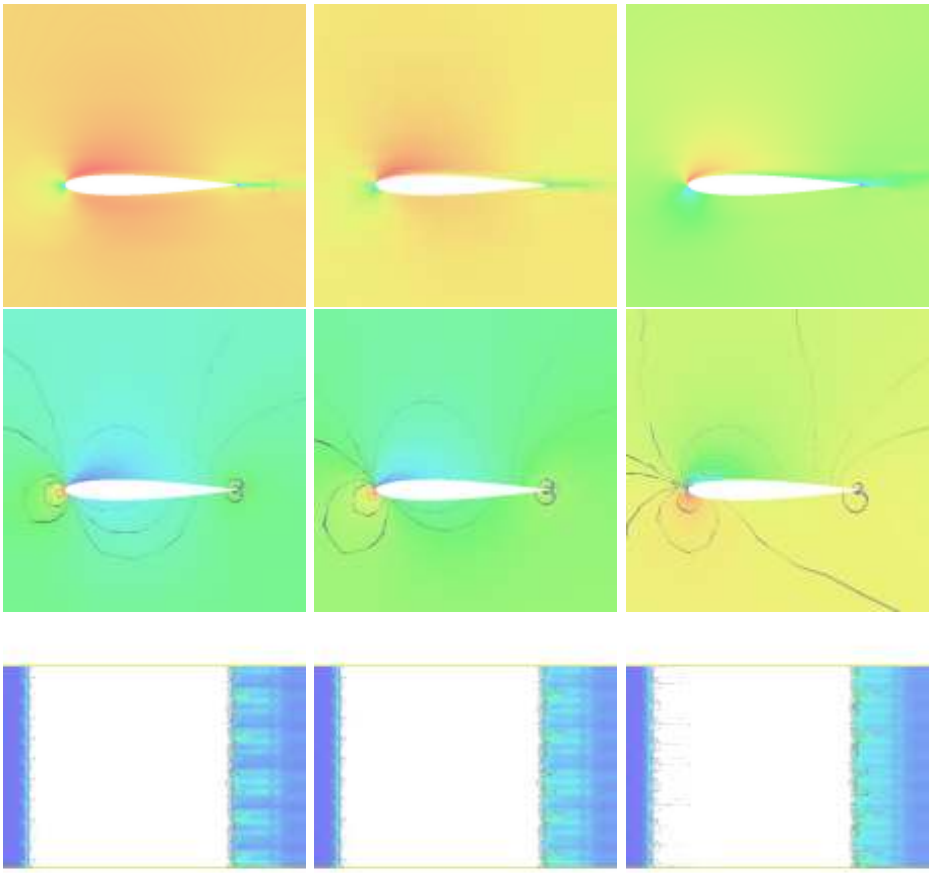


FIGURE 174.4. G2 computation of velocity magnitude (upper), pressure (middle), and non-transversal vorticity (lower), for angles of attack 2, 4, and 8° (from left to right). Notice in particular the rolls of streamwise vorticity at separation.

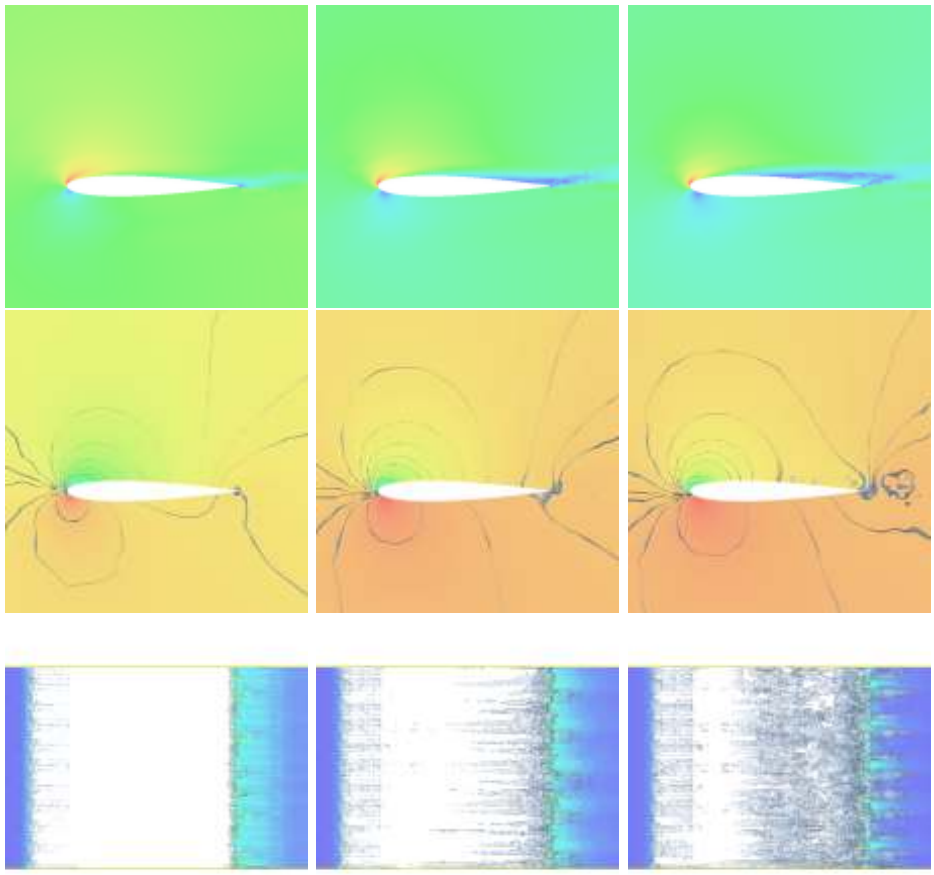


FIGURE 174.5. G2 computation of velocity magnitude (upper), pressure (middle), and topview of non-transversal vorticity (lower), for angles of attack 10, 14, and 18° (from left to right). Notice in particular the rolls of streamwise vorticity at separation.

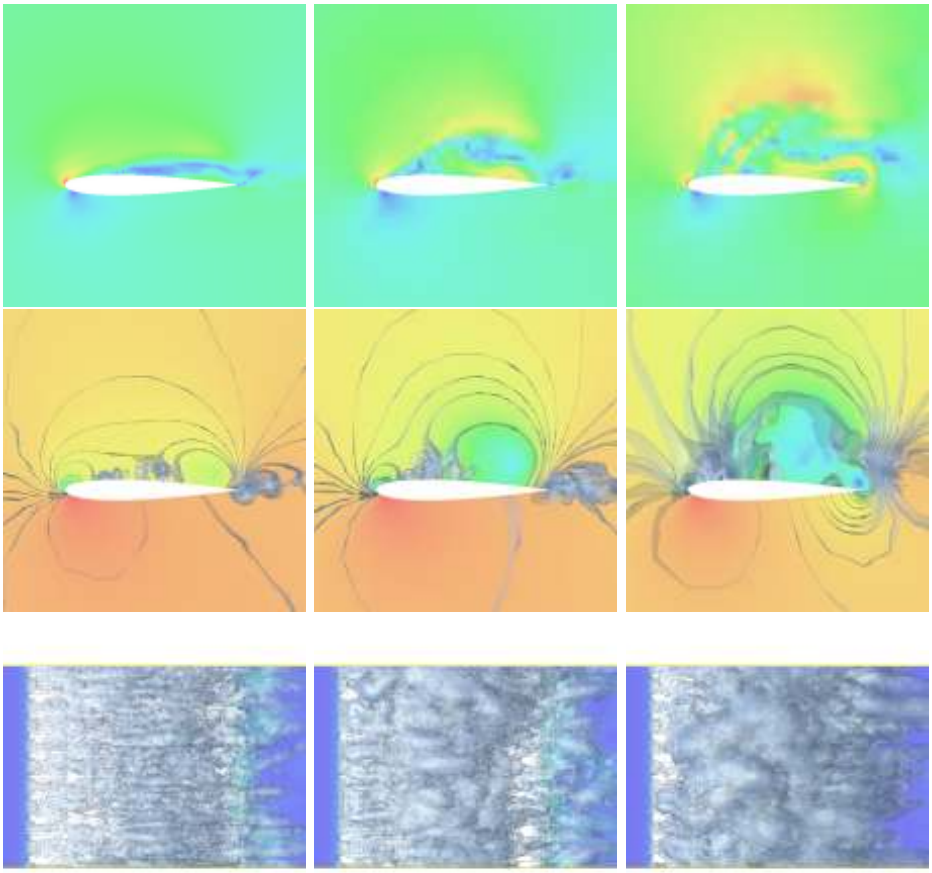


FIGURE 174.6. G2 computation of velocity magnitude (upper), pressure (middle), and non-transversal vorticity (lower), for angles of attack 20, 22, and 24° (from left to right).

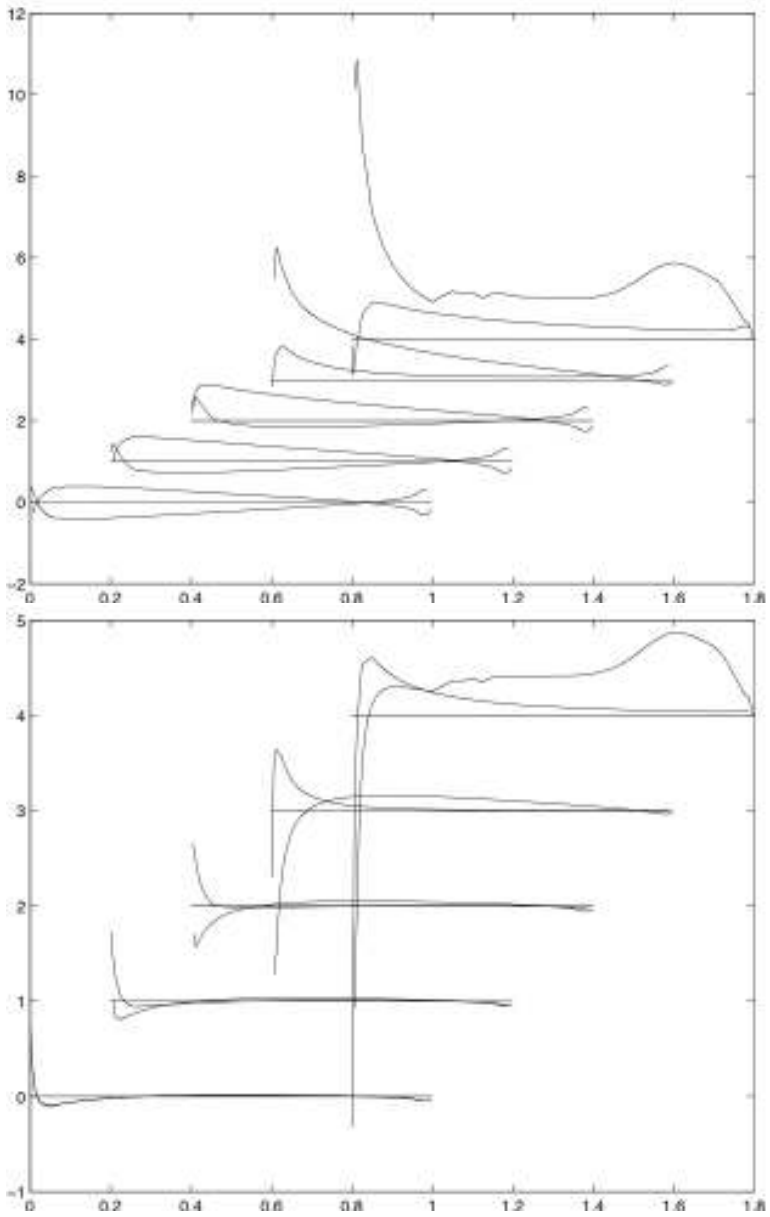


FIGURE 174.7. G2 computation of normalized local lift force (upper) and drag force (lower) contributions acting along the lower and upper parts of the wing, for angles of attack 0, 2 ,4 ,10 and 18°, each curve translated 0.2 to the right and 1.0 up, with the zero force level indicated for each curve.

174.12 Lift and Drag Distribution Curves

The distributions of lift and drag forces over the wing resulting from projecting the pressure acting perpendicular to the wing surface onto relevant directions, are plotted in Fig.174.7. The total lift and drag results from integrating these distributions around the wing. In potential flow computations (with circulation according to Kutta-Zhukovsky), only the pressure distribution or c_p-distribution is considered to carry releveant information, because a potential solution by construction has zero drag. In the perspective of Kutta-Zhukovsky, it is thus remarkable that the projected c_p-curves carry correct information for both lift and drag.

The lift generation in Phase 1 and 3 can rather easily be envisioned, while both the lift and drag in Phase 2 results from a (fortunate) intricate interplay of stability and instability of potential flow: The main lift comes from upper surface suction arising from a turbulent boundary layer with small skin friction combined with rear separation instability generating low-pressure streamwise vorticity, while the drag is kept small by negative drag from the leading edge.

174.13 Comparing Computation with Experiment

Comparing G2 computations with about 150 000 mesh points with experiments [20, 40], we find good agreement with the main difference that the boost of the lift coefficient in phase 3 is lacking in experiments. This is probably an effect of smaller Reynolds numbers in experiments, with a separation bubble forming on the leading edge reducing lift at high angles of attack. The oil-film pictures in [20] show surface vorticity generating streamwise vorticity in accordance with [26, 27, 23].

A jumbojet can only be tested in a wind tunnel as a smaller scale model, and upscaling test results is cumbersome because boundary layers do not scale. This means that computations can be closer to reality than wind tunnel experiments. Of particular importance is the maximal lift coefficient, which cannot be predicted by Kutta-Zhukovsky nor in model experiments, which for Boeing 737 is reported to be 2.73 in landing, corresponding to the maximal lift of 2.5 in computation for a long wing and not a full aircraft. In take-off the maximal lift is reported to be 1.75 with 1.5 in computation at a somewhat smaller angle of attack.

We compute turbulent solutions of the Navier-Stokes equations using a stabilized finite element method with *a posteriori error control* of lift and drag, referred to as *General Galerkin* or *G2*, available in executable open source from [25]. The stabilization in G2 acts as an automatic turbulence model, and thus offers a model for *ab initio* computational simulation of the turbulent flow around a wing with the only input being the geometry

of the wing. The computations performed on a single workstation show good agreement with experiments. We are now performing computations on super-computers allowing more precise comparisons and paramater studies.

174.14 Navier-Stokes with Force Boundary Conditions

For the reader interested in the mathematics we now present the Navier-Stokes equations along with a stability analysis exhibiting the basic instability mechanism at separation which is crucial for the generation of lift, at the expense of some drag.

The Navier-Stokes equations for an incompressible fluid of unit density with *small viscosity* $\nu > 0$ and *small skin friction* $\beta \geq 0$ filling a volume Ω in $\mathbb{R}^3$ surrounding a solid body with boundary Γ over a time interval $I = [0, T]$, read as follows: Find the velocity $u = (u_1, u_2, u_3)$ and pressure p depending on $(x, t) \in \Omega \cup \Gamma \times I$, such that

$$\begin{aligned} \dot{u} + (u \cdot \nabla)u + \nabla p - \nabla \cdot \sigma &= f && \text{in } \Omega \times I, \\ \nabla \cdot u &= 0 && \text{in } \Omega \times I, \\ u_n &= g && \text{on } \Gamma \times I, \\ \sigma_s &= \beta u_s && \text{on } \Gamma \times I, \\ u(\cdot, 0) &= u^0 && \text{in } \Omega, \end{aligned} \tag{174.1}$$

where u_n is the fluid velocity normal to Γ, u_s is the tangential velocity, $\sigma = 2\nu\epsilon(u)$ is the viscous (shear) stress with $\epsilon(u)$ the usual velocity strain, σ_s is the tangential stress, f is a given volume force, g is a given inflow/outflow velocity with $g = 0$ on a non-penetrable boundary, and u^0 is a given initial condition. We notice the skin friction boundary condition coupling the tangential stress σ_s to the tangential velocity u_s with the friction coefficient β with $\beta = 0$ for slip, and $\beta >> 1$ for no-slip. We note that β is related to the standard *skin friction coeffieient* $c_f = \frac{2\tau}{U^2}$ with τ the tangential stress per unit area, by the relation $\beta = \frac{U}{2} c_f$. In particular, β tends to zero with c_f (if U stays bounded).

Prandtl insisted on using a no-slip velocity boundary condition with $u_s = 0$ on Γ, because his resolution of d'Alembert's paradox hinged on discriminating potential flow by this condition. On the oher hand, with the new resolution of d'Alembert's paradox, relying instead on instability of potential flow, we are free to choose instead a friction force boundary condition, if data is available. Now, experiments show [47, 22] that the skin friction coefficient decreases with increasing Reynolds number Re as $c_f \sim Re^{-0.2}$, so that $c_f \approx 0.0005$ for $Re = 10^{10}$ and $c_f \approx 0.007$ for $Re = 10^5$. Accordingly we model a turbulent boundary layer by a friction boundary condition with a friction parameter $\beta \approx 0.03 U Re^{-0.2}$. For very

large Reynolds numbers, we can effectively use $\beta = 0$ in G2 computation corresponding to slip boundary conditions.

As developed in more detail in [27], we make a distinction between laminar (boundary layer) separation modeled by no-slip and turbulent (boundary layer) separation modeled by slip/small friction. Note that laminar separation cannot be modeled by slip, since a laminar boundary layer needs to be resolved with no-slip to get correct (early) separation. On the other hand, as will be seen below, in turbulent (but not in laminar) flow the interior turbulence dominates the skin friction turbulence indicating that the effect of a turbulent boundary layer can be modeled by slip/small friction, which can be justified by an posteriori sensitivity analysis as shown in [27].

We thus assume that the boundary layer is turbulent and is modeled by slip/small friction, which effectively includes the case of laminar separation followed by reattachment into a turbulent boundary layer.

174.15 Potential Flow

Potential flow (u, p) with velocity $u = \nabla\varphi$, where φ is harmonic in Ω and satisfies a homogeneous Neumann condition on Γ and suitable conditions at infinity, can be seen as a solution of the Navier-Stokes equations for slightly viscous flow with slip boundary condition, subject to

- perturbation of the volume force $f = 0$ in the form of $\sigma = \nabla\cdot(2\nu\epsilon(u))$,
- perturbation of zero friction in the form of $\sigma_s = 2\nu\epsilon(u)_s$,

with both perturbations being small because ν is small and a potential flow velocity u is smooth. Potential flow can thus be seen as a solution of the Navier-Stokes equations with small force perturbations tending to zero with the viscosity. We can thus express d'Alembert's paradox as the zero lift/drag of a Navier-Stokes solution in the form of a potential solution, and resolve the paradox by realizing that potential flow is unstable and thus cannot be observed as a physical flow.

Potential flow is like an inverted pendulum, which cannot be observed in reality because it is unstable and under infinitesimal perturbations turns into a swinging motion. A stationary inverted pendulum is a fictious mathematical solution without physical correspondence because it is unstable. You can only observe phenomena which in some sense are stable, and an inverted pendelum or potential flow is not stable in any sense.

Potential flow has the following crucial property which partly will be inherited by real turbulent flow, and which explains why a flow over a wing subject to small skin friction can avoid separating at the crest and thus generate downwash, unlike viscous flow with no-slip, which separates at the crest without downwash. We will conclude that gliding flight is possible

only in slightly viscous incompressible flow. For simplicity we consider two-dimensional potential flow around a cylindrical body such as a long wing (or cylinder).

Theorem. Let φ be harmonic in the domain Ω in the plane and satisfy a homogeneous Neumann condition on the smooth boundary Γ of Ω. Then the streamlines of the corresponding velocity $u = \nabla\varphi$ can only separate from Γ at a point of stagnation with $u = \nabla\varphi = 0$.

Proof. Let ψ be a harmonic conjugate to φ with the pair (φ, ψ) satisfying the Cauchy-Riemann equations (locally) in Ω. Then the level lines of ψ are the streamlines of φ and vice versa. This means that as long as $\nabla\varphi \neq 0$, the boundary curve Γ will be a streamline of u and thus fluid particles cannot separate from Γ in bounded time.

174.16 Exponential Instability

Subtracting the NS equations with $\beta = 0$ for two solutions (u, p, σ) and $(\bar{u}, \bar{p}, \bar{\sigma})$ with corresponding (slightly) different data, we obtain the following linearized equation for the difference $(v, q, \tau) \equiv (u - \bar{u}, p - \bar{p}, \sigma - \bar{\sigma})$ with :

$$\begin{aligned} \dot{v} + (u\cdot\nabla)v + (v\cdot\nabla)\bar{u} + \nabla q - \nabla\cdot\tau &= f - \bar{f} && \text{in } \Omega\times I,\\ \nabla\cdot v &= 0 && \text{in } \Omega\times I,\\ v\cdot n &= g - \bar{g} && \text{on } \Gamma\times I,\\ \tau_s &= 0 && \text{on } \Gamma\times I,\\ v(\cdot,0) &= u^0 - \bar{u}^0 && \text{in } \Omega, \end{aligned} \tag{174.2}$$

Formally, with u and $\bar{u}$ given, this is a linear convection-reaction-diffusion problem for (v, q, τ) with the reaction term given by the 3×3 matrix $\nabla\bar{u}$ being the main term of concern for stability. By the incompressiblity, the trace of $\nabla\bar{u}$ is zero, which shows that in general $\nabla\bar{u}$ has eigenvalues with real value of both signs, of the size of $|\nabla\bar{u}|$ (with $|\cdot|$ som matrix norm), thus with at least one exponentially unstable eigenvalue.

Accordingly, we expect local exponential perturbation growth of size $\exp(|\nabla u|t)$ of a solution (u, p, σ), in particular we expect a potential solution to be illposed. This is seen in G2 solutions with slip initiated as potential flow, which subject to residual perturbations of mesh size h, in $\log(1/h)$ time develop into turbulent solutions. We give computational evidence that these turbulent solutions are wellposed, which we rationalize by cancellation effects in the linearized problem, which has rapidly oscillating coefficients when linearized at a turbulent solution.

Formally applying the curl operator $\nabla\times$ to the momentum equation of (174.1), with $\nu = \beta = 0$ for simplicity, we obtain the *vorticity equation*

$$\dot{\omega} + (u\cdot\nabla)\omega - (\omega\cdot\nabla)u = \nabla\times f \quad \text{in } \Omega, \tag{174.3}$$

which is a convection-reaction equation in the vorticity $\omega = \nabla \times u$ with coefficients depending on u, of the same form as the linearized equation (174.2), with similar properties of exponential perturbation growth $\exp(|\nabla u|t)$ referred to as *vortex stretching*. Kelvin's theorem formally follows from this equation assuming the initial vorticity is zero and $\nabla \times f = 0$ (and $g = 0$), but exponential perturbation growth makes this conclusion physically incorrect: We will see below that large vorticity can develop from irrotational potential flow even with slip boundary conditions.

174.17 Energy Estimate with Turbulent Dissipation

The standard *energy estimate* for (174.1) is obtained by multiplying the momentum equation

$$\dot{u} + (u \cdot \nabla)u + \nabla p - \nabla \cdot \sigma - f = 0,$$

with u and integrating in space and time, to get in the case $f = 0$ and $g = 0$,

$$\int_0^t \int_\Omega R_\nu(u,p) \cdot u \, dxdt = D_\nu(u;t) + B_\beta(u;t) \tag{174.4}$$

where

$$R_\nu(u,p) = \dot{u} + (u \cdot \nabla)u + \nabla p$$

is the *Euler residual* for a given solution (u,p) with $\nu > 0$,

$$D_\nu(u;t) = \int_0^t \int_\Omega \nu |\epsilon(u(\bar{t},x))|^2 dxd\bar{t}$$

is the *internal turbulent viscous dissipation*, and

$$B_\beta(u;t) = \int_0^t \int_\Gamma \beta |u_s(\bar{t},x)|^2 dxd\bar{t}$$

is the *boundary turbulent viscous dissipation*, from which follows by standard manipulations of the left hand side of (174.4),

$$K_\nu(u;t) + D_\nu(u;t) + B_\beta(u;t) = K(u^0), \quad t > 0, \tag{174.5}$$

where

$$K_\nu(u;t) = \frac{1}{2}\int_\Omega |u(t,x)|^2 dx.$$

This estimate shows a balance of the *kinetic energy* $K(u;t)$ and the *turbulent viscous dissipation* $D_\nu(u;t) + B_\beta(u;t)$, with any loss in kinetic energy appearing as viscous dissipation, and vice versa. In particular,

$$D_\nu(u;t) + B_\beta(u;t) \le K(u^0),$$

and thus the viscous dissipation is bounded (if $f = 0$ and $g = 0$).

Turbulent solutions of (174.1) are characterized by *substantial internal turbulent dissipation*, that is (for t bounded away from zero),

$$D(t) \equiv \lim_{\nu \to 0} D(u_\nu; t) >> 0, \tag{174.6}$$

which is *Kolmogorov's conjecture* [18]. On the other hand, the *skin friction dissipation* decreases with decreasing friction

$$\lim_{\nu \to 0} B_\beta(u; t) = 0, \tag{174.7}$$

since $\beta \sim \nu^{0.2}$ tends to zero with the viscosity ν and the tangential velocity u_s approaches the (bounded) free-stream velocity. We thus find evidence that the interior turbulent dissipation dominates the skin friction dissipation, which supports the use of slip as a model of a turbulent boundray layer, but which is not in accordance with Prandtl's (unproven) conjecture that substantial drag and turbulent dissipation originates from the boundary layer.

Kolmogorov's conjecture (174.6) is consistent with

$$\|\nabla u\|_0 \sim \frac{1}{\sqrt{\nu}}, \quad \|R_\nu(u,p)\|_0 \sim \frac{1}{\sqrt{\nu}}, \tag{174.8}$$

where $\|\cdot\|_0$ denotes the $L_2(Q)$-norm with $Q = \Omega \times I$. On the other hand, it follows by standard arguments from (204.4) that

$$\|R_\nu(u,p)\|_{-1} \leq \sqrt{\nu}, \tag{174.9}$$

where $\|\cdot\|_{-1}$ is the norm in $L_2(I; H^{-1}(\Omega))$. Kolmogorov thus conjectures that the Euler residual $R_\nu(u,p)$ for small ν is strongly (in L_2) large, while being small weakly (in H^{-1}).

Altogether, we understand that the resolution of d'Alembert's paradox of explaining substantial drag from vanishing viscosity, consists of realizing that the internal turbulent dissipation D can be positive under vanishing viscosity, while the skin friction dissipation B will vanish. In contradiction to Prandtl, we conclude that drag does not result from boundary layer effects, but from internal turbulent dissipation, originating from instability at separation.

174.18 G2 Computational Solution

We show in [4, 26, 30] that the Navier-Stokes equations (174.1) can be solved by G2 producing turbulent solutions characterized by substantial turbulent dissipation from the least squares stabilization acting as an automatic turbulence model, reflecting that the Euler residual cannot be made

pointwise small in turbulent regions. G2 has a posteriori error control based on duality and shows output uniqueness in mean-values such as lift and drag [4, 23, 24]

We find that G2 with slip is capable of modeling slightly viscous turbulent flow with $Re > 10^6$ of relevance in many applications in aero/hydro dynamics, including flying, sailing, boating and car racing, with hundred thousands of mesh points in simple geometry and millions in complex geometry, while according to state-of-the-art quadrillions is required [42]. This is because a friction-force/slip boundary condition can model a turbulent boundary layer, and interior turbulence does not have to be resolved to physical scales to capture mean-value outputs [4].

The idea of circumventing boundary layer resolution by relaxing no-slip boundary conditions introduced in [23, 4], was used in [39, 5] in the form of weak satisfaction of no-slip, which however misses the main point of using a force condition instead of a velocity condition in a model of a turbulent boundary layer.

A G2 solution (U, P) on a mesh with local mesh size $h(x,t)$ according to [4], satisfies the following energy estimate (with $f = 0$, $g = 0$ and $\beta = 0$):

$$K(U(t)) + D_h(U;t) = K(u^0), \tag{174.10}$$

where

$$D_h(U;t) = \int_0^t \int_\Omega h|R_h(U,P)|^2\, dxdt, \tag{174.11}$$

is an analog of $D_\nu(u;t)$ with $h \sim \nu$, where $R_h(U,P)$ is the Euler residual of (U,P). We see that the G2 turbulent viscosity $D_h(U;t)$ arises from penalization of a non-zero Euler residual $R_h(U,P)$ with the penalty directly connecting to the violation (according the theory of criminology). A turbulent solution is characterized by substantial dissipation $D_h(U;t)$ with $\|R_h(U,P)\|_0 \sim h^{-1/2}$, and

$$\|R_h(U,P)\|_{-1} \leq \sqrt{h} \tag{174.12}$$

in accordance with (174.8) and (174.9).

174.19 Wellposedness of Mean-Value Outputs

Let $M(v) = \int_Q v\psi\, dxdt$ be a *mean-value output* of a velocity v defined by a smooth weight-function $\psi(x,t)$, and let (u,p) and (U,P) be two G2-solutions on two meshes with maximal mesh size h. Let (φ,θ) be the solution to the *dual linearized problem*

$$\begin{array}{rcll} -\dot\varphi - (u\cdot\nabla)\varphi + \nabla U^\top\varphi + \nabla\theta &=& \psi & \text{in } \Omega\times I,\\ \nabla\cdot\varphi &=& 0 & \text{in } \Omega\times I,\\ \varphi\cdot n &=& g & \text{on } \Gamma\times I,\\ \varphi(\cdot,T) &=& 0 & \text{in } \Omega, \end{array} \tag{174.13}$$

where $\top$ denotes transpose. Multiplying the first equation by $u - U$ and integrating by parts, we obtain the following output error representation [4]:

$$M(u) - M(U) = \int_Q (R_h(u,p) - R_h(U,P)) \cdot \varphi \, dxdt \tag{174.14}$$

where for simplicity the dissipative terms are here omitted, from which follows the a posteriori error estimate:

$$|M(u) - M(U)| \leq S(\|R_h(u,p)\|_{-1} + \|R_h(U,P)\|_{-1}), \tag{174.15}$$

where the stability factor

$$S = S(u,U,M) = S(u,U) = \|\varphi\|_{H^1(Q)}. \tag{174.16}$$

In [4] we present a variety of evidence, obtained by computational solution of the dual problem, that for global mean-value outputs such as drag and lift, $S << 1/\sqrt{h}$, while $\|R_h\|_{-1} \sim \sqrt{h}$, allowing computation of of drag/lift with a posteriori error control of the output within a tolerance of a few percent. In short, mean-value outputs such as lift and drag are wellposed and thus physically meaningful.

We explain in [4] the crucial fact that $S << 1/\sqrt{h}$, heuristically as an effect of *cancellation* of rapidly oscillating reaction coefficients of turbulent solutions combined with smooth data in the dual problem for mean-value outputs. In smooth potential flow there is no cancellation, which explains why zero lift/drag cannot be observed in physical flows.

As an example, we show in Fig.174.8 turbulent G2 flow around a car with substantial drag in qualitative accordance with wind-tunnel experiments. We see a pattern of streamwise vorticity forming in the rear wake. We also see surface vorticity forming on the hood transversal to the main flow direction. We see similar features in the flow of air around a wing.

174.20 Scenario for Separation without Stagnation

We now present a scenario for transition of potential flow into turbulent flow, based on identifying perturbations of strong growth in the linearized equations (174.2) and (174.3) at separation generating rolls of low pressure streamwise vorticity changing the pressure distribution to give both lift and drag of a wing.

As a model of potential flow at rear separation, we consider the potential flow $u(x) = (x_1, -x_2, 0)$ in the half-plane $\{x_1 > 0\}$. Assuming x_1 and x_2 are small, we approximate the v_2-equation of (174.2) by

$$\dot{v}_2 - v_2 = f_2,$$

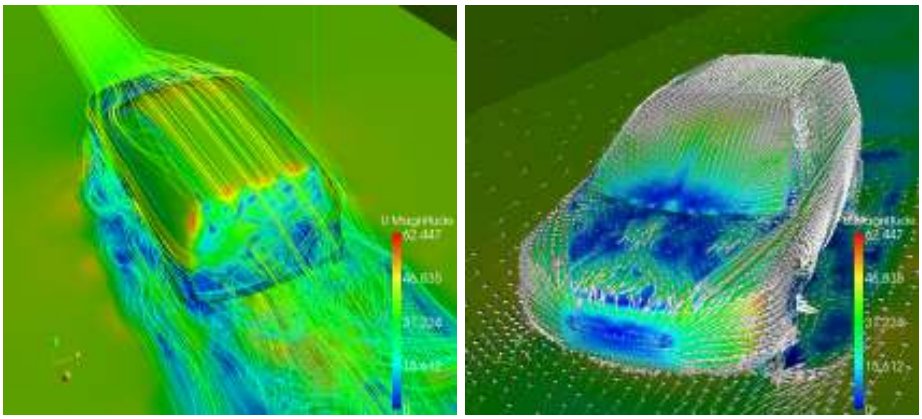


FIGURE 174.8. Velocity of turbulent G2 flow with slip around a car with courtesy of geometry Volvo Cars and computations by Murtazo Nasarov.

where $f_2 = f_2(x_3)$ is an oscillating mesh residual perturbation depending on x_3 (including also a pressure-gradient), for example $f_2(x_3) = h\sin(x_3/\delta)$, with $\delta > 0$. It is natural to assume that the amplitude of f_2 decreases with δ. We conclude, assuming $v_2(0, x) = 0$, that

$$v_2(t, x_3) = t\exp(t)f_2(x_3),$$

and for the discussion, we assume $v_3 = 0$. Next we approximate the ω_1-vorticity equation for x_2 small and $x_1 \geq \bar{x}_1 > 0$ with $\bar{x}_1$ small, by

$$\dot{\omega}_1 + x_1\frac{\partial\omega_1}{\partial x_1} - \omega_1 = 0,$$

with the "inflow boundary condition"

$$\omega_1(\bar{x}_1, x_2, x_3) = \frac{\partial v_2}{\partial x_3} = t\exp(t)\frac{\partial f_2}{\partial x_3}.$$

The equation for ω_1 thus exhibits exponential growth, which is combined with exponential growth of the "inflow condition". We can see these features in Fig. 174.9 showing how opposing flows on the back generate a pattern of co-rotating surface vortices which act as initial conditions for vortex stretching into the fluid generating rolls of low-pressure streamwise vorticity, in the case of a wing attaching to the trailing edge.

Altogether we expect $\exp(t)$ perturbation growth of residual perturbations of size h, resulting in a global change of the flow after time $T \sim \log(1/h)$, which can be traced in the computations.

We thus understand that the formation of streamwise streaks as the result of a force perturbation oscillating in the x_3 direction, which in the retardation of the flow in the x_2-direction creates exponentially increasing

vorticity in the x_1-direction, which acts as inflow to the ω_1-vorticity equation with exponential growth by vortex stretching. Thus, we find exponential growth at rear separation in both the retardation in the x_2-direction and the accelleration in the x_1 direction. This scenario is illustrated in principle and computation in Fig.174.9. Note that since the perturbation is convected with the base flow, the absolute size of the growth is related to the length of time the perturbation stays in a zone of exponential growth. Since the combined exponential growth is independent of δ, it follows that large-scale perturbations with large amplitude have largest growth, which is also seen in computations with δ the distance between streamwise rolls as seen in Fig.174.3 which does not seem to decrease with decreasing h.

Notice that at forward attachment of the flow the retardation does not come from opposing flows, and the zone of exponential growth of ω_2 is short, resulting in much smaller perturbation growth than at rear separation.

We can view the occurence of the rear surface vorticities as a mechanism of separation with non-zero tangential speed, by diminishing the normal pressure gradient of potential flow, which allows separation only at stagnation. The surface vorticities thus allow separation without stagnation but the price is generation of a system of low-pressure tubes of streamwise vorticity creating drag in a form of "separation trauma" or "cost of divorce".

The scenario for separation can summarized as follows: Velocity instability in retardation as opposing flows meet in the rear of the cylinder, generates a zig-zag pattern of surface vorticity shown in Fig.174.9, allowing separation into counter-rotating low-pressure rolls, attaching to the trailing edge in the case of a wing, as shown in Fig. 174.1.

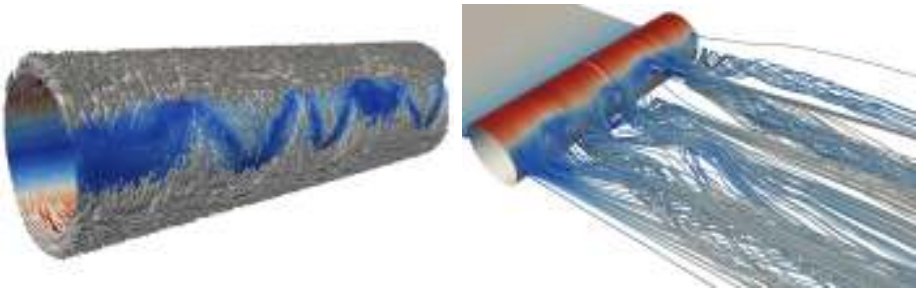


FIGURE 174.9. Turbulent separation by surface vorticity forming counter-rotating low-pressure rolls in flow around a circular cylinder, illustrating separation at the trailing edge of a wing [23].

174.21 Stability of the Streamwise Vorticity Perturbed Flow

The rolls of streamwise vorticity swirling flow appearing at separation because of the instability of potential flow represent a more stable flow pattern. An indication of stability is given by an analysis of the stability of the rotating flow field $u = (0, x_3, -x_2)$ with linearized problem of the form

$$\dot{v}_2 + v_3 = 0, \quad \dot{v}_3 - v_2 = 0 \tag{174.17}$$

which does not have any exponentially unstable solutions. The swirling flow at separation is similar to the vortex seen at the drain of a bathtub.

174.22 Sailing

Both the sail and keel of a sailing boat under tacking against the wind, act like wings generating lift and drag, but the action, geometrical shape and angle of attack of the sail and the keel are different. The effective angle of attack of a sail is typically 15-20 degrees and that of a keel 5-10 degrees, for reasons which we now give.

The boat is pulled forward by the sail, assuming for simplicity that the beam is parallel to the direction of the boat at a minimal tacking angle, by the component $L\sin(15)$ of the lift L, as above assumed to be perpendicular to the effective wind direction, but also by the following contributions from the drag assumed to be parallel to the effective wind direction: The negative drag on the leeeward side at the leading edge close to the mast gives a positive pull which largely compensates for the positive drag from the rear leeward side, while there is less positive drag from the windward side of the sail as compared to a wing profile, because of the difference in shape. The result is a forward pull $\approx \sin(15)L \approx 0.2L$ combined with a side (heeling) force $\approx L\cos(15) \approx L$, which tilts the boat and needs to be balanced by lift from the the keel in the opposite direction. Assuming the lift/drag ratio for the keel is 13, the forward pull is then reduced to $\approx (0.2 - 1/13)L \approx 0.1L$, which can be used to overcome the drag from the hull minus the keel.

The shape of a sail is different from that of a wing which gives smaller drag from the windward side and thus improved forward pull, while the keel has the shape of a symmetrical wing and acts like a wing. A sail with aoa $15 - 20$ degrees gives maximal pull forward at maximal heeling/lift with contribution also from the rear part of the sail, like for a wing just before stall, while the drag is smaller than for a wing at 15-20 degrees aoa (for which the lift/drag ratio is 4-3), with the motivation given above. The lift/drag curve for a sail is thus different from that of wing with lift/drag ratio at aoa 15-20 much larger for a sail. On the other hand, a keel with aoa 5-10 degrees has a lift/drag ratio about 13. A sail at aoa 15-20 thus

gives maximal pull at strong heeling force and small drag, which together with a keel at aoa 5-10 with strong lift and small drag, makes an efficient combination. This explains why modern designs combine a deep narrow keel acting efficiently for small aoa, with a broader sail acting efficiently at a larger aoa.

Using a symmetrical wing as a sail would be inefficient, since the lift/drag ratio is poor at maximal lift at aoa 15-20. On the other hand, using a sail as a wing can only be efficient at a large angle of attack, and thus is not suitable for cruising. This material is developed in more detail in [30].

References

[1] How Airplanes Fly: A Physical Description of Lift, http://www.allstar.fiu.edu/ aero/airflylvl3.htm.

[2] J. D. Anderson, *A History of Aerodynamics*, Cambridge Aerospace Series 8, Cambridge University Press, 1997.

[3] http://www.aviation-history.com/theory/lift.htm

[4] AVweb, http://www.avweb.com/news/airman/183261-1.html.

[5] Y. Bazilevs, C. Michler, V.M. Calo and T.J.R. Hughes, Turbulence without Tears: Residual-Based VMS, Weak Boundary Conditions, and Isogeometric Analysis of Wall-Bounded Flows, Preprint 2008.

[6] W. Beaty, Airfoil Lifting Force Misconception Widespread in K-6 Textbooks, Science Hobbyist, http://www.eskimo.com/ billb/wing/airfoil.htm#L1.

[7] Garret Birkhoff, *Hydrodynamics: a study in logic, fact and similitude*, Princeton University Press, 1950.

[8] K. Chang, Staying Aloft; What Does Keep Them Up There?, New York Times, Dec 9, 2003.

[9] S. Cowley, Laminar boundary layer theory: A 20th century paradox, Proceedings of ICTAM 2000, eds. H. Aref and J.W. Phillips, 389-411, Kluwer (2001).

[10] G. M. Craig, Stop Abusing Bernoulli! - How Airplanes Really Fly, Regenerative Press, 1998.

[11] A. Crook, Skin friction estimation at high Reynolds numbers and Reynolds-number effects for transport aircraft, Center for Turbulence Research, 2002.

[12] D'Alembert's paradox, en.wikipedia.org/wiki/D'Alembert's_paradox.

[13] 3rd CFD AIAA Drag Prediction Workshop, aaac.larc.nasa.gov/tfcab/cfdlarc/aiaa-dpw.

[14] O. Darrigol, World of Flow, A History of hydrodynamics from the Bernouillis to Prandtl, Oxford University Press.

[15] J. Hadamard, Sur les problmes aux drives partielles et leur signification physique, Princeton University Bulletin, 49–52, 1902.

[16] A. Fage and L.F. Simmons, An investigation of the air-flow pattern in the wake of an airfoil of finite span, Rep.Memor.aero.REs.Coun.,Lond.951, 1925.

[17] The FEniCS Project, www.fenics.org.

[18] U. Frisch, Turbulence: The Legacy of A. N. Kolmogorov. Cambridge University Press, 1995.

[19] S. Goldstein, Fluid mechanics in the first half of this century, in Annual Review of Fluid Mechanics, Vol 1, ed. W. R. Sears and M. Van Dyke, pp 1-28, Palo Alto, CA: Annuak Reviews Inc.

[20] N. Gregory and C.L. O'Reilly, Low-Speed Aerodynamic Characteristics of NACA 0012 Aerofoil Section, including the Effects of Upper-Surface Roughness Simulating Hoar Frost, Aeronautical Research Council Reports and Memoranda, http://aerade.cranfield.ac.uk/ara/arc/rm/3726.pdf.

[21] J.Hoffman, Simulation of turbulent flow past bluff bodies on coarse meshes using General Galerkin methods: drag crisis and turbulent Euler solutions, Comp. Mech. 38 pp.390-402, 2006.

[22] J. Hoffman, Simulating Drag Crisis for a Sphere using Friction Boundary Conditions, Proc. ECCOMAS, 2006.

[23] J.Hoffman and Niclas Jansson, A computational study of turbulent flow separation for a circular cylinder using skin friction boundary conditions, Proc. Quality and Reliability of Large Eddy Simulation, Pisa, 2009.

[24] J. Hoffman and C. Johnson, Blowup of Euler solutions, BIT Numerical Mathematics, Vol 48, No 2, 285-307.

[25] J. Hoffman and C. Johnson, *Computational Turbulent Incompressible Flow*, Springer, 2007, www.bodysoulmath.org/books.

[26] J. Hoffman and C. Johnson, Resolution of d'Alembert's paradox, Journal of Mathematical Fluid Mechanics, Online First, Dec 10, 2008.

[27] J. Hoffman and C. Johnson, Separation in slightly viscous flow, to appear.

[28] J. Hoffman and C. Johnson, Is the Clay Navier-Stokes Problem Well-posed?, NADA 2008, http://www.nada.kth.se/∼cgjoh/hadamard.pdf, http://knol.google.com/k/the-clay-navier-stokes-millennium-problem.

[29] J. Hoffman and C. Johnson, The Secret of Flight, http://www.nada.kth.se/∼ cgjoh/ambsflying.pdf

[30] J. Hoffman and C. Johnson, The Secret of Sailing, http://www.nada.kth.se/∼ cgjoh/ambssailing.pdf

[31] J. Hoffman and C. Johnson, Movies of take-off of Naca0012 wing, http://www.csc.kth.se/ctl

[32] http://knol.google.com/k/claes-johnson/dalemberts-paradox/yvfu3xg7d7wt/2.

[33] http://knol.google.com/k/claes-johnson/why-it-is-possible-to-fly/yvfu3xg7d7wt/18.

[34] HowStuffWorks, http://science.howstuffworks.com/airplane7.htm.

[35] G.S. Jones, J.C. Lin, B.G. Allan, W.E. Milholen, C.L. Rumsey,R.C. Swanson, Overview of CFD Validation Experiments for Circulation Control Applications at NASA.

[36] R. Kunzig, An old, lofty theory of how airplanes fly loses some altitude, Discover, Vol. 22 No. 04, April 2001.

[37] F. W. Lanchester, *Aerodynamics*, 1907.

[38] Experiments in Aerodynamics, Smithsonian Contributions to Knowledge no. 801, Washinton, DC, Smithsonian Institution.

[39] W. Layton, Weak imposition of no-slip boundary conditions in finite element methods, Computers and Mathematics with Applications, 38 (1999), pp. 129142.

[40] W. J. McCroskey, A Critical Assessment of Wind Tunnel Results for the NACA 0012 Airfoil, NASA Technical Memorandum 10001, Technical Report 87-A-5, Aeroflightdynamics Directorate, U.S. Army Aviation Research and Technology Activity, Ames Research Center, Moffett Field, California.

[41] http://web.mit.edu/16.00/www/aec/flight.html.

[42] P. Moin and J. Kim, Tackling Turbulence with Supercomputers, Scientific American Magazine, 1997.

[43] http://www.grc.nasa.gov/WWW/K-12/airplane/lift1.html.

[44] http://www.planeandpilotmag.com/aircraft/specifications/diamond/2007-diamond-star-da40-xl/289.html

[45] L. Prandtl, On Motion of Fluids with Very Little, in Verhandlungen des dritten internationalen Mathematiker-Kongresses in Heidelberg 1904, A. Krazer, ed., Teubner, Leipzig, Germany (1905), p. 484. English trans. in Early Developments of Modern Aerodynamics, J. A. K. Ackroyd, B.P. Axcell, A.I. Ruban, eds., Butterworth-Heinemann, Oxford, UK (2001), p. 77.

[46] L. Prandtl and O Tietjens, *Applied Hydro- and Aeromechanics*, 1934.

[47] H. Schlichting, Boundary Layer Theory, McGraw-Hill, 1979.

[48] K. Stewartson, D'Alembert's Paradox, SIAM Review, Vol. 23, No. 3, 308-343. Jul., 1981.

[49] B. Thwaites (ed), *Incompressible Aerodynamics*, An Account of the Theory and Observation of the Steady Flow of Incompressible Fluid pas Aerofoils, Wings and other Bodies, Fluid Motions Memoirs, Clarendon Press, Oxford 1960, Dover 1987, p 94.

[50] R. von Mises, Theory of Flight, McGraw-Hill, 1945.

[51] D. You and P. Moin, Large eddy simulation of separation over an airfoil with synthetic jet control, Center for Turbulence Research, 2006.

175
The Secret of Thermodynamics

ABSTRACT We test the functionality of FEniCS on the challenge of computational thermodynamics in the form of the EG2 finite element solver of the Euler equations expressing conservation of mass, momentum and energy. We show that computational solutions satisfy a 2nd Law formulated in terms of kinetic energy, internal (heat) energy, work and shock/turbulent dissipation, without reference to entropy. We show that the 2nd Law expresses an irreversible transfer of kinetic energy to heat energy in shock/turbulent dissipation arising because the Euler equations lack pointwise solutions, and thus explains the occurence of irreversibility in formally reversible systems as an effect of instability with blow-up of Euler residuals combined with finite precision computation, without resort to statistical mechanics or ad hoc viscous regularization. We simulate the classical Joule or Joule-Thompson experiment of a gas expanding from rest under temperature drop followed by temperature recovery by turbulent dissipation until rest in the double volume. We present the FEniCS implementation of EG2 including applications to bluff body flow.

175.1 FEniCS as Computational Science

The goal of the FEniCS project is to develop software for automated computational solution of differential equations based on a finite element methodology combining generality with efficiency. Thermodynamics is a basic area of continuum mechanics with many important applications, which

however is feared by both teachers, students and engineers as being difficult to understand and to apply, principally because of the apperance of turbulence. In this article we show that turbulent thermodynamics can be made understandable and useful by automated computational solution, as a demonstration of the capability of FEniCS.

The biggest mystery of classical thermodynamics is the 2nd Law about entropy and automation cannot harbor any mystery. Expert systems are required for mysteries and FEniCS is not an expert system. Automation requires a continuum mechanics formulation of thermodynamics with a transparent 2nd Law. We present a formulation of thermodynamics based on finite precision computation with a 2nd Law without reference to entropy, which we show can serve as a basis for automated computational simulation of complex turbulent thermodynamics and thus can open to new insight and design, a main goal of FEniCS. In this setting the digital finite element model becomes the real model of the physics of thermodynamics viewed as a form of analog finite precision computation, a model which is open to inspection and analysis because solutions can be computed and put on the table. This represents a new kind of science in the spirit of Dijkstra [6] and Wolfram [29], which can be explored using FEniCS and which we present in non-technical form in My Book of Knols [32].

175.2 The 1st and 2nd Laws of Thermodynamics

> Heat, a quantity which functions to animate, derives from an internal fire located in the left ventricle. (Hippocrates, 460 B.C.)

Thermodynamics is fundamental in a wide range of phenomena from macroscopic to microscopic scales. Thermodynamics essentially concerns the interplay between *heat energy* and *kinetic energy* in a *gas* or *fluid*. Kinetic energy, or *mechanical energy*, may generate heat energy by *compression* or *turbulent dissipation*. Heat energy may generate kinetic energy by *expansion*, but not through a *reverse* process of turbulent dissipation. The industrial society of the 19th century was built on the use of *steam engines*, and the initial motivation to understand thermodynamics came from a need to increase the efficiency of steam engines for conversion of heat energy to useful mechanical energy. Thermodynamics is closely connected to the dynamics of *slightly viscous* and *compressible* gases, since substantial compression and expansion can occur in a gas, but less in fluids (and solids).

The development of classical thermodynamics as a rational science based on logical deduction from a set of axioms, was initiated in the 19th century by Carnot [4], Clausius [3] and Lord Kelvin [20], who formulated the basic axioms in the form of the *1st Law* and the *2nd Law* of thermodynamics. The 1st Law states (for an isolated system) that the *total energy*, the sum of

kinetic and heat energy, is conserved. The 1st Law is naturally generalized to include also conservation of mass and Newton's law of conservation of momentum and then can be expressed as the *Euler equations* for a gas/fluid with *vanishing viscosity*.

The 2nd Law has the form of an inequality $dS \geq 0$ for a quantity named *entropy* denoted by S, with dS denoting change thereof, supposedly expressing a basic feature of real thermodynamic processes. The classical 2nd Law states that the entropy cannot decrease; it may stay constant or it may increase, but it can never decrease (for an isolated system).

The role of the 2nd Law is to give a scientific basis to the many observations of *irreversible* processes, that is, processes which cannot be reversed in time, like running a movie backwards. Time reversal of a process with strictly increasing entropy, would correspond to a process with strictly decreasing entropy, which would violate the 2nd Law and therefore could not occur. A perpetum mobile would represent a reversible process and so the role of the 2nd Law is in particular to explain *why* it is imposssible to construct a perpetum mobile, and *why* time is moving forward in the direction an *arrow of time*, as expressed by Max Planck [10, 27, 28]: *Were it not for the existence of irreversible processes, the entire edifice of the 2nd Law would crumble.*

While the 1st Law in the form of the Euler equations expressing conservation of mass, momentum and total energy can be understood and motivated on rational grounds, the nature of the 2nd Law is mysterious. It does not seem to be a consequence of the 1st Law, since the Euler equations seem to be time reversible, and the role of the 2nd Law is to explain irreversibility. Thus questions are lining up: nIf the 2nd Law is a new independent law of Nature, how can it be justified? What is the physical significance of that quantity named entropy, which Nature can only get more of and never can get rid of, like a steadily accumulating heap of waste? What mechanism prevents Nature from recycling entropy? How can irreversiblity arise in a reversible system? How can viscous dissipation arise in a system with vanishing viscosity? Why is there no *Maxwell demon* [24]? Why can a gas by itself expand into a larger volume, but not by itself contract back again, if the motion of the gas molecules is governed by the reversible Newton's laws of motion? Why is there an arrow of time? This article presents answers.

175.3 The Enigma

> Those who have talked of "chance" are the inheritors of antique superstition and ignorance...whose minds have never been illuminated by a ray of scientific thought. (T. H. Huxley)

These were the questions which confronted scientists in the late 19th century, after the introduction of the concept of entropy by Clausius in 1865,

and these showed to be tough questions to answer. After much struggle, agony and debate, the agreement of the physics community has become to view *statistical mechanics* based on an assumption of *molecular chaos* as developed by Boltzmann [1], to offer a rationalization of the classical 2nd Law in the form of a tendency of (isolated) physical processes to move from improbable towards more probable states, or from ordered to less ordered states. Boltzmann's assumption of molecular chaos in a dilute gas of colliding molecules, is that two molecules about to collide have independent velocities, which led to the *H-theorem* for *Boltzmann's equations* stating that a certain quantity denoted by H could not decrease and thus could serve as an entropy defining an arrow of time. Increasing disorder would thus represent increasing entropy, and the classical 2nd Law would reflect the eternal pessimistists idea that things always get more messy, and that there is really no limit to this, except when everything is as messy as it can ever get. Of course, experience could give (some) support this idea, but the trouble is that it prevents things from ever becoming less messy or more structured, and thus may seem a bit too pessimistic. No doubt, it would seem to contradict the many observations of *emergence* of ordered non-organic structures (like crystals or waves and cyclons) and organic structures (like DNA and human beings), seemingly out of disordered chaos, as evidenced by the physics Nobel Laureate Robert Laughlin [21].

Most trained thermodynamicists would here say that emergence of order out of chaos, in fact does not contradict the classical 2nd Law, because it concerns "non-isolated systems". But they would probably insist that the Universe as a whole (isolated system) would steadily evolve towards a "heat-death" with maximal entropy/disorder (and no life), thus fulfilling the pessimists expectation. The question from where the initial order came from, would however be left open.

The standard presentation of thermodynamics based on the 1st and 2nd Laws, thus involves a mixture of deterministic models (Boltzmann's equations with the H-theorem) based on statistical assumptions (molecular chaos) making the subject admittedly difficult to both learn, teach and apply, despite its strong importance. This is primarily because the question *why* necessarily $dS \geq 0$ and never $dS < 0$, is not given a convincing understandable answer. In fact, statistical mechanics allows $dS < 0$, although it is claimed to be very unlikely. The basic objective of statistical mechanics as the basis of classical thermodynamics, thus is to (i) give the entropy a physical meaning, and (ii) to motivate its tendency to (usually) increase. Before statistical mechanics, the 2nd Law was viewed as an experimental fact, which could not be rationalized theoretically. The classical view on the 2nd Law is thus either as a statistical law of large numbers or as a an experimental fact, both without a rational deterministic mechanistic theoretical foundation. The problem with thermodynamics in this form is that it is understood by very few, if any:

- *Every mathematician knows it is impossible to understand an elementary course in thermodynamics.* (V. Arnold)
- *...no one knows what entropy is, so if you in a debate use this concept, you will always have an advantage.* (von Neumann to Shannon)
- *As anyone who has taken a course in thermodynamics is well aware, the mathematics used in proving Clausius' theorem (the 2nd Law) is of a very special kind, having only the most tenous relation to that known to mathematicians.* (S. Brush [2])
- *Where does irreversibility come from? It does not come form Newton's laws. Obviously there must be some law, some obscure but fundamental equation. perhaps in electricty, maybe in neutrino physics, in which it does matter which way time goes.* (Feynman [9])
- *For three hundred years science has been dominated by a Newtonian paradigm presenting the World either as a sterile mechanical clock or in a state of degeneration and increasing disorder...It has always seemed paradoxical that a theory based on Newtonian mechanics can lead to chaos just because the number of particles is large, and it is subjectivly decided that their precise motion cannot be observed by humans... In the Newtonian world of necessity, there is no arrow of time. Boltzmann found an arrow hidden in Nature's molecular game of roulette.* (Paul Davies [5])
- *The goal of deriving the law of entropy increase from statistical mechanics has so far eluded the deepest thinkers.* (Lieb [22])
- *There are great physicists who have not understood it.* (Einstein about Boltzmann's statistical mechanics)

175.4 Computational Foundation

In this note we present a foundation of thermodynmaics, further elaborated in [12, 5], where the basic assumption of statistical mechanics of molecular chaos, is replaced by *deterministic finite precision computation*, more precisely by a *least squares stabilized finite element method* for the Euler equations, referred to as *Euler General Galerkin* or *EG2*. In the spirit of Dijkstra [6], we thus view EG2 as the physical model of thermodynamics, that is the Euler equations together with a computational solution procedure, and not just the Euler equations without constructive solution procedure as in a classical non-computational approach.

Using EG2 as a model of thermodynamics changes the questions and answers and opens new possibilities of progress together with new challenges to mathematical analysis and computation. The basic new feature is that EG2 solutions are computed and thus are available to inspection. This means that the analysis of solutions shifts from *a priori* to *a posteriori*; after the solution has been computed it can be inspected.

Inspecting computed EG2 solutions we find that they are *turbulent* and have *shocks*, which is identified by pointwise large Euler residuals, reflecting

that pointwise solutions to the Euler equations are lacking. The enigma of thermodynamics is thus the enigma of turbulence (since the basic nature of shocks is understood). Computational thermodynamics thus essentially concerns computational turbulence. In this note and [5] we present evidence that EG2 opens to a resolution of the enigma of turbulence and thus of thermodynamics.

The fundamental question concerns *wellposedness* in the sense of Hadamard, that is what aspects or *outputs* of turbulent/shock solutions are stable under perturbations in the sense that small perturbations have small effects. We show that wellposedness of EG2 solutions can be tested a posteriori by computationally solving a *dual linearized problem*, through which the output sensitivity of non-zero Euler residuals can be estimated. We find that mean-value outputs such as drag and lift and total turbulent dissipation are wellposed, while point-values of turbulent flow are not. We can thus a posteriori in a case by case manner, assess the quality of EG2 solutions as solutions of the Euler equations.

We formulate a *2nd Law* for EG2 without the concept of entropy, in terms of the basic physical quantities of kinetic energy K, heat energy E, rate of *work* W and shock/turbulent dissipation $D > 0$. The new 2nd Law reads

$$\dot{K} = W - D, \quad \dot{E} = -W + D, \tag{175.1}$$

where the dot indicates time differentiation. Slightly viscous flow always develops turbulence/shocks with $D > 0$, and the 2nd Law thus expresses an irreversible transfer of kinetic energy into heat energy, while the total energy $E + K$ remains constant.

With the 2nd Law in the form (175.1), we avoid the (difficult) main task of statistical mechanics of specifying the physical significance of entropy and motivating its tendency to increase by probabilistic considerations based on (tricky) combinatorics. Thus using *Ockham's razor* [25], we rationalize a scientific theory of major importance making it both more understandable and more useful. The new 2nd Law is closer to classical Newtonian mechanics than the 2nd Law of statistical mechanics, and thus can be viewed to be more fundamental.

The new 2nd Law is a consequence of the 1st Law in the form of the Euler equations combined with EG2 finite precision computation effectively introducing viscosity and viscous dissipation. These effects appear as a consequence of the non-existence of pointwise solutions to the Euler equations reflecting instablities leading to the development shocks and turbulence in which large scale kinetic energy is transferred to small scale kinetic energy in the form of heat energy. The viscous dissipation can be interpreted as a penalty on pointwise large Euler residuals arising in shocks/turbulence, with the penalty being directly coupled to the violation following a principle of criminal law exposed in [11]. EG2 thus explains the 2nd Law as a consequence of the non-existence of pointwise solutions with small Eu-

ler residuals. This offers an understanding to the emergence of irreversible solutions of the formally reversible Euler equations. If pointwise solutions had existed, they would have been reversible without dissipation, but they don't exist, and the existing computational solutions have dissipation and thus are irreversible.

175.5 Viscosity Solutions

An EG2 solution can be viewed as particular *viscosity solution* of the Euler equations, which is a solution of *regularized Euler equations* augmented by additive terms modeling viscosity effects with small viscosity coefficients. The effective viscosity in an EG2 solution typically may be comparable to the mesh size.

For incompressible flow the existence of viscosity solutions, with suitable solution dependent viscosity coefficients, can be proved a priori using standard techniques of analytical mathematics. Viscosity solutions are pointwise solutions of the regularized equations. But already the most basic problem with constant viscosity, the incompressible Navier-Stokes equations for a Newtonian fluid, presents technical difficulties, and is one of the open Clay Millennium Problems.

For compressible flow the technical complications are even more severe, and it is not clear which viscosities would be required for an analytical proof of the existence of viscosity solutions [8] to the Euler equations. Furthermore, the question of wellposedness is typically left out, as in the formulation of the Navier-Stokes Millennium Problem, with the motivation that first the existence problem has to be settled. Altogether, analytical mathematics seems to have little to offer a priori concerning the existence and wellposedness of solutions of the compressible Euler equations. In contrast, EG2 computational solutions of the Euler equations seem to offer a wealth of information a posteriori, in particular concerning wellposedness by duality.

An EG2 solution thus can be viewed as a specific viscosity solution with a specific regularization from the least squares stabilization, in particular of the momentum equation, which is necessary because pointwise momentum balance is impossible to achieve in the presence of shocks/turbulence. The EG2 viscosity can be viewed to be the minimal viscosity required to handle the contradiction behind the non-existence of pointwise solutions. For a shock EG2 could then be directly interpreted as a certain physical mechanism preventing a shock wave from turning over, and for turbulence as a form of automatic computational turbulence model.

EG2 thermodynamics can be viewed as form of deterministic chaos, where the mechanism is open to inspection and can be used for prediction. On the other hand, the mechanism of statistical mechanics is not

open to inspection and can only be based on ad hoc assumption, as noted by e.g. Einstein [7]. If Boltzmann's assumption of molecular chaos cannot be justified, and is not needed, why consider it at all, [23]?

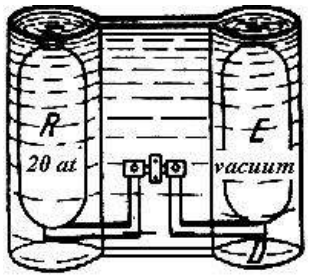


Fig. 358 Concerning overflowing experiment of Joule (Scientific Papers). *R* contains at first air compressed to 20 atm, *E* is initially a vacuum, *D* the tube

FIGURE 175.1. Joule's 1845 experiment

175.6 Joule's 1845 Experiment

To illustrate basic aspects of thermodynamics, we recall Joule's experiment from 1845 with a gas initially at rest, or in equilibrium, at a certain temperature and density in a certain volume immersed into a container of water, see Fig. 175.1. At initial time a valve was opened and the gas was allowed to expand into the double volume while the temperature change in the water was carefully measured by Joule. To the great surprise of both Joule and the scientific community, no change of the temperature of the water could be detected, in contradiction with the expectation that the gas would cool off under expansion. Moreover, the expansion was impossible to reverse; the gas had no inclination to contract back to the original volume.

We simulate Joule's experiment computationally using EG2: At initial time a valve is opened in a channel connecting two cubical chambers, a left and a right chamber, filled with gas of the same temperature but different density/pressure with high density/pressure in the left and low in the right chamber. Fig. 175.2 and 175.3 displays the time-evolution of mean temperature, density, kinetic energy and pressure in the left and right chambers, while Fig. 175.4 and 175.5 give snapshots of the distribution of temperature and speed at an intermediate time.

We see that temperature drop in the left chamber as the gas expands with heat energy transforming to kinetic energy with a maximal tempera-

ture drop in the channel. When the cool expanding gas hits the wall opposite to the channel inlet in the right chamber, it is heated in recompression and returns along the walls into a vortical turbulent flow with additional heating from turbulent dissipation. The net effect is that the mean temperature in the right chamber increases. The mean temperature thus drops in the left chamber and increases in the right and after a slight rebounce settles to a remaining density/temperature gap as the gas comes to rest with the same pressure in the left and right chambers and the same total heat energy as before expansion. Joule measured the total heat energy of the initial and final equilibrium states and found them to be equal. Joule did not seek to measure the dynamics of the process, nor the remaining temperature/density gap.

From the 1st Law alone there are many different possible end states with varying gaps in density/temperature. It is the 2nd Law which determines the size of the gap, which relates to the amount of turbulent/shock dissipation in the left and right chambers, which is determined by the dynamics of the process including the distribution of turbulence/shock dissipation.

Classical thermodynamics focussing on equiblium states does not tell which from a range of possible equlibrium end states with varying gaps, will actually be realized, because the true end state depends on the dynamics of the process. If anything, classical thermodynmics would predict an end state with zero gap, which we have seen is incorrect. In short, classical equilibrium thermodynamics excluding dynamics cannot correctly predict equlibrium end states, and thus has little practical value.

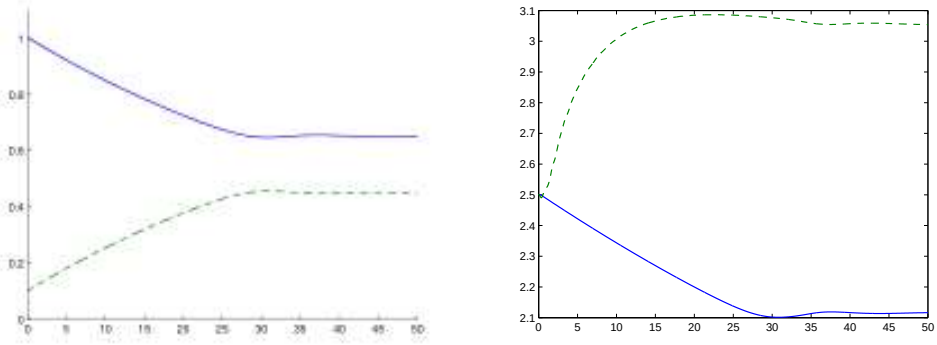


FIGURE 175.2. Density and temperature in left and right chambers

The 2nd Law states that reversal of the process with the gas contracting back to the original small volume, is impossible because the only way the gas can be put into motion without external forcing is by expansion: Self-expansion is possible, but not self-constraction.

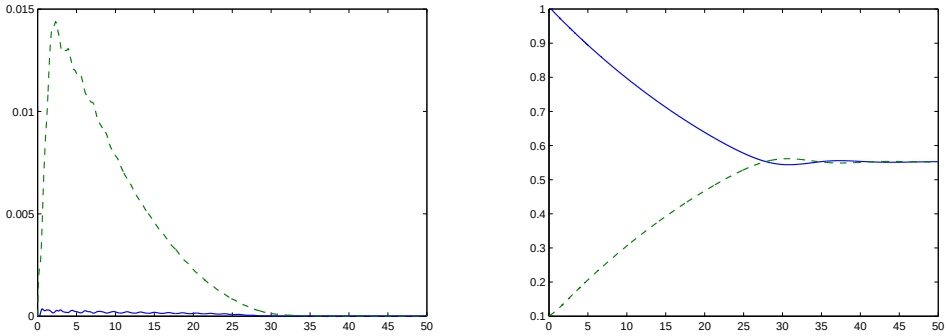


FIGURE 175.3. Kinetic energy and pressure in left and right chambers

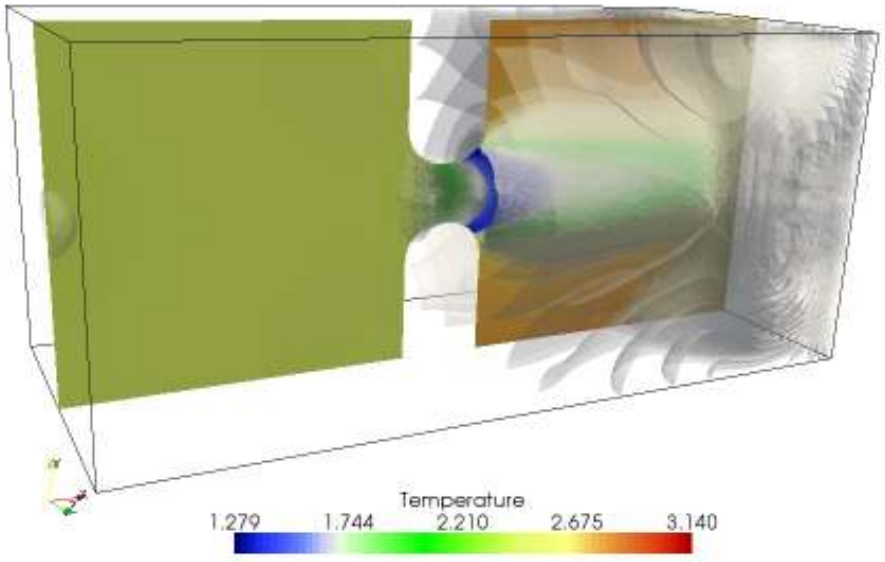


FIGURE 175.4. Distribution of gas temperature at $T = 3$

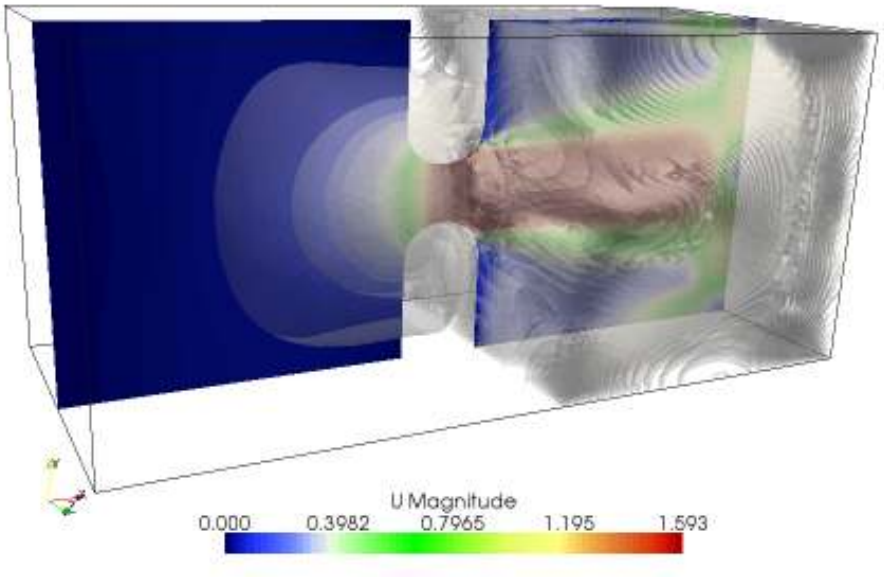


FIGURE 175.5. Distribution of gas speed at $T = 3$

We are thus able to analyze and understand the dynamics of the Joule experiment using the 1st and the new form of the 2nd law. The experiment displays the expansion phase of a compression refrigerator with heat being moved by expansion from the left chamber in contact with the inside of the refrigerator, into the right chamber in contact with the outside. The cycle is closed by recompression under outside cooling. The efficiency connects to the temperature drop in the left chamber and the gap, with efficiency suffering from rebounce to small gap.

175.7 The Euler Equations

We consider the Euler equations for an inviscid perfect gas enclosed in a volume Ω in $\mathbb{R}^3$ with boundary Γ over a time interval $I = (0, 1]$ expressing conservation of *mass density* ρ, *momentum* $m = (m_1, m_2, m_3)$ and *internal energy* e: Find $\hat{u} = (\rho, m, e)$ depending on $(x, t) \in Q \equiv \Omega \times I$ such that

$$\begin{aligned} R_\rho(\hat{u}) \equiv \dot{\rho} + \nabla \cdot (\rho u) &= 0 \quad \text{in } Q, \\ R_m(\hat{u}) \equiv \dot{m} + \nabla \cdot (mu + p) &= f \quad \text{in } Q, \\ R_e(\hat{u}) \equiv \dot{e} + \nabla \cdot (eu) + p\nabla \cdot u &= g \quad \text{in } Q, \\ u \cdot n &= 0 \quad \text{on } \Gamma \times I \\ \hat{u}(\cdot, 0) &= \hat{u}^0 \quad \text{in } \Omega, \end{aligned} \tag{175.2}$$

where $u = \frac{m}{\rho}$ is the velocity, $p = (\gamma - 1)e$ with $\gamma > 1$ a *gas constant*, f is a given volume force, g a heat source/sink and $\hat{u}^0$ a given initial state. We here express energy conservation in terms of the internal energy $e = \rho T$, with T the temperature, and not as conservation of the *total energy* $\epsilon = e + k$ with $k = \frac{\rho v^2}{2}$ the *kinetic energy*, in the form $\dot{\epsilon} + \nabla \cdot (\epsilon u) = 0$. Because of the appearance of shocks/turbulence, the Euler equations lack pointwise solutions, except possible for short time, and regularization is therefore necessary. For a mono-atomic gas $\gamma = 5/3$ and (179.1) then is a *parameter-free model*, the ideal form of mathematical model according to Einstein...

175.8 Energy Estimates for Viscosity Solutions

For the discussion we consider the following regularized version of (179.1) assuming for simplicity that $f = 0$ and $g = 0$: Find $\hat{u}_{\nu,\mu} \equiv \hat{u} = (\rho, m, e)$ such that

$$\begin{aligned} R_\rho(\hat{u}) &= 0 \quad \text{in } Q, \\ R_m(\hat{u}) &= -\nabla \cdot (\nu \nabla u) + \nabla(\mu p \nabla \cdot u) \quad \text{in } Q, \\ R_e(\hat{u}) &= \nu |\nabla u|^2 \quad \text{in } Q, \\ u &= 0 \qquad \text{on } \Gamma \times I, \\ \hat{u}(\cdot, 0) &= \hat{u}^0 \quad \text{in } \Omega, \end{aligned} \tag{175.3}$$

where $\nu > 0$ is a *shear viscocity* $\mu >> \nu \geq 0$ if $\nabla \cdot u > 0$ in expansion (with $\mu = 0$ if $\nabla \cdot u \leq 0$ in compression), is a small *bulk viscosity*, and we use the notation $|\nabla u|^2 = \sum_i |\nabla u_i|^2$. We shall see that the bulk viscosity is a safety feature putting a limit to the work $p \nabla \cdot u$ in expansion appearing in the energy balance.

We note that only the momentum equation is subject to viscous regulariztion. Further, we note that the shear viscosity term in the momentum equation multiplied by the velocity u (and formally integrated by parts) appears as a positive right hand side in the equation for the internal energy, reflecting that the dissipation from shear viscosity is transformed into internal heat energy. In contrast, the dissipation from the bulk viscosity represents another form of internal energy not accounted for as heat energy, acting only as a safety feature in the sense that its contribution to the energy balance in general will be small, while that from the shear viscosity in general will be substantial reflecting shock/turbulent dissipation.

Below we will consider instead regularization by EG2 with the advantage that the EG2 solution is computed and thus is available to inspection, while $\hat{u}_{\nu,\mu}$ is not. We shall see that EG2 regularization can be interpreted as a (mesh-dependent) combination of bulk and shear viscosity and thus (175.3) can be viewed as an analytical model of EG2 open to simple form of analysis in the form of energy estimates.

As indicated, the existence of a pointwise solution $\hat{u} = \hat{u}_{\nu,\mu}$ to the regularized equations (175.3) is an open problem of analytical mathematics, although with suitable additional regularization it could be possible to settle [8]. Fortunately, we can leave this problem aside, since EG2 solutions will be shown to exist a posteriori by computation. We thus formally assume that (175.3) admits a pointwise solution, and derive basic energy estimates which will be paralleled below for EG2. We thus use the regularized problem (175.3) to illustrate basic features of EG2, including the 2nd Law.

We shall prove now that a regularized solution $\hat{u}$ is an approximate solution of the Euler equations in the sense that $R_\rho(\hat{u}) = 0$ and $R_e(\hat{u}) \geq 0$ pointwise, $R_m(\hat{u})$ is weakly small in the sense that

$$\|R_m(\hat{u})\|_{-1} \leq \frac{\sqrt{\nu}}{\sqrt{\mu}} + \sqrt{\mu} << 1, \tag{175.4}$$

where $\|\cdot\|_{-1}$ denotes the $L_2(I; H^{-1}(\Omega))$-norm, and the following 2nd Law holds:

$$\dot{K} \leq W - D, \quad \dot{E} = -W + D, \tag{175.5}$$

where

$$K = \int_\Omega k\,dx, \quad E = \int_\Omega e\,dx, \quad W = \int_\Omega p\nabla\cdot u\,dx, \quad D = \int_\Omega \nu|\nabla u|^2\,dx.$$

Choosing $\nu << \mu$ we can assure that $\|R_m(\hat{u}_{\nu,\mu})\|_{-1}$ is small. We can view the 2nd Law as a compensation for the fact that the momentum equation is only satisfied in a weak sense, and the equation for internal energy with inequality.

The 2nd Law (179.7) states an irreversible transfer of kinetic energy to heat energy in the presence of shocks/turbulence with $D > 0$, which is the generic case. On the other hand, the sign of W is variable and thus the corresponding energy transfer may go in either direction.

The basic technical step is to multiply the momentum equation by u, and use the mass balance equation in the form $\frac{|u|^2}{2}(\dot{\rho} + \nabla\cdot(\rho u)) = 0$, to get

$$\dot{k} + \nabla\cdot(ku) + p\nabla\cdot u - \nabla(\mu p\nabla\cdot u)\cdot u - \nabla\cdot(\nu\nabla u)\cdot u = 0. \tag{175.6}$$

By integration in space it follows that $\dot{K} \leq W - D$, and similarly it follows that $\dot{E} = -W + D$ from the equation for e, which proves the 2nd Law. Adding next (175.6) to the equation for the internal energy e and integrating in space, gives

$$\dot{K} + \dot{E} + \int_\Omega \mu p(\nabla\cdot u)^2\,dx = 0,$$

and thus after integration in time

$$K(1) + E(1) + \int_Q \mu p(\nabla \cdot u)^2 \, dxdt = K(0) + E(0). \tag{175.7}$$

We now need to show that $E(1) \geq 0$ (or more generally that $E(t) > 0$ for $t \in I$), and to this end we rewrite the equation for the internal energy as follows:

$$D_u e + \gamma e \nabla \cdot u = \nu |\nabla u|^2,$$

where $D_u e = \dot{e} + u \cdot \nabla e$ is the material derivative of e following the fluid particles with velocity u. Assuming that $e(x, 0) > 0$ for $x \in \Omega$, it follows that $e(x, 1) > 0$ for $x \in \Omega$, and thus $E(1) > 0$. Assuming $K(0) + E(0) = 1$ the energy estimate (175.7) thus shows that

$$\int_Q \mu p(\nabla \cdot u)^2 \, dxdt \leq 1, \tag{175.8}$$

and also that $E(t) \leq 1$ for $t \in I$. Next, integrating (175.6) in space and time gives, assuming for simplicity that $K(0) = 0$,

$$K(1) + \int_Q \nu(\Delta u)^2 dxdt = \int_Q p \nabla \cdot u dxdt - \int_Q \mu p(\nabla \cdot u)^2 dxdt \leq \frac{1}{\mu} \int_Q p dxdt \leq \frac{1}{\mu},$$

where we used that $\int_Q p dxdt = (\gamma - 1) \int_Q e dxdt \leq \int_I E(t) dt \leq 1$. It follows that

$$\int_Q \nu |\nabla u|^2 dxdt \leq \frac{1}{\mu}. \tag{175.9}$$

By standard estimation (assuming that p is bounded), it follows from (175.8) and (175.9) that

$$\|R_m(\hat{u})\|_{-1} \leq C(\sqrt{\mu} + \frac{\sqrt{\nu}}{\sqrt{\mu}}),$$

with C a constant of moderate size, which completes the proof. As indicated, $\|R_m(\hat{u})\|_{-1}$ is estimated by computation, as shown below. The role of the analysis is thus to rationalize computational experience, not to replace it.

175.9 Compression and Expansion

The 2nd Law (179.7) states that there is a transfer of kinetic energy to heat energy if $W < 0$, that is under compression with $\nabla \cdot u < 0$, and a transfer from heat to kinetic energy if $W > 0$, that is under expansion with $\nabla \cdot u > 0$. Returning to Joule's experiment, we see by the 2nd Law that contraction back to the original volume from the final rest state in the double volume, is impossible, because the only way the gas can be set into motion is by expansion. To see this no reference to entropy is needed.

175.10 A 2nd Law witout Entropy

We note that the 2nd Law (179.7) is expressed in terms of the physical quantities of kinetic energy K, heat energy E, work W, and dissipation D and does not involve any concept of entropy. This relieves us from the task of finding a physical significance of entropy and justification of a classical 2nd Law stating that entropy cannot decrease. We thus circumvent the main difficulty of classical thermodynamics based on statistical mechanics, while we reach the same goal as statistical mechanics of explaining irreversibility in formally reversible Newtonian mechanics.

We thus resolve *Loschmidt's paradox* [23] asking how irreversibility can occur in a formally reversible system, which Boltzmann attempted to solve. But Loschmidt pointed out that Boltzmann's equations are not formally reversible, because of the assumption of molecular chaos that velocities are independent before collision, and thus Boltzmann effectively assumes what is to be proved. Boltzmann and Loschmidt's met in heated debates without conclusion, but after Boltzmann's tragic death followed by the experimental verification of the molecular nature of gases, Loschmidt's paradox evaporated as if it had been resolved, while it had not. Postulating molecular chaos still amounts to assume what is to be proved.

175.11 Comparison with Classical Thermodynamics

Classical thermodynamics is based on the relation

$$Tds = dT + pdv, \tag{175.10}$$

where ds represents change of entropy s per unit mass, dv change of volume and dT denotes the change of temperature T per unit mass, combined with a 2nd Law in the form $ds \geq 0$. On the other hand, the new 2nd Law takes the symbolic form

$$dT + pdv \geq 0, \tag{175.11}$$

effectively expressing that $Tds \geq 0$, which is the same as $ds \geq 0$ since $T > 0$. In symbolic form the new 2nd Law thus expresses the same as the classical 2nd Law, without referring to entropy.

Integrating the classical 2nd Law (175.10) for a perfect gas with $p = (\gamma - 1)\rho T$ and $dv = d(\frac{1}{\rho}) = -\frac{d\rho}{\rho^2}$, we get

$$ds = \frac{dT}{T} + \frac{p}{T}d(\frac{1}{\rho}) = \frac{dT}{T} + (1-\gamma)\frac{d\rho}{\rho},$$

and we conclude that with $e = \rho T$,

$$s = \log(T\rho^{1-\gamma}) = \log(\frac{e}{\rho^\gamma}) = \log(e) - \gamma\log(\rho) \tag{175.12}$$

up to a constant. Thus, the entropy $s = s(\rho, e)$ for a perfect gas is a function of the physical quantities ρ and $e = \rho T$, thus a *state function*, suggesting that s might have a physical significance, because ρ and e have. We thus may decide to introduce a quantity s defined this way, but the basic questions remains: (i) What is the physical significance of s? (ii) Why is $ds \geq 0$? What is the entropy non-perfect gas in which case s may not be a state function?

To further exhibit the connection between the classical and new forms of the 2nd Law, we observe that by the chain rule,

$$\rho D_u s = \frac{\rho}{e} D_u e - \gamma D_u \rho = \frac{1}{T}(D_u e + \gamma \rho T \nabla \cdot u) = \frac{1}{T}(D_u e + e\nabla \cdot u + (\gamma - 1)\rho T \nabla \cdot u)$$

since by mass conservation $D_u \rho = -\rho \nabla \cdot u$. It follows that the entropy $S = \rho s$ satisfies

$$\dot{S} + \nabla \cdot (Su) = \rho D_u s = \frac{1}{T}(\dot{e} + \nabla \cdot (eu) + p\nabla \cdot u) = \frac{1}{T} R_e(\hat{u}). \quad (175.13)$$

A solution $\hat{u}$ of the regularized Euler equations (175.3) thus satisfies

$$\dot{S} + \nabla \cdot (Su) = \frac{\nu}{T} |\nabla u|^2 \geq 0 \quad \text{in } Q, \quad (175.14)$$

where $S = \rho \log(e\rho^{-\gamma})$. In particular, in the case of the Joule experiment with T the same in the initial and final states, we have $s = \gamma \log(V)$ showing an increase of entropy in the final state with larger volume.

We sum up by noting that the classical and new form of the second law effectively express the same inequality $ds \geq 0$ or $Tds \geq 0$. The new 2nd law is expressed in terms of the fundamental concepts of of kinetic energy, heat energy and work without resort to any form of entropy and statistical mechanics with all its complications. Of course, the new 2nd Law readily extends to the case of a general gas.

175.12 EG2

EG2 in cG(1)cG(1)-form for the Euler equations (179.1), reads: Find $\hat{u} = (\rho, m, \epsilon) \in V_h$ such that for all $(\bar{\rho}, \bar{u}, \bar{\epsilon}) \in W_h$

$$\begin{aligned} ((R_\rho(\hat{u}), \bar{\rho})) + ((hu \cdot \nabla \rho, u \cdot \nabla \bar{\rho})) &= 0, \\ ((R_m(\hat{u}), \bar{u})) + ((hu \cdot \nabla m, u \cdot \nabla \bar{u})) + (\nu_{sc} \nabla u, \nabla \bar{u})) &= 0, \\ ((R_\epsilon(\hat{u}), \bar{e})) + ((hu \cdot \nabla \epsilon, u \cdot \nabla \bar{\epsilon})) &= 0, \end{aligned} \quad (175.15)$$

where V_h is a trial space of continuous piecewise linear functions on a space-time mesh of size h satisfying the initial condition $\hat{u}(0) = \hat{u}^0$ with $u \in V_h$ defined by nodal interpolation of $\frac{m}{\rho}$, and W_h is a corresponding test space

of function which are continuous piecewise linear in space and piecewise constant in time, all functions satisfying the boundary condition $u \cdot n = 0$ at the nodes on Γ. Further, $((\cdot,\cdot))$ denotes relevant $L_2(Q)$ scalar products, and $\nu_{sc} = h^2|R_m(\hat{u})|$ is a residual dependent *shock-capturing viscosity*, see [5]. We here use the conservation equation for the total energy ϵ rather than for the internal energy e.

EG2 combines a weak satisfaction of the Euler equations with a weighted least squares control of the residual $R(\hat{u}) \equiv (R_\rho(\hat{u}), R_m(\hat{u}), R_e(\hat{u}))$ and thus represents a midway between the Scylla of weak solution and Carybdis of least squares strong solution.

175.13 The 2nd Law for EG2

Subtracting the mass equation with $\bar{\rho}$ a nodal interpolant of $\frac{|u|^2}{2}$ from the momentum equation with $\bar{u} = u$ and using the heat energy equation with $\bar{e} = 1$, we obtain the following 2nd Law for EG2 (up to a $\sqrt{h}$-correction controled by the shockcapturing viscosity [18]):

$$\dot{K} = W - D_h, \quad \dot{E} = -W + D_h, \tag{175.16}$$

where

$$D_h = ((h\rho u \cdot \nabla u, u \cdot \nabla u)). \tag{175.17}$$

For solutions with turbulence/shocks, $D_h > 0$ expressing an irreversible transfer of kinetic energy into heat energy, just as above for regularized solutions. We note that in EG2 only the momentum equation is subject to viscous regularization, since D_h expresses a penalty on $u \cdot \nabla u$ appearing in the momentum residual.

175.14 The Stabilization in EG2

The stabilization in EG2 is expressed by the dissipative term D_h which can be viewed as a weighted least squares control of the term $\rho u \cdot \nabla u$ in the momentum residual. The rationale is that least squares control of a part of a residual which is large, effectively may give control of the entire residual, and thus EG2 gives a least squares control of the momentum residual. But the EG2 stabilization does not correspond to an ad hoc viscosity, as in classical regularization, but to a form of penalty arsing because Euler residuals of turbulent/shock solutions are not pointwise small. In particular the dissipative mecahnism of EG2 does not correspond to a simple shear viscosity, but rather to a form of "streamline viscosity" preventing fluid particles from colliding while allowing strong shear.

175.15 EG2 Implementation in FEniCS

FEniCS code + short info on a posteroiri error control. To be added by Murtazo.

References

[1] L. Boltzmann, Lectures on Gas Theory, Dover, 1964.

[2] Stephen Brush, *The Kind of Motion We Call Heat: Physics and the Atomists*, Elsevier Science, 1986.

[3] Rudolph Clausius, *Abhandlungen über die Mechanische Wäremtheorie*, Band 1-2, 1864.

[4] Sadi Carnot, in Reflections of the motive power of fire and on machines fitted to develop that power, 1824.

[5] Paul Davies, *The Cosmic Blueprint, New Discoveries in Nature's Creative Ability to Order the Universe*, Templeton Foundation Press.

[6] Dijkstra: *Originallly I viewed it as the function of the abstract machine to provide a truthful picture of the physical reality. Later, however, I learned to consider the abstract machine as the* true *one, because that is the only one we can think; it is the physical machine's purpose to supply a working model, a (hopefully) sufficiently accurate physical simulation of the true, abstract machine.*

[7] Einstein: *Neither Herr Boltzmann nor Herr Planck has given a definition of* W *....Usually* W *is put equal to the number of complexions. In order to calculate* W*, one needs a complete molecular-mechanical theory of the system under consideration. Therefore it is dubious whether the Boltzmann principle has any meaning without a complete molecular-mechanical theory or some other theory which describes the elementary processes (and such a theory is missing).*

[8] Eduard Feireisl, *Dynmaics of Viscous Compressible Fluids*, Oxford University Press, 2004.

[9] Richard Feynman, The Feynman Lectures on Physics, Caltech, 1963.

[10] www.fenics.org

[11] Michel Foucault, Discipline and Punishment, The Birth of the Prison, Vintage books, 1991.

[12] J. Hoffman, A general Galerkin finite element method for turbulent compressible flow, Finite Element Center Preprint 2006-13.

[13] J. Hoffman and C. Johnson, Finally, Resolution of d'Alembert's Paradox, Journal of Mathematical Fluid Mechanics, Online First, Dec 10 2008.

[14] My Book of Knols http://knol.google.com/k/claes-johnson/my-book-of-knols/yvfu3xg7d7wt/57

[15] J. Hoffman and C. Johnson, Computational Turbulent Incompressible Flow, Applied Mathematics Body and Soul Vol 4, Springer, 2007.

[16] J. Hoffman and C. Johnson, Computational Thermodynamics, Applied Mathematics Body and Soul Vol 5, Springer, 2008.

[17] C. Johnson, Adaptive finite element methods for conservation laws, Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, Springer Lecture Notes in Mathematics 1697 (1998), 269-323.

[18] C. Johnson, Adaptive finite element methods for conservation laws, in *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*, Springer Lecture Notes in Mathematics, Springer, 1998.

[19] C. Johnson, The Clock and the Arrow: A Brief Theory of Time, www.csc.kth.se/ cgjoh.

[20] Thomson, W. (1851) On the dynamical theory of heat; with numerical results deduced from Mr. Joule's equivalent of a thermal unit and M. Regnault's observations on steam, Math. and Phys. Papers vol.1, p.179.

[21] Robert Laughlin, A Different Universe, 2005.

[22] Elliot Lieb and Jakob Yngvason, The Physics and Mathematics of the Second Law of Thermodynamics, Physics Reports 310, 1-96 (1999).

[23] J. Loschmidt, Sitzungsber. Kais. Akad. Wiss. Wien, Math. Naturwiss. Classe 73, 128-142, (1876).

[24] Leff, H.S. and Rex, A.F. (eds), Maxwell's Demon: Entropy, Information, Computing. Bristol: Adam-Hilger, 1990.

[25] William Ockham (1285-1349): “Entia non sunt multiplicanda praeter necessitatem” (Entities should not be multiplied unnecessarily).

[26] Max Planck, Über den zweiten Hauptsatz der mechanischen Wärmetheorie, Dissertation, Berlin, 1879.

[27] Max Planck, Vorlesungen über Thermodynamik, 1897.

[28] Max Planck, Acht Vorlesungen über Theoretische Physik, Fünfte Vorlesung: Wärmestrahlung und Elektrodynamische Theorie, Leipzig, 1910.

[29] Stephen Wolfram, A New Kind of Science, http://www.wolframscience.com/

175.16 FEniCS Implementation

176
Computational Blackbody Radiation

> All these fifty years of conscious brooding have brought me no nearer to the answer to the question, "What are light quanta?". Nowadays every Tom, Dick and Harry thinks he knows it, but he is mistaken. (Einstein 1954)

176.1 Watch

- Black Bodies

176.2 Wave-Particle Duality and Modern Physics

Maxwell's equations represent a culmination of classical mathematical physics by offering a compact mathematical formulation of all of electromagnetics including the propagation of light and radiation, as electromagnetic waves. But like in a Greek tragedy, the success of Maxwell's equations prepared for the collapse of classical mathematical physics and the rise of modern physics based on a concept *wave-particle duality* with a resurrection of Newton's old idea of light as a stream of light particles or photons, in its modern version combined with statistics.

But elevating wave-particle duality to a physical principle is a cover-up of a contradiction [3, 4, 11]: As a reasonable human being you may sometimes act like a fool, but duality is here called schizophrenia, and schizophrenic

science is crazy science. In our time this may be represented by string theory, quantum loop gravity, multiverse theory, and possibly also CO_2 climate alarmism ultimately based on radiation as streams of particles. The purpose of this note is to show that particle statistics can be replaced by deterministic finite precision computational wave mechanics. We thus seek to open a door to restoring rational physics including climate physics, without any contradictory wave-particle duality.

176.3 Climate Alarmism, Greenhouse Effect and Backradiation

In particular, the objective is to show that the "greenhouse effect" of climate alarmism claimed to arise from "backradiation" of particle streams as depicted by NASA in Fig. 5, cannot have a real physical meaning, as little as "backconduction" or "backdiffusion". This because such processes are inherently unstable and thus cannot occur in real physics, only in imagination. This removes a main source of energy from CO_2 climate alarmism, in the sense that various feedbacks will have to start from zero rather than an alarming warming from radiation alone. We first give a popular science description in words and then a mathematical one using formulas.

To express physics in precise terms it is necessary to use the language of mathematics, but main ideas can be captured also in ordinary language helping understanding, and so the two forms of expression complement each other. In particular we shall find that the term "backradiation" which can be contemplated without mathematics, when expressed mathematically reveals its true unstable nature, which makes it into a fictitious unphysical phenomenon without reality. We shall find that it represents the same form of fiction as a bubble-economy in real economic terms: Fictitious values without real substance from a circulating selfpropelling flow of paper money.

176.4 Blackbody Radiation in Words

A blackbody acts like a transformer of radiation which absorbs high-frequency radiation and emits low-frequency radiation. The temperature of the blackbody determines a *cut-off frequency* for the emission, which increases linearly with the temperature: The warmer the blackbody is, the higher frequencies it can and will emit. Thus only frequencies below cut-off are emitted, while all frequencies are being absorbed.

A blackbody thus can be seen as a system of resonators with different eigen-frequencies which are excited by incoming radiation and then emit

radiation. An ideal blackbody absorbs all incoming radiation and remits all absorbed radiation below cut-off.

Conservation of energy requires absorbed frequencies above cut-off to be stored in some form, more precisely as heat energy thus increasing the temperature of the blackbody.

As a transformer of radiation a blackbody thus acts in a very simple way: it absorbs all radiation, emits absorbed frequencies below cut-off, and uses absorbed frequencies above cut-off to increase its temperature. A blackbody thus acts as a semi-conductor transmitting only frequencies below cut-off, and grinding coherent frequencies above cut-off into heat in the form of incoherent high-frequency noise.

We here distinguish between coherent organized electromagnetic waves of different frequencies in the form of radiation or light, and incoherent high-frequency vibrations or noise, perceived as heat.

A blackbody thus absorbs and emits frequencies below cut-off without getting warmer, while absorbed frequencies above cut-off are not emitted but are instead stored as heat energy increasing the temperature.

A blackbody is like an amplifier with a restricted range of frequences, or high-pass filter, which remits/amplifies frequencies below a cut-off frequency and dampens frequencies above cut-off with the damped wave energy being turned into heat.

A blackbody acts like a censor which filters out coherent high-frequency (dangerous) information by transforming it into incoherent (harmless) noise. The IPCC acts like a blackbody by filtering coherent critical information transforming it into incoherent nonsense perceived as global warming.

The increase of the cut-off frequency with temperature can be understood as an increasing ability to emit coherent waves with increasing temperature/excitation or wave amplitude. At low temperature waves of small amplitude cannot carry a sharp signal. It is like speaking at $-40\,C$ with very stiff lips.

We can also compare with a common teacher-class situation with an excited/high temperature teacher emitting information over a range of frequencies from low (simple stuff) to high (difficult stuff), which by the class is absorbed and re-emitted/repeated below a certain cut-off frequency, while the class is unable to emit/repeat frequencies above cut-off, which are instead used to increase the temperature or frustration/interest of the class. The temperature of the class can then never exceed the temperature of the teacher, because all coherent information originates from the teacher. The teacher and student connect in two-way communication with a one-way flow of coherent information.

The net result is that a warm blackbody can heat a cold blackbody, but not the other way around. A teacher can teach a student but not the other way around. The hot Sun heats the colder Earth, but the Earth does not heat the Sun. A warm Earth surface can heat a cold atmospheric layer, but a cold atmosphere cannot heat a warm Earth surface. A blackbody is

heated only by frequencies which it cannot emit, but has to store as heat energy.

There is no "backradiation" from the atmosphere to the Earth. There is no "greenhouse effect" from "backradiation". Fig. 5 propagated by NASA thus displays fictional non-physical recirculating radiation with an Earth surface emitting 117% while absorbing 48% from the Sun.

We shall see that the reason recirculation of energy is non-physical is that it is unstable. The instability is of the same nature as that of an economy with income tax approaching 100%, or interest rate 0%, or benefits without limits from taxes without limits. An economy with fictitious money circling with increasing velocity creates financial bubbles which burst sooner or later from inherent instability, as we have been witnessing in recent times.

An atmosphere with circulating radiation would also be unstable and thus cannot exist over time.

There is no "backradiation" by the same reason as there is no "backconduction" or "backdiffusion", namely instability. "Backdiffusion" would correspond to restoring a blurred diffuse image using Photoshop, which you can easily convince yourself is impossible: Take a sharp picture and blurr it, and then try to restore it by sharpening and discover that this does not work, because of instability. Blurring or diffusion destroys fine details which cannot be recovered. Diffusion or blurring is like taking meanvalues of individual values, and the individual values cannot be recovered from mean values. Mixing milk into your coffee by stirring/blurring is possible but unmixing is impossile by unstirring/unblurring.

Radiative heat can be transmitted by electromagnetic waves from a warm blackbody to a colder blackbody, but not from a cold to a warmer, thus with a one-way direction of heat energy, while the electromagnetic waves propagate in both directions. We thus distinguish between two-way propagation of waves and one-way propagation of heat energy by waves.

A cold body can heat up by eating/absorbing high-frequency high temperature coherent waves in a catabolic process of destruction of coherent waves into incoherent heat energy. A warm body cannot heat up by eating/absorbing low-frequency low-temperature waves, because catabolism involves destruction of structure. Anabolism builds structure, but a blackbody is only capable of destructive catabolism (the metabolism of a living cell consists of destructive catabolism and constructive anabolism).

176.5 Planck's Law

The particle nature of light of frequency ν as a stream of *photons* of energy $h\nu$ with h Planck's constant, is supposed to be motivated by Einstein's model of the photoelectric effect [2] viewed to be impossible [1, 7] to explain assuming light is an electromagnetic wave phenomenon satisfying Maxwell's

FIGURE 176.1. A blackbody acts like a censor or high-pass filter which transforms coherent high-frequency high-interest information into incoherent noise, while it lets low-frequency low-interest information pass through.

equations. The idea of light in the form of energy quanta of size $h\nu$ was introduced by Planck [10] in "an act of despair" to explain the *radiation energy* $R_\nu(T)$ emitted by a *blackbody* as a function of frequency ν and temperature T, per unit frequency, surface area, viewing solid angle and time:

$$R_\nu(T) = \gamma T\nu^2\theta(\nu,T), \quad \gamma = \frac{2k}{c^2}, \tag{176.1}$$

with the *high-frequency cut-off* factor

$$\theta(\nu,T) = \frac{\frac{h\nu}{kT}}{e^{\frac{h\nu}{kT}}-1}, \tag{176.2}$$

where c is the speed of light in vacuum, k is Boltzmann's constant, with $\theta(\nu,T) \approx 0$ for $\frac{h\nu}{kT} > 10$ say and $\theta(\nu,T) \approx 1$ for $\frac{h\nu}{kT} < 1$. Since $h/k \approx 10^{-10}$, this effectively means that only frequencies $\nu \leq T10^{11}$ will be emitted, which fits with the common experience that a black surface heated by the high-frequency light from the Sun, will not itself shine like the Sun, but radiate only lower frequencies. We refer to $\frac{kT}{h}$ as the *cut-off* frequency, in the sense that frequencies $\nu > \frac{kT}{h}$ will be radiated subject to strong damping. We see that the cut-off frequency scales with T, which is *Wien's Displacement Law*.

The term *blackbody* is conventionally used to describe an idealized object which absorbs all electromagnetic radiation falling on it, hence appearing to be black. The analysis to follow will reveal some of the real truth of a

real blackbody such as the Earth radiating infrared light while absorbing light mainly in the visible spectrum from the Sun.

It is important to note that the constant $\gamma = \frac{2k}{c^2}$ is very small: With $k \approx 10^{-23}\, J/K$ and $c \approx 3 \times 10^8\, m/s$, we have $\gamma \approx 10^{-40}$. In particular, $\gamma\nu^2 << 1$ if $\nu \leq 10^{18}$ including the ultraviolet spectrum, a condition we will meet below.

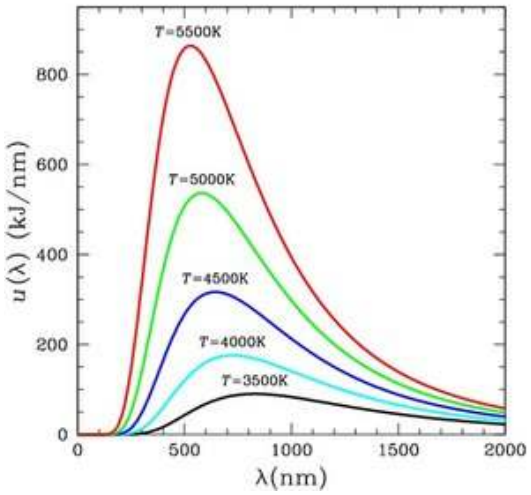


FIGURE 176.2. Radiation energy vs wave length/frequency at different temperatures of a radiating blackbody, per unit frequency. Observe that the cut-off shifts to higher frequency with higher temperature according to Wien's Displacement Law.

By integrating/summing over frequencies in Plancks radiation law (176.1), one obtains *Stefan-Boltmann's Law* stating the the total radiated energy $R(T)$ per unit surface area emitted by a black-body is proportional to T^4:

$$R(T) = \sigma T^4 \tag{176.3}$$

where $\sigma = \frac{2\pi^5 k^4}{15c^2h^3} = 5.67 \times 10^{-8}\, W^{-1}m^{-2}K^{-4}$ is *Stefan-Boltzmann's constant.*

On the other hand, the classical *Rayleigh-Jeans Radiation Law* $R_\nu(T) \sim \gamma T\nu^2$ without the cut-off factor, results in an "ultra-violet catastrophy" with infinite total radiated energy, since $\gamma T \int_1^n \nu^2\, d\nu \sim \gamma T n^3 \to \infty$ as $n \to \infty$.

Stefan-Boltzmann's Law fits (reasonably well) to observation, while the Rayleigh-Jeans Law leads to an absurdity and so must somehow be incorrect. The Rayleigh-Jeans Law was derived viewing light as electromagnetic waves governed by Maxwell's equations, which forced Planck in his "act of despair" to give up the wave model and replace it by statistics of "quanta"

viewing light as a stream of particles or photons. But the scientific cost of abandoning the wave model is very high, and we now present an alternative way of avoiding the catastrophy by modifying the wave model by *finite precision computation*, instead of resorting to particle statistics.

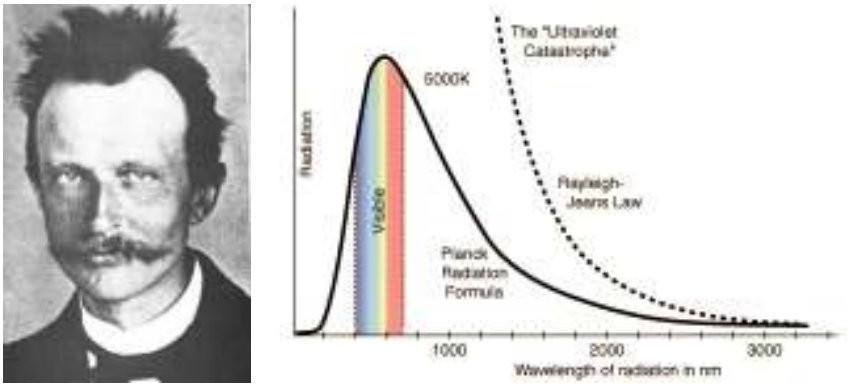


FIGURE 176.3. Planck on the ultraviolet catastrophy in 1900: *...the whole procedure was an* **act of despair** *because a theoretical interpretation had to be found at any price, no matter how high that might be...Either the quantum of action was a fictional quantity, then the whole deduction of the radiation law was essentially an illusion representing only an empty play on formulas of no significance, or the derivation of the radiation law was based on sound physical conception.* Planck in 1909: *Mechanically, the task seems impossible, and we will just have to get used to it (quanta).*

We shall see that finite precision computation introduces a high-frequency cut-off in the spirit of the finite precision computational model for thermodynamics presented in [5].

The scientific price of resorting to statistical mechanics is high, as was clearly recognized by Planck and Einstein, because the basic assumption of statistical mechanics of microscopic games of roulette seem both scientifically illogical and impossible to verify experimentally. Thus statistical mechanics runs the risk of representing *pseudo-science* because of obvious difficulties of testability of basic assumptions.

The purpose of this note is to present an alternative to particle statistics for black-body radiation based on determinsitic finite precision computation in the form of *General Galerkin G2* [4, 5].

To observe individual photons as "particles" without both mass and charge seems impossible, and so the physical reality of photons has remained hypothetical with the main purpose of explaining black-body radiation and the photoelectric effect. If explanations can be given by wave mechanics, both the contradiction of wave-particle duality and the mist of statistical mechanics can be avoided, thus fulfilling a dream of the late Einstein [3, 4].

176.6 The Enigma

The basic enigma of blackbody radiation can be given different formulations:

- Why is a blackbody black/invisible, by emitting infrared radiation when "illuminated" by light in the visible spectrum?
- Why is radiative heat transfer between two bodies always directed from the warmer body to the colder?
- Why can high frequency radiation transform to heat energy?
- Why can heat energy transform to radiation of a certain frequency only if the temperature is high enough?

We shall find that the answer is *resonance in a system of oscillators* (oscillating molecules/charges):

- incoming radiation is absorbed by resonance,
- absorbed incoming radiation is emitted as outgoing radiation, or is stored as internal/heat energy,
- outgoing radiation has a frequency spectrum $\sim T\nu^2$ for $\nu \lesssim T$, assuming all frequencies ν have the same temperature T, with a cut-off to zero for $\nu \gtrsim T$,
- incoming frequencies below cut-off are emitted,
- incoming frequencies above cut-off are stored as internal heat energy.

176.7 Waves vs Particles in Climate Science

We shall find answers to these questions using a wave model where we can separate between propagation of waves and propagation of heat energy by waves, which allows two-way propagation of waves with one-way propagation of heat energy. In a particle model this separation is impossible since the heat energy is tied to the particles. Radiation as a stream of particles thus leads to an idea of "backradiation" with two-way propagation of heat energy carried by two-way propagation of particles. We argue that such two-way propagation is unstable because it requires cancellation, and cancellation in massive two-way flow of heat energy is unstable to small perturbations and thus is unphysical. We thus find that the supposed scientific basis of climate alarmism is unstable and therefore will collapse under perturbations, even small ones, with climategate representing a perturbation which is big rather than small...

176.8 A Wave Equation with Radiation

> There are no quantum jumps, nor are there any particles. (H.D. Zeh [12])

176.8.1 A Basic Radiation Model

We consider the wave equation with radiation, for simplicity in one space dimension assuming periodicity: Find $u = u(x,t)$ such that

$$\ddot{u} - u'' - \gamma \dddot{u} = f, \quad -\infty < x,\, t < \infty \tag{176.4}$$

where (x,t) are space-time coordinates, $\dot{v} = \frac{\partial v}{\partial t}$, $v' = \frac{\partial v}{\partial x}$, $f(x,t)$ models forcing in the form of incoming waves, and the term $-\gamma\dddot{u}$ models outgoing radiation with $\gamma > 0$ a small constant.

This models, in the spirit of Planck [10] before collapsing to statistics of quanta, a continuous string of vibrating charges absorbing energy from the forcing f of intensity f^2 and radiating energy of intensity $\gamma\ddot{u}^2$. The radiation term has the form $-\gamma\dddot{u} \sim \dot{F}$, where $F \sim \ddot{u}$ represents the electrical field generated by an oscillating charge at position x with accelleration $\ddot{u}(x,t)$.

176.8.2 Basic Energy Balance

Multiplying (244.12) by $\dot{u}$ and integrating by parts over a space period, we obtain

$$\int (\ddot{u}\dot{u} + \dot{u}'u')\,dx + \int \gamma\ddot{u}^2\,dx = \int f\dot{u}\,dx,$$

which we can write

$$\dot{E} = A - R \tag{176.5}$$

where

$$E(t) \equiv \frac{1}{2}\int (\dot{u}(x,t)^2 + u'(x,t)^2)\,dx \tag{176.6}$$

is the internal energy viewed as heat energy, and

$$A(t) = \int f(x,t)\dot{u}(x,t)\,dx, \quad R(t) = \int \gamma\ddot{u}(x,t)^2 dx, \tag{176.7}$$

is the absorbed and radiated energy, respectively, with their difference $A-R$ driving changes of internal energy E.

If the incoming wave is an emitted wave $f = -\gamma\dddot{U}$ of amplitude U, then

$$\dot{E} = \int (f\dot{u} - \gamma\ddot{u}^2)dx = \int \gamma(\ddot{U}\ddot{u} - \ddot{u}^2)\,dx \le \frac{1}{2}(R_{in} - R), \tag{176.8}$$

with $R_{in} = \int \gamma\ddot{U}^2\,dx$ the incoming radiation energy, and R the outgoing. We conclude that if $\dot{E} \ge 0$, then $R \le R_{in}$, that is, in order for energy to be

stored as internal/heat energy, it is required that the incoming radiation energy is bigger than the outgoing.

Of course, this is what is expected from conservation of energy. It can also be viewed as a 2nd Law of Radiation stating that radiative heat transfer is possible only from warmer to cooler. We shall see this basic law expressed differently more precisely below.

176.9 Computational Rayleigh-Jeans Law

> But the conception of localized light-quanta out of which Einstein got his equation must still be regarded as far from established. Whether the mechanism of interaction between ether waves and electrons has its seat in the unknown conditions and laws existing within the atom, or is to be looked for primarily in the essentially corpuscular Thomson-Planck-Einstein conception of radiant energy, is the all-absorbing uncertainty upon the frontiers of modern Physics. (Robert A. Millikan [8])

176.9.1 Spectral Analysis of Radiation

We shall show that the Rayleigh-Jeans radiation law $R_\nu(T) = \gamma T\nu^2$ is a direct consequence of the form of the radiation term $-\gamma\dddot{u}$, assuming that all frequencies have the same temperature T. This is elementary.

We shall also show that if the intensity of the forcing f in the model (244.12) has a Rayleigh-Jeans spectrum $\sim T\nu^2$, then so has the corresponding radiation energy $R_\nu(T)$. More precisely, we show as a main result that

$$R_\nu(T) \sim \overline{f_\nu^2} \tag{176.9}$$

with the bar denoting integration in time. This is less elementary and results from a (quite subtle) phenomenon of near resonance.

The prove this we first make a spectral decomposition in x, assuming periodicity with period 2π:

$$\ddot{u}_\nu + \nu^2 u_\nu - \gamma\dddot{u}_\nu = f_\nu, \quad -\infty < t < \infty, \quad \nu = 0, \pm 1, \pm 2, ..., \tag{176.10}$$

into a set of damped linear oscillators with

$$u(x,t) = \sum_{\nu=-\infty}^{\infty} u_\nu(t)e^{i\nu x}.$$

We then use Fourier transformation in t,

$$u_\nu(t) = \int_{-\infty}^{\infty} u_{\nu,\omega}e^{i\omega t}d\omega, \quad u_{\nu,\omega} = \frac{1}{2\pi}\int_{-\infty}^{\infty} u_\nu(t)e^{-i\omega t}\,dt,$$

to get, assuming $u_\nu^{(3)}$ can be replaced by $-\nu^2\dot{u}_\nu$:

$$(-\omega^2+\nu^2)u_{\nu,\omega}+i\omega\gamma\nu^2u_{\nu,\omega}=f_{\nu,\omega}.$$

We have by Parseval's formula,

$$\overline{u_\nu^2}\equiv\int_{-\infty}^{\infty}|u_\nu(t)|^2\,dt=2\pi\int_{-\infty}^{\infty}|u_{\nu,\omega}|^2\,d\omega=2\pi\int_\omega\frac{|f_{\nu,\omega}|^2\,d\omega}{(\nu-\omega)^2(\nu+\omega)^2+\gamma^2\nu^4\omega^2}$$

$$\sim\frac{1}{\nu^2}\int_{-\infty}^{\infty}\frac{|f_{\nu,\omega}|^2\,d\omega}{(\nu-\omega)^2+\gamma^2\nu^4}\sim\frac{1}{\gamma\nu^4}\int_{-\infty}^{\infty}\frac{|f_{\nu,\nu+\gamma\nu^2\bar{\omega}}|^2\,d\bar{\omega}}{\bar{\omega}^2+1},$$

where we used the change of integration variable $\omega=\nu+\gamma\nu^2\bar{\omega}$, and we hide constants using $\sim$ to denote proportionality (with constant close to 1).

We now assume that $|f_{\nu,\nu+\gamma\nu^2\bar{\omega}}|^2\sim\overline{f_\nu^2}$ for $|\bar{\omega}|\le 1$, which means that frequencies ω with $|\nu-\omega|\lesssim\gamma\nu^2$ contribute more or less equally to the excitation of the frequency ν, because the resonance term $(\nu-\omega)^2$ then is dominated by the radiation term $\gamma^2\nu^4$. This reflects that the radiation term acts like diffusion effectively blurring the ω-reading of the forcing $f_{\nu,\omega}$. With this assumption we get

$$\overline{u_\nu^2}\sim\frac{1}{\gamma\nu^4}\overline{f_\nu^2}$$

that is

$$R_\nu\equiv\gamma\overline{\ddot{u}_\nu^2}\approx\gamma\nu^4\overline{u_\nu^2}=\gamma T_\nu\nu^2\sim\overline{f_\nu^2},\tag{176.11}$$

where $R_\nu=R_\nu(T_\nu)$ is the intensity of the radiated wave of frequency ν, and we view $T_\nu=\frac{1}{2}(\overline{\dot{u}_\nu^2}+\nu^2\overline{u_\nu^2})\approx\overline{\dot{u}_\nu^2}$ as the temperature of the corresponding frequency.

We read from (176.11) that

$$R_\nu(T_\nu)\approx\gamma T_\nu\nu^2,\tag{176.12}$$

which is the Rayleigh-Jeans Law. Further, if $\overline{f_\nu^2}\sim T\nu^2$, then also $R_\nu(T_\nu)\sim T\nu^2$ with $T_\nu\sim T$. The emitted radiation will thus mimic an incoming Rayleigh-Jeans spectrum, in temperature equilibrium with $T_\nu\sim T$ for all frequencies ν.

We note that the constant of proportionality in $R_\nu\sim\overline{f_\nu^2}$ is independent of γ and ν which reflects that the string has a certain absorbitivity (greater or equal to its emissivity).

Summing over frequencies we get

$$R\equiv\frac{1}{2\pi}\int_0^{2\pi}\gamma\overline{\ddot{u}^2}\,dx\sim\frac{1}{2\pi}\int_0^{2\pi}\overline{f^2}\,dx=\|f\|^2,\tag{176.13}$$

that is, the intensity of the total outgoing radiation R is proportional to the intensity of the incoming radiation as measured by $\|f\|^2$, thus $R\sim\|f\|^2$. We summarize in

Theorem 176.1 *The radiation* $R_\nu = \overline{\gamma\ddot{u}_\nu^2}$ *of the damped oscillator (176.10) with forcing* f_ν *satisfies* $R_\nu \sim \overline{f_\nu^2}$*, or after summation* $R \sim \|f\|^2$*. In particular, if* $\overline{f_\nu^2} \sim \gamma T\nu^2$ *then* $R_\nu = R_\nu(T_\nu) \sim \gamma T\nu^2$ *with* $T_\nu = T$*.*

176.9.2 Radiation from Near-Resonance

We have seen radiation resulting from forcing by a phenomeon of near-resonance in a damped oscillator of the form

$$\ddot{u}_\nu + \nu^2 u_\nu + \gamma\nu^2\dot{u}_\nu = f_\nu, \tag{176.14}$$

where the forcing f_ν is balanced by the dynamics of the oscillator $\ddot{u}_\nu+\nu^2u_\nu$ and the radiator $\gamma\nu^2\dot{u}_\nu$ with an effect of dissipative damping (with $\gamma\nu^2 \leq 1$). In the case of large damping with $\gamma\nu^2 \approx 1$, then f_ν is mainly balanced by the radiator, that is, $\gamma\nu^2\dot{u}_\nu \approx \dot{u}_\nu \approx f_\nu$ with the result that $R_\nu = \overline{f_\nu\dot{u}_\nu} \approx \overline{f_\nu^2}$. We see that in this case $\dot{u}_\nu$ is *in-phase* with the forcing f_ν, and there is little resonance with the oscillator.

We next consider the case $\gamma\nu^2 << 1$ with small damping and thus near-resonance. The relation $R_\nu = \overline{f_\nu\dot{u}_\nu} \sim \overline{f_\nu^2}$ tells us that in this case f_ν is balanced by the dynamics of both oscillator and radiator with u_ν in-phase and thus $\dot{u}_\nu$ *out-of-phase*. This is because if not, then $\gamma\nu^2\dot{u}_\nu \approx f_\nu$ with $\dot{u}_\nu$ in-phase, which would give the contradicting $R_\nu = \overline{f_\nu\dot{u}_\nu} \sim \frac{\overline{f_\nu^2}}{\gamma\nu^2} >> \overline{f_\nu^2}$.

176.9.3 Absorption vs Emission

In the wave model (244.12) we have associated the term $-\gamma\dddot{u}$ with radiation, but if we just read the equation, we only see a dissipative term absorbing energy without information how this energy is dispensed with e.g. by being radiated away. The model thus describes *absorption by* the vibration string under forcing, and as written the process of *emission from* the string.

However, if we switch the roles of f and $-\gamma\dddot{u}$ and view $-\gamma\dddot{u}$ as input, then we can view f as an emitted wave, which can act as forcing on another system. For frequencies with $\gamma\nu^2 << 1$, we will then have

$$\overline{f_\nu^2} \sim \gamma\overline{\ddot{u}_\nu^2} >> \overline{(\gamma\dddot{u})^2} \approx \gamma\nu^2\gamma\overline{\ddot{u}_\nu^2}$$

with thus emission boosted by resonance, as in the resonant amplification of a musical instrument (e.g the body of a guitar).

In both cases, the relation $R_\nu \sim \overline{f^2}$ expresses that the energy of the incoming absorbed radiation is equal to the outgoing emitted radiation.

176.10 Computational Planck Law

> Would it not be possible to replace the hypothesis of light quanta by another assumption that would also fit the known phenomena?

> If it is necessary to modify the elements of the theory, would it not be possible to retain at least the equations for the propagation of radiation and conceive only the elementary processes of emission and absorption differently than they have been until now? (Einstein)

176.10.1 The Gordian Cut-Off by Planck

The Rayleigh-Jeans Law leads to an "ultraviolet catastrophe" because without some form of high-frequency limitation, the total raditation will be unbounded. Classical wave mechanics thus appears to lead to an absurdity, which has to be resolved in one way or the other. In an "act of despair" Planck escaped the catastrophy by cutting the Gordian Knot simply replacing classical wave mechanics with a new statistical mechanics where high frequencies were assumed to be rare; "a theoretical interpretation had to be found at any price, no matter how high that might be...". It is like kicking out a good old horse which has served fine for many purposes, just because it has a tendency to "go to infinity" at a certain stimulus, and replacing it with a completely new wild horse which you don't understand and cannot control.

The price of throwing out classical wave mechanics is very high, and it is thus natural to ask if this is really necessary.Is there a form of classical mechanics without the ultraviolet catastrophe? Can a cut-off of high frequencies be performed without an Gordian Cut-off?

We believe this is possible, and it is certainly highly desirable, because statistical mechanics is difficult to both understand and apply. We shall thus present a resolution where Planck's statistical mechanics is replaced by deterministic mechanics viewing physics as a form of *analog computation with finite precision* with a certain dissipative diffusive effect, which we model by digital computational mechanics associated with a certain numerical dissipation.

It is natural to model finite precision computation as a dissipative/diffusive effect, since finite precision means that small details are lost as in smoothing by damping of high frequencies which is the effect of dissipation by diffusion.

We consider computational mechanics in the form of the *General Galerkin (G2)* method for the wave equation, where the dissipative mechanism arises from a weighted least squares residual stabilization [4]. We shall first consider a simplified form of G2 with least squares stabilization of one of the residual terms and corresponding simplified diffusion model. We then comment on full G2 residual stabilization.

176.10.2 Wave Equation with Radiation and Dissipation

We consider the wave equation (244.12) with radiation augmented by (simplified) G2 diffusion:

$$\begin{aligned} &\ddot{u} - u'' - \gamma\dddot{u} - \delta^2\dot{u}'' = f, \quad -\infty < x, t < \infty, \\ &\dot{E} = \int f\dot{u}\,dx - \int \gamma\ddot{u}^2\,dx, \quad -\infty < t < \infty, \end{aligned} \tag{176.15}$$

where $-\delta^2\dot{u}''$ models dissipation/diffusion from velocity gradients, $\delta = h/T$ represents a *smallest coordination length* with h a *precision* or *smallest detectable change*, and T is temperature related to the internal energy E by $T = \sqrt{E}$.

The relation $\delta = \frac{h}{T}$ takes the form $|\dot{u}|\delta \sim h$ with $T \sim |\dot{u}|$. A signal with $|\dot{u}|\delta < h$ cannot be represented in coherent form and thus cannot be emitted. This is like the "Mexican Wave" around a stadium which cannot be sustained unless people raise their arms properly; the smaller the "lift" is (with lift as temperature), the longer is the required coordination length or wave length.

We see that the wave equation is here augumented by an equation for the internal energy E, which thus has a contribution from the dissipation $\int \delta^2(\dot{u}')^2\,dx$ (obtained as above by multiplication by $\dot{u}$). In particular we have as above if the incoming wave is an emitted wave $f = -\gamma\dddot{U}$ of amplitude U, then

$$\dot{E} = \int \gamma(\ddot{U}\ddot{u} - \ddot{u}^2)\,dx \le \frac{1}{2}\int \gamma(\ddot{U}^2 - \ddot{u}^2)\,dx. \tag{176.16}$$

We assume that incoming frequencies are bounded by a certain maximal frequency ν_{max}, we choose $\gamma = \nu_{max}^{-2}$ and assume $\nu_{max}^{-1} >> \delta^2 = \nu_{cut}^{-2} >> \gamma$, where $\nu_{cut} < \nu_{max}$ is a certain cut-off frequency.

We motivate this set up as follows: If u is a wave of frequency ν in x, then for $\nu > \nu_{cut} = \frac{T}{h} = \frac{1}{\delta}$, we have

$$\delta^2\dot{u}'' \sim \frac{h^2\nu^2}{T^2}\dot{u}$$

which signifies the presence of considerable damping in (176.15) from the dissipative term since $\frac{h^2\nu^2}{T^2} \ge 1$. Alternatively, we have by a spectral decomposition as above

$$\delta^2\nu^2\dot{u}_\nu^2 \sim f_\nu^2$$

and thus since $\gamma << \delta^2$

$$R_\nu = \frac{\gamma}{\delta^2}\delta^2\nu^2\dot{u}_\nu^2 << f_\nu^2.$$

Thus absorbed waves with $\nu > \nu_{cut}$ are damped and not fully radiated with the corresponding missing energy contributing to the internal/heat energy E and increasing temperature T.

We will also find cut-off for lower frequencies due to the design of the dissipative term $\delta^2\dot{u}''$ corresponding to a simplified form of G2 discretization. In real G2 computations the cut-off will have little effect on frequencies smaller than ν_{cut}. In the analysis we assume this to be the case, which corresponds to allowing δ to depend on ν so that effectively $\delta = 0$ for $\nu \leq \nu_{cut} = \frac{1}{\delta}$. We then obtain a Planck Law of the form

$$R_\nu(T) = \gamma T\nu^2\theta_h(\nu, T) = \gamma T \min(\nu^2, \nu_{cut}^2) \tag{176.17}$$

with a computational high-frequency cut-off factor $\theta_h(\nu, T) = 1$ for $\nu \leq \nu_{cut}$ and $\theta_h(\nu, T) = \frac{\nu_{cut}^2}{\nu^2}$ for $\nu_{cut} < \nu < \nu_{max}$ with $\nu_{cut} = \frac{T}{h}$.

Clearly, it is possible to postulate different cut-off functions $\theta_h(\nu, T)$ for example exponential cut-off functions with the effect that $\theta_h(\nu, T) \approx 0$ for $\nu >> \nu_{cut}$. In the next section we study the cut-off in G2.

The net result is that absorbed frequencies above cut-off will heat the string, while absorbed frequencies below cut-off will be radiated without heating (in the ideal case with the dissipation only acting above cut-off).

If the incoming radiation has a Rayleigh-Jeans spectrum $\sim T\nu^2$, then so has the outgoing radiated spectrum $R_\nu(T_\nu) \sim T\nu^2$ with $T\nu \sim T$ for $\nu \leq \nu_{cut}$. In particular, the outgoing radiated spectrum is equilibrated with all colors having the same temperature, if the incoming spectrum is equilibrated.

Another way of expressing this fundamental property of the vibrating string model is to say that frequencies below cut-off will be absorbed and radiated as *coherent* waves, while frequencies above cut-off will be absorbed transformed into internal energy in the form of incoherent waves. which are not radiated. High frequencies thus may heat the body and thereby decrease the coordination length and thereby allow absorption and emission of higher frequencies.

Note that the internal energy E is the sum over the internal energies E_ν of frequencies $\nu \leq \nu_{cut} \sim T$ with $E_\nu \sim T$ assuming equilibration in temperature, and thus $E \sim T^2$ motivating the relation $T = \sqrt{E}$.

176.10.3 Cut-Off by Residual Stabilization

The discertization in G2 is accomplished by residual stabilization of a Galerkin variational method and may take the form: Find $u \in V_h$ such that for all $v \in V_h$

$$\int (A(u) - f)v\,dxdt + \delta^2 \int (A(u) - f)A(V)\,dxdt = 0, \tag{176.18}$$

where $A(u) = \ddot{u} - u'' - \gamma\dddot{u}$ and V is a primitive function to v (with $\dot{V} = v$), and V_h is a space-time finite element space continuous in space and discontinuous in time over a sequence of discrete time levels.

Here $A(u) - f$ is the residual and the residual stabilization requires $\delta^2(A(u) - f)^2$ to be bounded, which should be compared with the dissipation $\delta\ddot{u}^2$ in the analysis with $\ddot{u}^2$ being one of the terms in the expression $(A(u) - f)^2$. Full residual stabilization has little effect below cut-off, acts like simplified stabilization above cut-off, and effectively introduces cut-off to zero for $\nu \geq \nu_{max}$ since then $\gamma|\dddot{u}| \sim \gamma\nu^2|\dot{u}| = \frac{\nu^2}{\nu_{max}^2}|\dot{u}| \geq |\dot{u}|$, which signifies massive dissipation.

176.10.4 The Sun and the Earth

If an incoming spectrum of temperature T_{in} is attenuated by a factor $\kappa << 1$ (representing a solid viewing angle $<< 180°$), so that the incoming radiation $f_\nu^2 = \kappa\gamma T_{in}\nu^2$ with cut-off for $\nu > \frac{T_{in}}{h}$ (and not for $\nu > \frac{\kappa T_{in}}{h} << \frac{T_{in}}{h}$).

This may represent the incoming radiation from the Sun to the Earth with $\kappa \approx (\frac{R}{D})^2 \approx 0.005^2$ the viewing angle of the Sun seen from the Earth, R the radius of the Sun and D the distance from the Sun to the Earth. The amplitude of the incoming radiation is thus reduced by the factor κ, while the cut-off of the spectrum is still $\frac{\hat{T}}{h}$.

The Earth at temperature T acting like the vibrating string will convert absorbed radiation into heat for frequencies $\nu > \frac{T}{h}$, that is as long a $T < \hat{T}$, while radiating $\sim \gamma T^4$ while absorbing $\sim \kappa T_{in}^4$ thus reaching equilibrium with $\frac{T^4}{T_{in}} \approx \kappa$. With $T_{in} = 5778$ K and $\kappa = 0.005^2$, this gives $T \approx 273K$ (including a factor 4 from the fact that the the disc area of the Sun is πR^2 and the Earth surface area $4\pi r^2$ with r the Earth diameter).

The amplitude of the radiation/light emitted from the surface of the Sun at 5778 K when viewed from the Earth is scaled by the viewing solid angle (scaling with the square of distance from the Sun to the Earth), while the light spectrum covering the visible spectrum centered at $0.5\,\mu m$ remains the same. The Earth emits infrared radiation (outside the visible spectrum) at an effective blackbody temperature of 255 K (at a height of 5 km), thus with almost no overlap with the incoming Sunlight spectrum. The Earth thus absorbs high-frequency reduced-amplitude radiation and emits low-frequency radiation, and thereby acts as a transformer of radiation from high to low frequency: Coherent high-frequency radiation is absorbed and dissipated into incoherent heat energy, which is then emitted as coherent low-frequency radiation.

The transformation only acts from high-frequency to low-frequency, and is an irreversible process representing a 2nd Law.

176.10.5 The Temperature of Radiation

The temperature T_{in} of incoming radiation with an attenuated Planck spectrum $R_\nu = \kappa\gamma T_{in}\nu^2$ with cut-off for $\nu > \frac{T_{in}}{h}$, can be read from the cut-off

(Wien's Law), while the amplitude does not carry this information unless the attenuation factor κ is known. For the outgoing spectrum $\gamma T\nu^2$, we noted that $T \leq T_{in}$ since heating requires dissipative cut-off after absorption, which requires that incoming radiation contains higher frequencies than outgoing and that is only possible if the temperature of the incoming radiation is bigger than the present temperature of the absorbing body, as also expressed in the basic energy balance (176.5): Energy is transferred only from warmer to cooler.

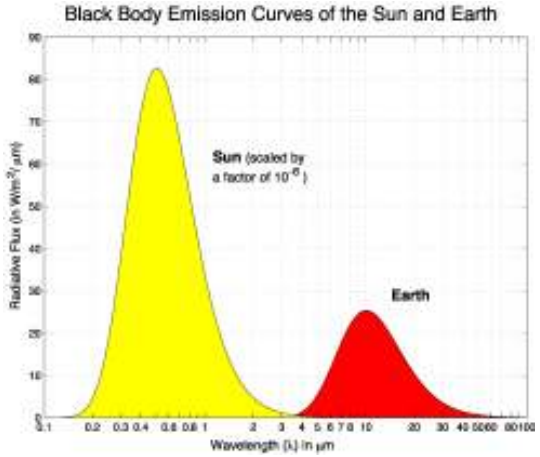


FIGURE 176.4. Blackbody spectrum of the Sun and the Earth.

176.10.6 A Fourier Law of Radiative Heat Transfer

Suppose an incoming radiation has a spectrum $\kappa\gamma T_{in}\nu$ of temperature T_{in} (with $\kappa \leq 1$) is absorbed and then emitted with spectrum $\gamma T\nu^2$. The heating effect from frequencies above cut-off at T, assuming $h = 1$, is then given by

$$\int_T^{T_{in}} \kappa\gamma T_{in}\nu^2\, d\nu \sim \kappa\gamma T_{in}(T_{in}^3 - T^3) \sim \kappa\gamma T_{in}^3(T_{in} - T) \qquad (176.19)$$

which can be viewed as a Fourier Law with heating proportional to temperature difference $T_{in} - T \geq 0$. Note that if $T_{in} < T$, then there is no heating since there is no cut-off: all of absorbed radiation is emitted.

176.11 The 2nd Law and Irreversibility

Radiative heating of a blackbody is an irreversible process, because the heating results from dissipation with coherent high frequency energy above

cut-off being transformed into internal heat energy. We have shown that radiative heating requires that the temperature of the incoming radiation is higher than that of the absorbing body.

We assume that the dissipation is only active above cut-off, while the radition is active over the whole spectrum. Below cut-off radiation is a reversible process since the same spectrum is emitted as absorbed. Formally, the radiation term is dissipative and thus would be expected to transform the spectrum, and the fact that it does not is a remarkable effect of the resonance.

176.12 Aspects of Radiative Heat Transfer

We can find aspects of radiative heating in many different settings, as heat conduction or communicating vessels with the flow always from higher level (temperature) to lower level. But radiative heat transfer is richer in the sense that it involves propagation of both waves and energy.

Let us try with a parallel in psychology: We know that trivial messages radiated from a parent may enter one ear of a child and go out through the other, while less trivial messages would not be listened to at all. However, the alertness of the child may be raised as a result of a "high temperature" outburst by the parent which could open the childs mind to absorbing/radiating less trivial messages. We would here distinguish between propagation of message and meaning.

176.13 Reflection vs Blackbody Absorption/Emission

A blackbody emits what it absorbs ($f^2 \to R$), and it is thus natural to ask what makes this process different from simple reflection (e.g. $f \to -f$ with $f^2 \to f^2$)? The answer is that the mathematics/physics of blackbody radiation $f \to \ddot{u} - u'' - \gamma\dddot{u}$, is fundamenatlly different from simple reflection $f \to -f$. The string representing a blackbody is brought to vibration in resonance with forcing and the vibrating string string emits resonant radiation. Incoming waves thus are absorbed into the blackbody/string and then are emitted depending on the body temperature. In simple reflection there is no absorbing/emitting body, just a reflective surface without temperature.

176.14 Blackbody as Transformer of Radiation

The Earth absorbs incident radiation from the Sun with a Planck frequency distribution characteristic of the Sun surface temperature of about 5778 K and an amplitude depending on the ratio of the Sun's diameter to the distance of the Earth from the Sun. The Earth as a blackbody transforms the incoming radiation to a outgoing blackbody radiation of temperature about 288 K, so that total incoming and outgoing energy balances.

The Earth thus acts as a transformer of radiation and transforms incoming high-frequency low-amplitude radiation to outgoing low-frequency high-amplitude radiation under conservation of energy.

This means that high-frequency incoming radition is transformed into heat which shows up as low-frequency outgoing infrared radiation, so that the Earth emits more infrared radiation than it absorbs from the Sun. This increase of outgoing infrared radiation is not an effect of backradiation, since it would be present also without an atmosphere.

The spectra of the incoming blackbody radiation from the Sun and the outgoing infrared blackbody radiation from the Earth have little overlap, which means that the Earth as a blackbody transformer distributes incoming high-frequency energy so that all frequencies below cut-off obtain the same temperature. This connects to the basic assumption of statistical mechanics of *equidistribution in energy* or thermal equilibrium with one common temperature.

In the above model the absorbing blackbody inherits the equidistribution of the incoming radition (below cut-off) and thereby also emits an equidistributed spectrum. To ensure that an emitted spectrum is equidistributed even if the forcing is not, requires a mechanism driving the system towards equidistribution or thermal equilibrium.

176.15 Connection to Turbulence

The computational dissipation in our radiative model acts like turbulent dissipation in slightly viscous flow, in which high frequency coherent kinetic energy is transformed into heat energy in the form of small scale incoherent kinetic energy. The small coefficient γ in radiation corresponds to a small viscosity coefficient in fluid flow.

Since γ is small, the emitted wave is in one sense a small perturbation, but this is compensated by the third order derivate in the radiation term, with the effect that the radiated energy is not small. Or expressed differently: temperature involves first derivates (squared) and radiated energy a second derivative multiplied by a small factor. Without the dissipative radiation term, the string cannot emit the energy absorbed and the temperature will

then increase without limit. With radiation, the temperature will be limited by the temperature of the incoming wave.

176.16 CO_2 Climate Alarmism and Backradiation

> It is virtually certain that increasing atmospheric concentrations of carbon dioxide and other greenhouse gases will cause global surface climate to be warmer. (American Geophysical Union)
>
> We know the science, we see the threat, and we know the time for action is now (Arnold Schwarzenegger)
>
> There are many who still do not believe that global warming is a problem at all. And it's no wonder: because they are the targets of a massive and well-organized campaign of disinformation lavishly funded by polluters who are determined to prevent any action to reduce the greenhouse gas emissions that cause global warming out of a fear that their profits might be affected if they had to stop dumping so much pollution into the atmosphere. (Al Gore)

Global climate can be described as a thermodynamic system with gravitation subject to radiative forcing by blackbody radiation. Understanding climate thus requires understanding blackbody radiation.

We have learned in this chapter that "backradiation" is unphysical because it is unstable. Since climate alarmism feeds on a "greeenhouse effect" based on "backradiation" as shown in NASA's energy budget in Fig. 176.16, removing backradiation removes the main energy source of CO_2 climate alarmism.

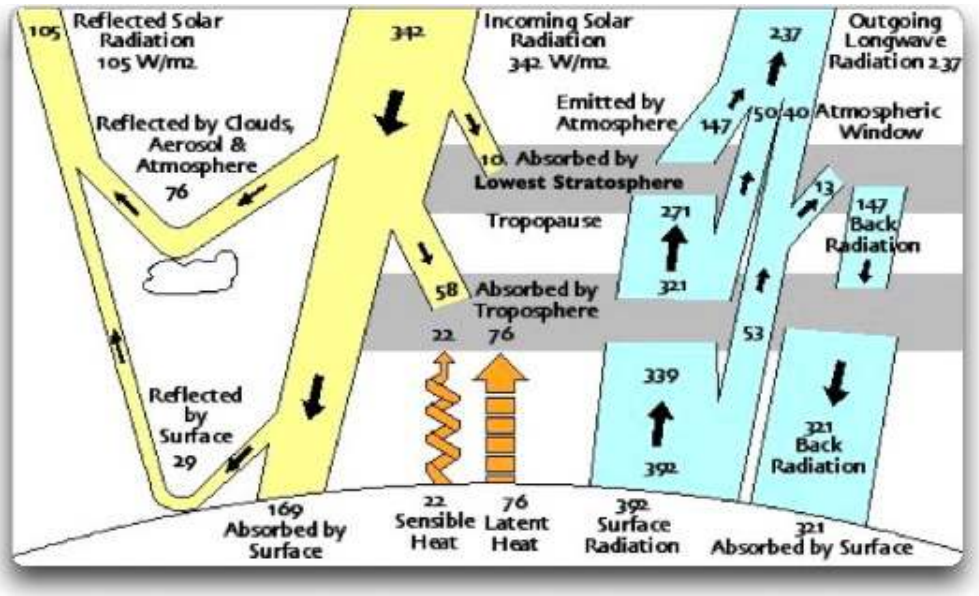


FIGURE 176.5. The Earth energy budget according to NASA [9] with incorrect unphysical 100% backradiation and 117% = 390 W/m^2 outgoing radiation from the Earth surface, but with correct physical 30% out of absorbed 48% transported by convection/evaporation from the Earth surface to the atmosphere.

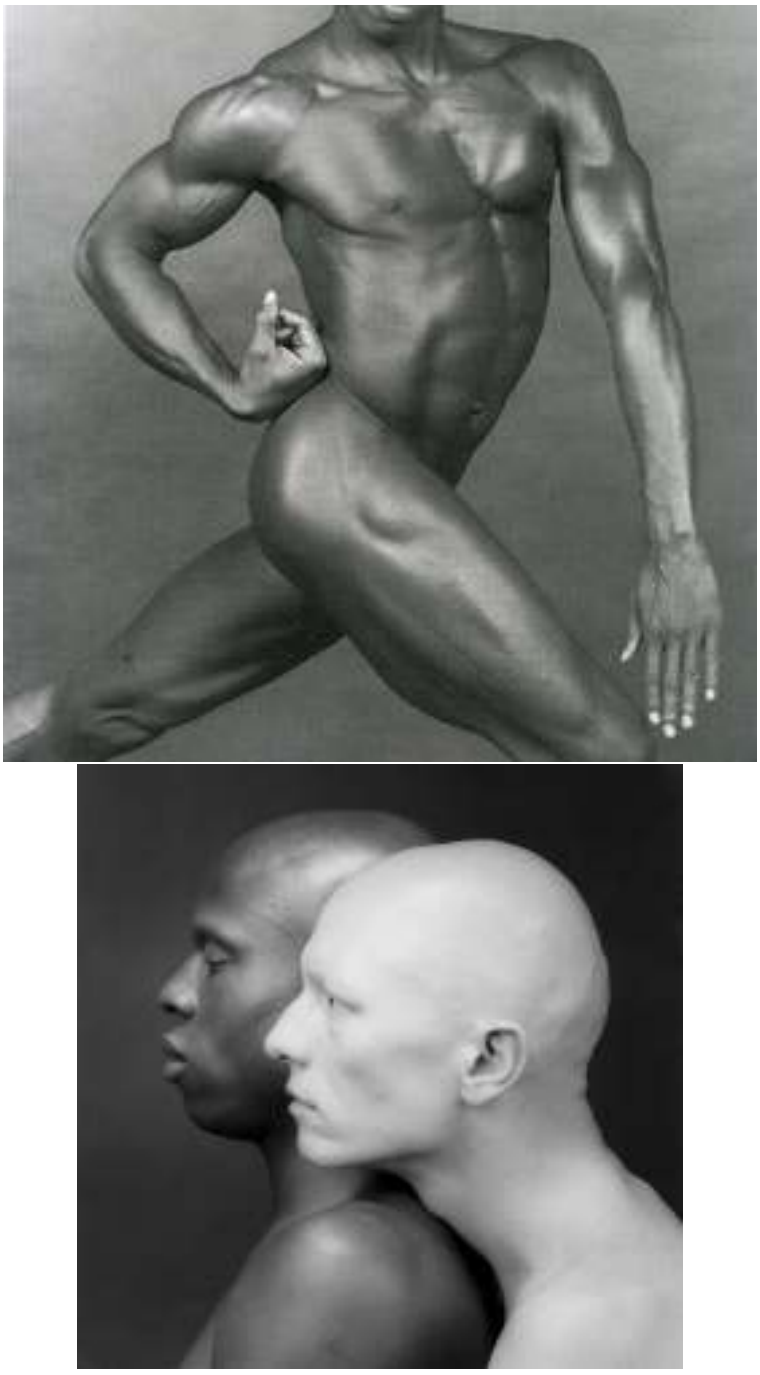


FIGURE 176.6. Black body by Mapplethorpe.

References

[1] Arthur C. Clarke: *If a scientist says that something is possible he is almost certainly right, but if he says that it is impossible he is probably wrong.*

[2] A. Einstein, On a Heuristic Point of View Toward the Emission and Transformation of Light, Ann. Phys. 17, 132, 1905.

[3] Einstein: *I consider it quite possible that physics cannot be based on the field concept, i.e., on continuous structures. In that case, nothing remains of my entire castle in the air, gravitation theory included, and of the rest of physics.* (Einstein 1954)

[4] Einstein: *What I wanted to say was just this: In the present circumstances the only profession I would choose would be one where earning a living had nothing to do with the search for knowledge.* (Einstein's last letter to Born Jan 17 1955 shortly before his death on the 18th of April, probably referring to Born's statistical interpretation of quantum mechanics).

[5] J. Hoffman and C. Johnson, *Computation Turbulent Incompressible Flow*, Springer 2008.

[6] J. Hoffman and C. Johnson, *Computational Thermodynamics*, http://www.nada.kth.se/cgjoh/ambsthermo.pdf

[7] Thomas Kuhn, Black-Body Theory and the Quantum Discontinuity, 1894-1912, Oxford Univ Press 1978.

[8] Robert A Millikan, *The electron and the light-quanta from the experimental point of view*, Nobel Lecture, May 23, 1923.

[9] NASA Earth Observatory $http://earthobservatory.nasa.gov/Features/EnergyBalance/$.

[10] Max Planck, Acht Vorlesungen über Theoretische Physik, Fünfte Vorlesung: Wärmestrahlung und Elektrodynamische Theorie, Leipzig, 1910.

[11] Schrodinger, The Interpretation of Quantum Physics. Ox Bow Press, Woodbridge, CN, 1995: *What we observe as material bodies and forces are nothing but shapes and variations in the structure of space. Particles are just schaumkommen (appearances). ...Let me say at the outset, that in this discourse, I am opposing not a few special statements of quantum physics held today (1950s), I am opposing as it were the whole of it, I am opposing its basic views that have been shaped 25 years ago, when Max Born put forward his probability interpretation, which was accepted by almost everybody.*

[12] H.D. Zeh, Physics Letters A 172, 189-192, 1993.

177
Human Speech

177.1 To Read

- Compressible flow with acoustics

177.2 To Watch

- KTH 3d vocal tract project
- Physical vocal tract model
- Vocal tract simulation
- Vocal folds
- Text to Speech.

177.3 Simulator

Construct a simulator for human speech based on the compressible Navier-Stokes equations in a variable domain of the vocal tract, mouth, tongue and lips. Model the generation of tones by fluid-structure interaction in the flow of air past the vocal folds.

We model the dynamical acoustics of human speech by the Euler equations for an inviscid perfect gas in a volume $\Omega(t)$ in $\mathbb{R}^3$ with boundary $\Gamma(t)$

changing with time t over a time interval $I = (0, 1]$, expressing conservation of *mass density* ρ, *momentum* $m = (m_1, m_2, m_3)$ and *internal energy* e: Find $\hat{u} = (\rho, m, e)$ depending on $(x, t) \in \Omega_I \equiv \cup_{t\in I}\Omega(t)$ such that

$$\begin{aligned} \dot{\rho} + \nabla\cdot(\rho u) &= 0 \quad \text{in } \Omega_I, \\ \dot{m} + \nabla\cdot(mu + p) &= f \quad \text{in } \Omega_I, \\ \dot{e} + \nabla\cdot(eu) + p\nabla\cdot u &= 0 \quad \text{in } \Omega_I, \\ u\cdot n &= 0 \quad \text{on } \Gamma_I \equiv \cup_{t\in I}\Gamma(t) \\ \hat{u}(\cdot, 0) &= \hat{u}^0 \quad \text{in } \Omega(0), \end{aligned} \tag{177.1}$$

where $u = \frac{m}{\rho}$ is the velocity, $p = (\gamma - 1)e = \bar{\gamma}e$ is the pressure with $1 > \bar{\gamma} = \gamma - 1 > 0$ a *gas constant* with $T = e/\rho$ temperature, f is a given volume force and $\hat{u}^0$ a given initial state. The domain Ω_I contains the time-variation of vocal chords, vocal tract with mouth, teeth and lips and volumes for air entrance and exit.

The vowels and consonants of human speech are produced by pulmonary pressure provided by the lungs. The vowels result from interaction of the glottis in the larynx with the air flow, which generate sound waves which are modified by the vocal tract. The consonants result from interaction of the air flow with the tongue, teeth and lips into plosives and fricatives. Human speech consists of time sequences of vowels in the form of sound waves of density-momentum variation and consonants in the form of turbulent aerodynamics.

In this note we report on simulations of human speech by computing time-dependent solutions of the Euler equations over a domain i space which varies with time. We focus here on the aerodynamics including sound waves, and consider fluid-structure interaction of glottis, vocal tract, tongue and lips in an upcoming report.

177.4 The Compressible Euler Equations with Acoustics

The sound waves appear as small variations in density-momentum. To capture this effect in computational simulation we augument the Euler equations for aerodynamics with a linearized wave equation for density-momentum into a system of the following form: Find $\hat{u} = (\rho, m, e, \rho_a, m_a)$ such that

$$\begin{aligned} \dot{\rho} + \nabla\cdot(\rho u) &= 0 \quad \text{in } \Omega_I, \\ \dot{m} + \nabla\cdot(mu + p) &= f \quad \text{in } \Omega_I, \\ \dot{e} + \nabla\cdot(eu) + p\nabla\cdot u &= 0 \quad \text{in } \Omega_I, \\ \dot{\rho}_a + \nabla\cdot m_a &= 0 \quad \text{in } \Omega_I, \\ \dot{m}_a + \nabla\cdot(m_a u) + \bar{\gamma}T\nabla\rho_a &= 0 \quad \text{in } \Omega_I, \\ u\cdot n &= 0 \quad \text{on } \Gamma_I \\ \hat{u}(\cdot, 0) &= \hat{u}^0 \quad \text{in } \Omega(0), \end{aligned} \tag{177.2}$$

with $p = \gamma\rho T$, where ρ_a and m_a represent variations of density and momentum. The acoustic signal is represented by the total pressure given by $p = \bar{\gamma}(\rho+\rho_a)T$ at exit. In the linearized density-momentum wave equation, we do not here account for variations in u and T.

177.5 Incompressible Aerodynamics with Sound Waves

Since the Mach number of the airflow of human speech is small it may be computationally cost effective to replace the compressible aerodynamics by incompressible aerodynamics into the following model: Find $\hat{u} = (u, p, \rho_a, m_a)$ such that

$$\begin{aligned} \dot{u} + u\cdot\nabla u + \nabla p &= f \quad \text{in } \Omega_I, \\ \nabla\cdot u &= 0 \quad \text{in } \Omega_I, \\ \dot{\rho}_a + \nabla\cdot m_a &= 0 \quad \text{in }, \\ \dot{m}_a + \nabla\cdot(m_a u) + \bar{\gamma}T\nabla\rho_a &= 0 \quad \text{in } \Omega_I, \\ u\cdot n &= 0 \quad \text{on } \Gamma_I \\ \hat{u}(\cdot,0) &= \hat{u}^0 \quad \text{in } \Omega(0), \end{aligned} \tag{177.3}$$

where T is given and the acoustic signal is represented by the total pressure given by $p + \bar{\gamma}\rho_a T$ at exit.

178
Global Circulation Models

178.1 General Circulation Models and GCMG2

Global climate simulation is based on computational solution of a system of partial differential equations describing the coupled ocean/atmosphere circulation of water/air driven by Solar radiation and Coriolis forces from the rotation of the Earth, in a *General Circulation Model GCM* [37]. A main objective is to predict *climate sensitivity* as the change of global temperature upon doubling of the concentration of CO_2 in the atmosphere related to *Anthropogenic Global Warming AGW* by burning of fossil fuels, but many other aspects of global climate can also be subject to studies by GCMs.

The *International Panel of Climate Change IPCC* predicts [33] climate sensitivities between 1 and 5 degrees Celcius, which without CO_2 emission control cold be reached in 2050. AGW of 5 degrees is viewed to be catastropical, and IPCC has put pressure on global politics to enforce limits on emissions of CO_2. The G-20 group of industrial and developing countries have agreed to limit AGW to 2 degrees, without however any agreement of how this could be accomplished in case it would be necessary.

The global climate system is also subject to natural variations of largely unknown nature and magnitude and the impact of atmospheric CO_2 is also largely unknown. Altogether, reliaable prediction of the global climate 100 years ahead seems as impossible as daily weather prediction 100 days ahead. In any case assessment of the relibility under different scenarios of natural variation is necessary, if GCM is going to serve as science and not

just politics. Of course, the real value of a prediction of climate sensitivity between 1 and 5 degrees is questionable, but that does not prevent IPCC from recommending strict emission control, which for some reasons is embraced by in particular EU politicians.

The basic GCM for the coupled ocean/atmosphere is the *Navier-Stokes equations* the flow of water/air viewed as a fluid of *variable density* which is incompressible in water and compressible in air, combined with a *transport equation* for *salinity* in the ocean and *moisture* in the atmosphere. We assume

We apply the *General Galerkin G2* [4] finite element method to this model to obtain a computational GCM named *GCMG2*. We show GCMG2 to be a versatile tool allowing efficient simulation of certain aspects of global climate making use of the following features and capabilities:

- full 3d incompressible Naviers-Stokes equations without Boussinesq or hydrostatic approximation [39, 40],
- automatic adaptive duality-based a posteriori error control allowing objective assessment of reliability,
- automatic modeling of interior turbulence by finite element stabilization,
- modeling of turbulent boundary layer by skin friction
- automatic seamless coupling of ocean and atmosphere or by skin friction,
- automatic handling of moving ocean/atmosphere interface allowing also breaking waves,
- vertically moving meshes for enhanced wave propagation,
- automatized efficient open-source implementation in FEniCS/Unicorn [25].

We illustate in a sequence of test problems including Rayleigh-Taylor instability, breaking wave, thermohaline circulation and hurricane formation.

Ocean simulation requires input of the density as a function of temperature and salinity, which can be determined experimentally. Atmosphere simulation requires modeling of cloud formation influencing incoming and ougoing radiation, which largely is a open problem. GCM with coupled ocean/circulation thus largely is an open problem. In this note we focus on reliability with respect to computational discretization and leave out open modeling problems.

178.2 State of the Art and Beyond

The state of the art of climate modeling is represented by Atmosphere-Ocean General Circulation Models or AOGCMs including coupling of ocean-atmosphere and models for radiation, cloud formation, sea ice, plants, soil et cet. These codes use finite difference or spectral discretization on uniform horisontal grids of mesh size about 250x250 km and 30 vertical layers possibly following the seafloor and Earth surface using curvilinear coordinates, with a total number of mesh points of about one million. The codes solve the so called *primitive equations*, which are simplified Navier-Stokes equations with approximate vertical momentum balance dominated by hydrostatic pressure and Boussinesq approximation with variable density only in bouyancy terms.

GCMG2 offers the new features listed above not present in existing GCMs. In coming publications we will further explore the capability of GCMG2 including extensions to fluid-structure interaction in problems on human scales such as simulation of a complete sailing boat with hull, sail and free water surface, breaking dam...

178.3 The Navier-Stokes Equations as GCM

We describe coupled ocean/water circulation by the *Navier-Stokes equations* for a slightly viscous fluid of variable density filling the volume Ω in $\mathbb{R}^3$ occupied by water/air with boundary Γ representing the seafloor, over a time interval $I = [0, \bar{t}]$. The basic dependent variables of the Navier-Stokes equation are the water/air fluid density, velocity u, pressure p, total energy $\epsilon = \rho|u|^2/2 + e$ with ρ density and e heat energy, and salinity/moisture s combined with a constitutive law for density $\rho = \rho(T, S, p)$ given as a function of $T = e/\rho$ temperature and $S = s/\rho$ salinity per unit mass, and also pressure if the fluid is compressible. We assume water to be incompressible and air to be a perfect gas with $p = \rho T$ modulo dependence on S.

We consider the following system of equations expressing conservation of mass, momemtum, total energy and salinity/moisture combined with incompressibility/perfect gas law and boundary/intial values: Find $\hat{u} = (\rho, u, e, s)$ depending on $(x, t) \in \Omega \cup \Gamma \times I$, such that with $m = \rho u$ momentum

and $Q \equiv \Omega \times I$

$$\begin{aligned}
\dot\rho + \nabla\cdot(\rho u) &= 0 && \text{in } Q,\\
\dot m + \nabla\cdot(mu) + \nabla p - \nabla\cdot\sigma - \rho g - f &= 0 && \text{in } Q,\\
\dot\epsilon + \nabla\cdot(\epsilon u + pu) - \rho g\cdot u &= r && \text{in } Q,\\
\dot s + \nabla\cdot(su) &= 0 && \text{in } Q,\\
\nabla\cdot u = 0 \,\text{in water and } p &= \rho T \,\text{in air} && \text{in } Q\\
u_n &= 0 && \text{on } \Gamma\times I,\\
\sigma_s &= \beta u_s && \text{on } \Gamma\times I,\\
\hat u(\cdot,0) &= \hat u^0 && \text{in } \Omega,
\end{aligned} \tag{178.1}$$

where $\dot u = \frac{\partial u}{\partial t}$, u_n is the fluid velocity normal to Γ, u_s is the tangential velocity, $\sigma = 2\nu\epsilon(u)$ is viscous stress with $\epsilon(u)$ the usual velocity strain and ν the fluid viscosity, σ_s is the tangential stress, β is boundary skin friction, g is gravitational force per unit mass, f is Coriolis force, r is radiative heat source and $\hat u^0$ is a given initial condition. For simplicity, we set heat and salinity/moisture diffusivities to zero, motivated by the fact that these coefficients are small. The skin friction boundary condition models a turbulent boundary layer [4] and can also be used in explicit coupling of ocean and atmosphere.

If we assume also air to be incompressible, which can be motivated for problems on smaller scales, then we can rewrite the system (178.8) using conservation of mass into: Find $\hat u = (u, e, S)$ and p such that

$$\begin{aligned}
\rho(\dot u + (u\cdot\nabla)u) + \nabla p - \nabla\cdot\sigma - g\rho - f &= 0 && \text{in } Q,\\
\nabla\cdot u &= 0 && \text{in } Q,\\
\dot e + (u\cdot\nabla)e - (\nabla\cdot\sigma)\cdot u &= r && \text{in } Q,\\
\dot S + (u\cdot\nabla)S &= 0 && \text{in } Q,\\
u_n &= 0 && \text{on } \Gamma\times I,\\
\sigma_s &= \beta u_s && \text{on } \Gamma\times I,\\
\hat u(\cdot,0) &= \hat u^0 && \text{in } \Omega,
\end{aligned} \tag{178.2}$$

where here $\rho = \rho(T, S)$ is a given function. This also models large scale ocean circulation since water is nearly incompressible also for large pressures.

Without salinity this model is close to the Navier-Stokes equations for variable density incompressible flow: Find $\hat u = (u, \rho)$ and p such that

$$\begin{aligned}
\dot\rho + (u\cdot\nabla)\rho &= 0 && \text{in } Q,\\
\rho(\dot u + (u\cdot\nabla)u) + \nabla p - \nabla\cdot\sigma - g\rho - f &= 0 && \text{in } Q,\\
\nabla\cdot u &= 0 && \text{in } Q,\\
u_n &= 0 && \text{on } \Gamma\times I,\\
\sigma_s &= \beta u_s && \text{on } \Gamma\times I,\\
\hat u(\cdot,0) &= \hat u^0 && \text{in } \Omega.
\end{aligned} \tag{178.3}$$

Here, temperature could be computed a posteriori using the equation for internal energy in (178.9) with in particular a contribution from turbuent dissipation included through $-(\nabla\cdot\sigma)\cdot u$.

178.4 Bouyancy Stability-Instability

A basic problem concerns the stability of an incompressible variable density fluid at a zero-velocity rest state. We are familiar with the fact that a heavier fluid on top of a ligther fluid is unstable, referred to as Rayleigh-Taylor instability, and we now seek a mathematical explanation of this effect: Linearizing the equations for conservation of mass and momentum (178.3), assuming $\sigma = 0$, we obtain the following equations for perturbations (r, v, q) of $(\rho, u = 0, p)$:

$$\begin{aligned} \dot r + v\cdot\nabla\rho &= 0,\\ \rho\dot v + \nabla q - gr &= 0,\\ \nabla\cdot v &= 0. \end{aligned} \tag{178.4}$$

Simplifying to only dependence on the vertical coordinate, we have

$$\begin{aligned} \dot r + v_3\frac{\partial\rho}{\partial x_3} &= 0,\\ \rho\dot v_3 - g_3 r &= 0, \end{aligned} \tag{178.5}$$

with x_3 the vertical coordinate directed upwards, we find a system which is unstable if $\frac{\partial\rho}{\partial x_3} > 0$ and marginally stable if $\frac{\partial\rho}{\partial x_3}\le 0$, which we wanted to show.

178.5 GCMG2

GCMG2 is obtained by direct application of G2 to (178.9) with automatized implementation in [25], with trial functions being continuous and piecewise linear in space-time, and test function being continuous piecewise linear in space and piecewise constant in time, on a space-time mesh of mesh size h, assuming velocity trial/test-fuctions v satisfy $v\cdot n = 0$ on Γ. Denoting the corresponding finite element spaces by U_h and V_h respectively GCMG2 takes the form: Find $\hat u\in U_h$ with $\hat u(\cdot,0)$ given, such that

$$B(\hat u,\hat v) = 0 \quad \text{for all } \hat v\equiv(v,q,\tau,s)\in W_h, \tag{178.6}$$

where

$$\begin{aligned} B(\hat u,\hat v) &= (\rho(\dot u + u\cdot\nabla u + \nabla p - \rho g - f), v + \delta(\rho u\cdot\nabla v + \nabla q))_Q + (\nabla\cdot u, q)_Q\\ &\quad + (\dot T + u\cdot\nabla T, \tau + \delta u\cdot\nabla\tau)_Q + (\dot S + u\cdot\nabla S, s + \delta u\cdot\nabla s)_Q \end{aligned} \tag{178.7}$$

with $(\cdot,\cdot)_Q$ appropriate $L_2(Q)$ scalar products and $\delta = h/|u|$ a stabilization parameter. Standard shock-capturing [4] is used at the ocean ocean-atmosphere interface in the case of breaking waves.

The input to FEniCS is the form $B(\hat{u},\hat{v})$, the function $\rho = \rho(T,S)$, the Coriolis force $f = 2\Omega \times v$ with Ω a given rotation, and the intial value $\hat{u}(\cdot,0)$.

With comprressible atmosphere GMCG2 uses the above formulation for the ocean combined with G2 for compressible flow as presented in [5].

178.6 Thermohaline Circulation

Thermohaline circulation refers to the large-scale circulation of the Ocean Conveyor Belt driven by density gradients resulting from varying temperature and salinity as the warm surface water of the Northbound Gulf Stream cools off and saltifies by evaporation and sinks to form the North Atlantic Deep Water moving South.

We study here a model of thermohaline circulation driven by a source and sink of density in the form of variable-density incompressible NS.

178.7 The Salter Sink Model

178.8 General Circulation Models

Global and local climate simulation is based on computational solution of a system of partial differential equations describing the coupled ocean/atmosphere circulation of water/air driven by Solar radiation and Coriolis forces from the rotation of the Earth, in a *General Circulation Model GCM* [37].

The basic GCM for the coupled ocean/atmosphere is the *Navier-Stokes equations* for the *turbulent flow* of water/air viewed as a fluid of *variable density* and *small viscosity*, which is *incompressible* in water and *compressible* in air, combined with a *transport equation* for *salinity* in the ocean and *moisture* in the atmosphere [39], [40].

We report on computational simulation of a projected device consisting of a vertical tube immersed in the ocean with top inlet of warm surface water and bottom outlet in cooler deeper water, driven by incoming waves, for the purpose of preventing formation of hurricanes by lowering the temperature of the surface water, referred to as the *Salter Sink*.

178.9 The Navier-Stokes Equations as GCM

We describe coupled ocean/water circulation by the *Navier-Stokes equations* for a slightly viscous fluid of variable density filling the volume Ω in $\mathbb{R}^3$ occupied by water/air with boundary Γ representing the seafloor, over a time interval $I = [0, \bar{t}]$. The basic dependent variables of the Navier-Stokes equation are the water/air fluid density, velocity u, pressure p, total energy $\epsilon = \rho|u|^2/2 + e$ with ρ density and e heat energy, and salinity/moisture s combined with a constitutive law for density $\rho = \rho(T, S, p)$ given as a function of $T = e/\rho$ temperature and $S = s/\rho$ salinity per unit mass, and also pressure if the fluid is compressible. We assume water to be incompressible and air to be a perfect gas with $p = (\gamma - 1)\rho T$ modulo dependence on S, with γ a gas constant $= 0.4$ for air.

We consider the following system of equations expressing conservation of mass, momentum, total energy and salinity/moisture combined with incompressibility/perfect gas law and boundary/initial values: Find $\hat{u} = (\rho, u, \epsilon, s)$ depending on $(x, t) \in \Omega \cup \Gamma \times I$, such that with $m = \rho u$ momentum and $Q \equiv \Omega \times I$

$$
\begin{aligned}
\dot{\rho} + \nabla \cdot (\rho u) &= 0 && \text{in } Q,\\
\dot{m} + \nabla \cdot (mu) + \nabla p - \nabla \cdot \sigma - \rho g - f &= 0 && \text{in } Q,\\
\dot{\epsilon} + \nabla \cdot (\epsilon u + pu) - \rho g \cdot u &= R && \text{in } Q,\\
\dot{s} + \nabla \cdot (su) &= 0 && \text{in } Q,\\
\nabla \cdot u = 0 \text{ in water and } p &= (\gamma - 1)\rho T \text{ in air} && \text{in } Q\\
u_n &= 0 && \text{on } \Gamma \times I,\\
\sigma_s &= \beta u_s && \text{on } \Gamma \times I,\\
\hat{u}(\cdot, 0) &= \hat{u}^0 && \text{in } \Omega,
\end{aligned}
\tag{178.8}
$$

where $\dot{u} = \frac{\partial u}{\partial t}$, u_n is the fluid velocity normal to Γ, u_s is the tangential velocity, $\sigma = 2\nu\epsilon(u)$ is viscous stress with $\epsilon(u)$ the usual velocity strain and ν the fluid viscosity, σ_s is the tangential stress, β is boundary skin friction, g is gravitational force per unit mass, f is Coriolis force, R is radiative heat source and $\hat{u}^0$ is a given initial condition. For simplicity, we set heat and salinity/moisture diffusivities to zero, motivated by the fact that these coefficients are small. The skin friction boundary condition models a turbulent boundary layer [4] and can also be used in explicit coupling of ocean and atmosphere.

If we assume also air to be incompressible, which can be motivated for problems on smaller scales, and here leave out dependence on salinity for simplicity, then we can rewrite the system (178.8) as the Navier-Stokes equations for variable density incompressible flow: Find $\hat{u} = (\rho, u, p)$ such

that

$$\begin{aligned} \dot\rho + (u\cdot\nabla)\rho &= 0 && \text{in } Q,\\ \rho(\dot u + (u\cdot\nabla)u) + \nabla p - \nabla\cdot\sigma - g\rho - f &= 0 && \text{in } Q,\\ \nabla\cdot u &= 0 && \text{in } Q,\\ u_n &= 0 && \text{on } \Gamma\times I,\\ \sigma_s &= \beta u_s && \text{on } \Gamma\times I,\\ \rho(\cdot,0)=\rho^0,\quad u(\cdot,0) &= u^0 && \text{in } \Omega. \end{aligned} \tag{178.9}$$

In this case the temperature $T = e/\rho$ can be computed a posteriori through the following equation for the internal energy e

$$\dot e + (u\cdot\nabla)e = -(\nabla\cdot\sigma)\cdot u + R \text{ in } Q, \tag{178.10}$$

where $-(\nabla\cdot\sigma)\cdot u$ is a source of heat from turbulent dissipation, and $T(\cdot,0)$ is coupled to $\rho(\cdot,0)$ by a relation e.g. determined by experiment.

178.10 G2 for Variable-Density Incompressible Flow

We apply the G2 finite element method [4] to (178.9) with $\beta = \nu = 0$ and $f = 0$ with trial functions being continuous and piecewise linear in space-time, and test function being continuous piecewise linear in space and piecewise constant in time, on a space-time mesh of mesh size h, assuming velocity trial/test-fuctions v satisfy $v\cdot n = 0$ on Γ, refereed to as the cG(1)cG(1) variant of G2. Denoting the corresponding finite element spaces by U_h and V_h respectively GCMG2 takes the form: Find $\hat u \in U_h$ with $u(\cdot,0)$ and $\rho(\cdot,0)$ given, such that

$$B(\hat u,\hat v) = 0 \quad \text{for all } \hat v \equiv (r,v,q) \in W_h, \tag{178.11}$$

where

$$\begin{aligned} B(\hat u,\hat v) = (\rho(\dot u + u\cdot\nabla u) + \nabla p - \rho g, v)_Q + (\nabla\cdot u, q)_Q + (\dot\rho + u\cdot\nabla\rho, r)_Q\\ +(\delta(\rho u\cdot\nabla u + \nabla p - \rho g), \rho u\cdot\nabla v + \nabla q))_Q + (\delta u\cdot\nabla\rho, u\cdot\nabla r)_Q \end{aligned} \tag{178.12}$$

with $(\cdot,\cdot)_Q$ appropriate $L_2(Q)$ scalar products and $\delta = h/|u|$ a stabilization parameter.

178.11 Simulations of Sink Circulation

The Salter Sink is a device for cooling an ocean surface by mixing warm surface water with deeper cooler water for the purpose of preventing the development of hurricanes. In this study it consists of a vertical tube of length

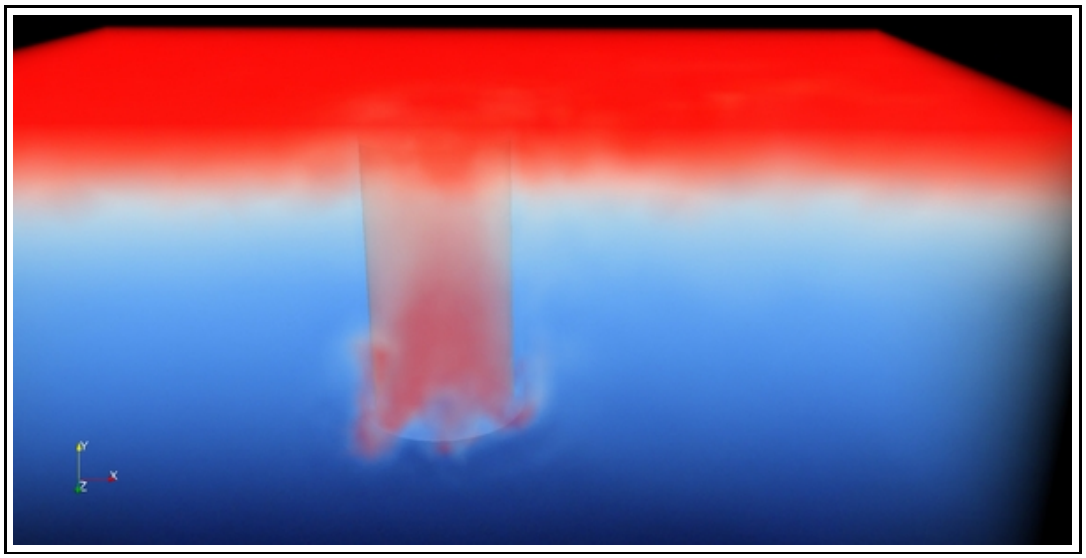


FIGURE 178.1. Simulation snapshot of the Salter Sink, showing a flow of warm surface water down the tube to get mixed with cooler deeper water, and then ascending by bouyancy to replace warm surface water with a cooling effect.

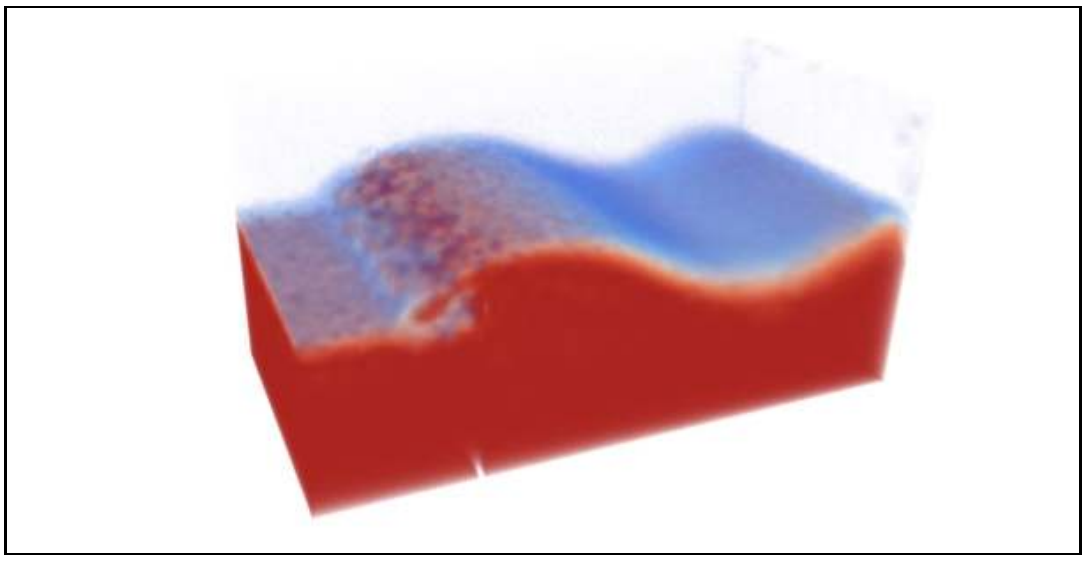


FIGURE 178.2. Simulation snapshot of a wave breaking over a wall and elevating the water surface behind the wall.

200 m and diameter 45 m immersed into an ocean of depth 600 m water in which light warm water of temperature 28 degrees Celscius is driven down the tube to get mixed with cooler water at a temperature of 10 degrees

at the bottom of the ocean, by an elevated water level inside the tube maintained by incoming water waves. We model the flow computationally by G2 for variable density incompressible Navier-Stokes equations (178.9) with a given relation between initial temperature and density determined experimentally.

We first simulate the circulation in the sink driven by a vertical force inside the sink modeling the pressure increase from an elevated water surface inside the sink of height 0.15 m and 0.3 m. In this simulation the water surface is fixed and the computation is restricted to the water, as displayed in Fig. 178.1. The sink is immersed in an ocean current of 0.1 m/s. The temperature variation was determined from the density variation by the experimental relation, instead of a posteriori solving the energy equation as indicated above.

We then simulate the breaking of a water wave as a variable density air-water system as shown in Fig. 178.2.

The next step is to simulate the complete action of the sink with breaking waves maintaining an elevated water surface inside the sink, and in a further step take also fluid-structure interaction into account into a full simulation including all basic aspects.

178.12 Thermohaline Circulation

178.13 General Circulation Models with G2

Global climate simulation is based on computational solution of a system of partial differential equations describing the coupled ocean/atmosphere circulation of water/air driven by Solar radiation and Coriolis forces from the rotation of the Earth, in a *General Circulation Model GCM* [37]. A main objective is to predict *climate sensitivity* as the change of global temperature upon doubling of the concentration of CO_2 in the atmosphere related to *Anthropogenic Global Warming AGW* by burning of fossil fuels, but many other aspects of global climate can also be subject to studies by GCMs.

The basic GCM for the coupled ocean/atmosphere is the *Navier-Stokes equations* the flow of water/air viewed as a fluid of *variable density* which is incompressible in water and compressible in air, combined with a *transport equation* for *salinity* in the ocean and *moisture* in the atmosphere.

In this note we focus on thermohaline circulation (THC), which is the large-scale ocean circulation named the Great Ocean Conveyor Belt driven by density gradients resulting from varying temperature and salinity as the warm surface water of the Northbound Gulf Stream cools off and saltifies by evaporation and sinks to form the North Atlantic Deep Water moving South.

We simulate THC by computational solution of the Navier-Stokes equations for variable density turbulent incompressible flow using the *General Galerkin G2* [4] finite element method with following features and capabilities:

- full 3d incompressible Naviers-Stokes equations without Boussinesq or hydrostatic approximation [39, 40],
- automatic adaptive duality-based a posteriori error control allowing objective assessment of reliability,
- automatic modeling of interior turbulence by finite element stabilization,
- modeling of turbulent boundary layer by skin friction
- automatized efficient open-source implementation in FEniCS/Unicorn [25].

For simplicity we here simulate variable salinity by sources and sinks of density.

178.14 The Navier-Stokes Equations for Variable Density Flow

We consider by the *Navier-Stokes equations* for a slightly viscous incompressible fluid of variable density filling the volume Ω in $\mathbb{R}^3$ with boundary Γ over a time interval $I = [0, \bar{t}]$: Find the velocity u, density ρ density and pressure p such that

$$\begin{aligned}
\dot{\rho} + (u \cdot \nabla)\rho &= d && && \text{in } Q,\\
\rho(\dot{u} + (u \cdot \nabla)u) + \nabla p - \nabla \cdot \sigma - g\rho - f &= 0 && && \text{in } Q,\\
\nabla \cdot u &= 0 && && \text{in } Q,\\
u_n &= 0 && && \text{on } \Gamma \times I,\\
\sigma_s &= \beta u_s && && \text{on } \Gamma \times I,\\
u(\cdot, 0) &= u^0 && \rho(\cdot, 0) = \rho^0 && \text{in } \Omega,
\end{aligned} \tag{178.13}$$

where $\dot{u} = \frac{\partial u}{\partial t}$, u_n is the fluid velocity normal to Γ, u_s is the tangential velocity, $\sigma = 2\nu\epsilon(u)$ is viscous stress with $\epsilon(u)$ the usual velocity strain and ν the fluid viscosity, σ_s is the tangential stress, β is boundary skin friction, g is gravitational force per unit mass, f is Coriolis force, d is a density source/sink and u^0 and ρ^0 are given initial values.

178.15 Bouyancy Stability-Instability

A basic problem concerns the stability of an incompressible variable density fluid at a zero-velocity rest state. We are familiar with the fact that a heavier fluid on top of a ligther fluid is unstable, referred to as Rayleigh-Taylor instability, and we now seek a mathematical explanation of this effect: Linearizing the equations for conservation of mass and momentum (178.13), assuming $\sigma = 0$, we obtain the following equations for perturbations (r, v, q) of $(\rho, u = 0, p)$:

$$\begin{aligned} \dot{r} + v \cdot \nabla\rho &= 0, \\ \rho\dot{v} + \nabla q + gr &= 0, \\ \nabla \cdot v &= 0. \end{aligned} \tag{178.14}$$

Simplifying to only dependence on the vertical coordinate, we have

$$\begin{aligned} \dot{r} + v_3 \frac{\partial rho}{\partial x_3} &= 0, \\ \rho\dot{v}_3 - g_3 r &= 0, \end{aligned} \tag{178.15}$$

with x_3 the vertical coordinate directed upwards, we find a system which is stable if $\frac{\partial\rho}{\partial x_3} < 0$ and marginally stable if $\frac{\partial\rho}{\partial x_3} \geq 0$, which we wanted to show.

178.16 G2 for Variable Density Flow

We apply G2 to (178.13) with automatized implementation in [25], with trial functions being continuous and piecewise linear in space-time, and test function being continuous piecewise linear in space and piecewise constant in time, on a space-time mesh of mesh size h, assuming velocity trial/test-fuctions v satisfy $v \cdot n = 0$ on Γ. Denoting the corresponding finite element spaces by U_h and V_h respectively G2 takes the form: Find $\hat{u} = (\rho, u, p) \in U_h$ with $u(\cdot, 0)$ and $\rho(\cdot, 0)$ given, such that

$$B(\hat{u}, \hat{v}) = 0 \quad \text{for all } \hat{v} \equiv (v, q, \tau, s) \in W_h, \tag{178.16}$$

where

$$\begin{aligned} B(\hat{u}, \hat{v}) = &(\rho(\dot{u} + u \cdot \nabla u + \nabla p - \rho g - f), v + \delta(\rho u \cdot \nabla v + \nabla q))_Q + (\nabla \cdot u, q)_Q \\ &+ (\dot{\rho} + u \cdot \nabla\rho - d, \tau + \delta u \cdot \nabla\tau)_Q \end{aligned} \tag{178.17}$$

with $(\cdot, \cdot)_Q$ appropriate $L_2(Q)$ scalar products and $\delta = h/|u|$ a stabilization parameter. The input to FEniCS is the form $B(\hat{u}, \hat{v})$, the Coriolis force $f = 2\Omega \times v$ with Ω a given rotation, and the intial values $u(\cdot, 0)$ and $\rho(\cdot, 0)$.

178.17 An Basic Model Example

As a basic model we consider THC in a flat rectangular box driven by a density source close to the surface at one end and a density sink close to the bottom at the other end. We find that a circulating stream is generated and and study velocity distribution and the turbulent mixing of the stream into the surrounding fluid.

References

[1] Global climate models, $http://en.wikipedia.org/wiki/Global_climate_model$

[2] G. Birkhoff, Hydrodynamics, Princeton University Press, 1950.

[3] S. Cowley, Laminar boundary layer theory: A 20th century paradox, Proceedings of ICTAM 2000, eds. H. Aref and J.W. Phillips, 389-411, Kluwer (2001).

[4] A. Crook, Skin friction estimation at high Reynolds numbers and Reynolds-number effects for transport aircraft, Center for Turbulence Research, 2002.

[5] J.Hoffman, Simulation of turbulent flow past bluff bodies on coarse meshes using General Galerkin methods: drag crisis and turbulent Euler solutions, Comp. Mech. 38 pp.390-402, 2006.

[6] J. Hoffman, Simulating Drag Crisis for a Sphere using Friction Boundary Conditions, Proc. ECCOMAS, 2006.

[7] FEniCs, www.fenics.org.

[8] J. Hoffman and C. Johnson, Blowup of Euler solutions, BIT Numerical Mathematics, Vol 48, No 2, 285-307.

[9] J. Hoffman and C. Johnson, Mathematical Theory of Flight, 2009.

[10] J. Hoffman and C. Johnson, Computational Turbulent Incompressible Flow, Springer 2007, home page at www.bodysoulmath.org/books.

[11] J. Hoffman, C. Johnson and M. Nazarov, Computational Thermodynamics, Icarus Ebooks.

[12] J. Hoffman and C. Johnson, Resolution of d'Alembert's paradox, Journal of Mathematical Fluid Mechanics, Online First Dec 10, 2008.

[13] J. Hoffman and C. Johnson, Modeling Turbulent Boundary Layers by Small Friction.

[14] J. Hoffman and Claes Johnson, Knol articles.

[15] IPCC Assessment Report 4, 2007, www.ipcc.ch.

[16] J. Kim and P. Moin, Tackling Turbulence with Supercomputer, Scientific American.

[17] Lions J L, Temam R, Wang S. New formulations of the primitive equations of atmosphere and applications. Nonlinearity, 1992, 5: 237288

[18] Lions J L, Temam R, Wang S. On the equations of the large scale ocean. Nonlinearity, 1992, 5: 10071053

[19] Global climate models, $http://en.wikipedia.org/wiki/Global_climate_model$

[20] G. Birkhoff, Hydrodynamics, Princeton University Press, 1950.

[21] S. Cowley, Laminar boundary layer theory: A 20th century paradox, Proceedings of ICTAM 2000, eds. H. Aref and J.W. Phillips, 389-411, Kluwer (2001).

[22] A. Crook, Skin friction estimation at high Reynolds numbers and Reynolds-number effects for transport aircraft, Center for Turbulence Research, 2002.

[23] J.Hoffman, Simulation of turbulent flow past bluff bodies on coarse meshes using General Galerkin methods: drag crisis and turbulent Euler solutions, Comp. Mech. 38 pp.390-402, 2006.

[24] J. Hoffman, Simulating Drag Crisis for a Sphere using Friction Boundary Conditions, Proc. ECCOMAS, 2006.

[25] FEniCs, www.fenics.org.

[26] J. Hoffman and C. Johnson, Blowup of Euler solutions, BIT Numerical Mathematics, Vol 48, No 2, 285-307.

[27] J. Hoffman and C. Johnson, Mathematical Theory of Flight, 2009.

[28] J. Hoffman and C. Johnson, Computational Turbulent Incompressible Flow, Springer 2007, home page at www.bodysoulmath.org/books.

[29] J. Hoffman, C. Johnson and M. Nazarov, Computational Thermodynamics, Icarus Ebooks.

[30] J. Hoffman and C. Johnson, Resolution of d'Alembert's paradox, Journal of Mathematical Fluid Mechanics, Online First Dec 10, 2008.

[31] J. Hoffman and C. Johnson, Modeling Turbulent Boundary Layers by Small Friction.

[32] J. Hoffman and Claes Johnson, Knol articles.

[33] IPCC Assessment Report 4, 2007, www.ipcc.ch.

[34] J. Kim and P. Moin, Tackling Turbulence with Supercomputer, Scientific American.

[35] Lions J L, Temam R, Wang S. New formulations of the primitive equations of atmosphere and applications. Nonlinearity, 1992, 5: 237288

[36] Lions J L, Temam R, Wang S. On the equations of the large scale ocean. Nonlinearity, 1992, 5: 10071053

[37] Global climate models, $http://en.wikipedia.org/wiki/Global_climate_model$

[38] J. Hoffman and C. Johnson, Computational Turbulent Incompressible Flow, Springer 2007, home page at www.bodysoulmath.org/books.

[39] Lions J L, Temam R, Wang S. New formulations of the primitive equations of atmosphere and applications. Nonlinearity, 1992, 5: 237288

[40] Lions J L, Temam R, Wang S. On the equations of the large scale ocean. Nonlinearity, 1992, 5: 10071053

179
Climate Thermodynamics

179.1 Global Climate by Navier-Stokes Equations

> Thermodynamics is a funny subject. The first time you go through it, you dont understand it at all. The second time you go through it, you think you understand it, except for one or two small points. The third time you go through it, you know you dont understand it, but by that time you are so used to it, it doesnt bother you any more. (Physicist Arnold Sommerfeld (1868-1951))

Global climate results from a thermodynamic interaction between the atmosphere and the ocean with radiative forcing from the Sun, gravitational forcing from the Earth (and the Moon) and dynamic Coriolis forcing from the rotation of the Earth. The thermodynamics is described by the *Navier-Stokes equations* (NSE) of fluid dynamics, for a variable density incompressible ocean and compressible atmosphere, expressing conservation of mass, momentum and energy.

The atmosphere transports heat energy absorbed by the Earth surface from the Sun to a top of the atmosphere TOA from where it is radiated to outer space, and thus acts as an air conditioner or heat engine [8] keeping the surface temperature constant under radiative forcing from the Sun. A basic question in climate science is the stability of this air conditioner under varying forcing, more specifically the change of surface temperature under doubled concentration of atmospheric CO_2 (from 0.028% to 0.056%) , referred to as *climate sensitivity*.

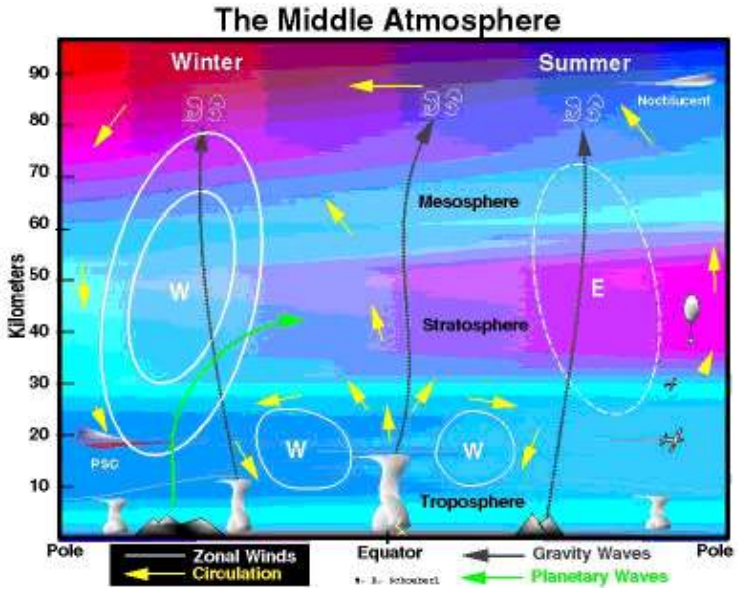


FIGURE 179.1. Thermodynamics of the atmosphere (NASA UARS Project).

The heat is transported by the atmosphere in a combination of thermodynamics (turbulent convection and phase change in evaporation/condensation) and radiation, roughly 2/3 by thermodynamics and 1/3 by radiation. The thermodynamics involves positive radiative forcing balanced by evaporation at low latitudes/altitudes from a warm ocean causing warm air to rise-expand-cool including poleward motion followed by negative radiative forcing balanced by condensation at high latitudes/altitudes causing cool air to descend-contract-warm closing a thermodynamic cycle, as indicated in Fig. 179.1, during polar winter.

179.2 The Illusory Greenhouse Effect

The main message to the World and its leaders from the 2007 IPCC Fourth Assessment Report (AR4) is a prediction of an alarming climate sensitivity in the range $1.5 - 4.5\,C$, with a "best estimate" of $3\,C$, as a result of a so-called *greenhouse effect*.

The physics of this effect is claimed to have been identified and scientifically described by Fourier[3] (1824), Tyndall[10] (1861) and Arrhenius[1] (1896). An inspection of these sources shows a rudimentary analysis based a simple model for radiation without any thermodynamics.

The so-called "greenhouse effect" is described with a double-meaning: It is both the combined total effect of the atmosphere on the Earth sur-

face temperature including both radiation and thermodynamics, and at the same time a hypothetical radiative effect of "greenhouse gases" including CO_2 without thermodynamics. In this way the "greenhouse effect" becomes real, because it is the total effect of the atmopshere and the atmosphere undeniably has an effect, an "atmosphere effect", while at the same time it can be linked to CO_2 apparently acting like a powerful "greenhouse gas" capable of global warming upon a very small increase of 0.028%.

The simplest version of the "greenhouse effect" is described by Stefan-Boltzmann's Law $Q = \sigma T^4$ (SBL), which in differented form

$$dQ = \sigma 4T^3 dT = 4\frac{Q}{T}dT \sim 4dT$$

with $Q \approx 280\,W/m^2$ and $T \approx 288\,K$, gives a climate sensitivty of about $1\,C$ by attributing a certain fictitious additional "radiative forcing" $dQ = 4\,W/m^2$ to doubled CO_2.

Since the total radiative forcing from the Sun is not assumed to change, the additional radiative forcing is supposed to result from a shift of the "characteristic emission level/altitude" to a higher level at lower temperature caused by less radiation escaping to space from lower levels by increasing absorption by CO_2. In this argument the outgoing radiation is connected to a *lapse rate* (decrease of temperature with increasing altitude) supposedly being determined by thermodynamics. With lower "characteristic emission temperature" at higher altitude the whole temperature profile will have to shifted upwards thus causing warming on the ground.

This is the starting point of the CO_2 climate alarmism propagated by IPCC, a basic climate sensitivity of $1\,C$, which then is boosted to $3\,C$ by various so-called (positive) "feed-backs". The basic argument is that since Stefan-Botzmann's Law cannot be disputed as such, and because CO_2 has certain properties of absorption/emission of radiation (light), which can be tested in a laboratory, the starting value of $1\,C$ is an "undeniable physical fact which cannot be disputed".

But wait! Science does not work that way, science only obeys facts and logical mathematical arguments, the essence of the scientific method, and let us now check if the basic postulate of a "greenhouse effect" with basic climate sensitivity of $1\,C$ can qualify as science.

179.3 Mathematical Climate Simulation

The language and methodology of science, in particular climate science, is mathematics: Physical laws are expressed as differential equations of the principal form $D(u) = F$, where F represents *forcing*, u represents the corressponding *system state* coupled to F through a differential operator $D(u)$ acting on u. With given forcing F, the corresponding state u can

be determined by solving the differential equation $D(u) = F$. This is the essence of the scientific method. Note that the differential equation $D(u) = F$ usually describes a cause-effect relation in the sense that the system state u responds to a known given forcing F in a (stable) *forward problem*. This corresponds to putting the horse in front of the wagon, and not the other way around which is referred to as an (unstable) *inverse problem* with the state u given and F the forcing being sought.

Consider now the following approaches to modelling and simulating global climate:

- (A) Thermodynamics with radiative forcing (NSE with SBL forcing).
- (B) Radiation $dQ \sim 4dT$ as differentiated form of (SBL).
- (C) Radiation $dQ \sim 4dT$ combined with thermodynamic lapse rate.
- (D) Radiation $dQ \sim 4dT$ combined with thermodynamic lapse rate and feed-back.

Here (A) is the (stable) forward problem described in the first section and studied below. (B) is self-referential without thermodynamics. (C-D) represent the IPCC approach as an (unstable) inverse problem of radiation with thermodynamic forcing with potentially large positive feed-backs and high climate sensitivity.

Altogether, (A) opens to a rational scientific approach as a stable forward problem, whereas the (C-D) of IPPC represents an unstable inverse problem of questionable value.

In its popular form the basic IPCC climate sensitivity of $1\,C$ is claimed to come from a "greenhouse gas" ability of CO_2 to "trap heat", which is supposed to convince the uneducated. In its more elaborate form intended for the educated, it is connected to a thermodynamic lapse rate and characteristic emission level, in order to account for an effect of additional radiative forcing without change of total radiative forcing. Both forms are severely simplistic and cannot count as science.

To follow (A) we must rid ourselves from the common misconception of thermodynamics expressed in the quote above by Sommerfeld, that it is beyond comprehension for mortals, in particular its 2nd Law. This is the reason why climate scientists have focussed on radiation only, as something understandable, backing away from thermodynamics as something nobody can grasp. But it is possible to give thermodynamics and the 2nd Law a fully understandable meaning as I show in [4, 5] and recall below. This insight opens to a rational approach to climate dynamics, as (A) thermodynamics with radiative forcing.

179.4 Lapse Rate and Global Warming/Cooling

> A theory is the more impressive the greater the simplicity of its premises, the more different kinds of things it relates to, and the more extended its area of applicability. This was therefore the deep impression that classical thermodynamics made upon me. It is the only physical theory of universal content which I am convinced will never be overthrown, within the framework of applicability of its basic concepts. (Einstein)

The effective blackbody temperature of the Earth with atmosphere is $-18\,C$, which can be allocated to a TOA at an altitude of $5\,km$ at a lapse rate of $6.5\,C/km$ connecting TOA to an Earth surface at $15\,C$ with a total warming of $5 \times 6.5 = 33\,C$. The lapse rate determines the surface temperature since the TOA temperature is determined to balance a basically constant insolation. What is then the main factor determining the lapse rate? Is it radiation or thermodynamics, or both?

Climate alarmism as advocated by IPCC is based on the assumption that radiation alone sets an initial lapse rate of $10\,C/km$, which then in reality is moderated by thermodynamics to an observed $6.5\,C/km$. Doubled CO_2 would then increase the initial lapse rate and with further positive thermodynamic feedback it is by IPCC predicted to reach an alarming climate sensitivity or global warming of $3\,C$. Climate alarmism skeptics like Richard Lindzen and Roy Spencer buy the argument of an intial rate of $10\,C/km$ determined by radiation, but suggest that negative thermodynamic feedback effectively reduces climate sensitivity to a harmless $0.5\,C$.

We will argue that an initial lapse rate of $g = 9.81\,C/km$ is instead determined by thermodynamics (and not by radiation) as an equilibrium state without heat transfer, which then in reality by thermodynamic heat transfer (turbulent convection/phase change) is decreased to the observed $6.5\,C/km$, with the heat transfer balancing the radiative heat forcing. More CO_2 would then require more heat transfer by thermodynamics and thus to a further decrease of the lapse rate rather than an increase. The atmopshere would then act like a boiling pot of water which under increased heating would boil more vigorously but not get any warmer.

In short: If thermodynamics is the main mechanism of the atmosphere as an air conditioner or heat transporter, then CO_2 will not cause warming, and IPCC climate alarmism collapses.

We thus identify a basic difference between atmospheric heat transport by radiation (similar to conduction) and by thermodynamics of convection/phase change. In radiation/conduction increased heat transport couples to increased lapse rate (warming). In convection/phase change increased heat transport couples to decreased lapse rate (cooling).

FIGURE 179.2. The atmosphere maintains a constant surface temperature under increasing radiative heat forcing by increasing vaporization and turbulent convection, like a boiling pot of water on a stove.

179.5 Euler Equations for the Atmosphere

> Every mathematician knows it is impossible to understand an elementary course in thermodynamics. (Mathematician V. Arnold)

The viscosity of both water and air is small, while the spatial dimensions of the ocean and atmopshere are large, which means that the Reynolds number $Re = \frac{UL}{\nu}$ is very large ($> 10^8$), where $U > 1\,m/s$ is a typical velocity, $L > 10^3\,m$ a length scale and $\nu < 10^{-5}$ a viscosity. Global climate thus results from turbulent flow at very large Reynolds numbers effectively in the form of turbulent solutions of the *Euler equations* as described in [4].

We focus now on the atmosphere and as a model we consider the Euler equations for a compressible prefect gas occupying a volume Ω representing e.g. the troposphere, here for simplicity without Coriolis force from rotation: Find (ρ, u, T) with ρ density, u velocity and T temperature depending on x and $t > 0$, such that for $x \in \Omega$ and $t > 0$:

$$\begin{aligned} D_u\rho + \rho\nabla\cdot u &= 0,\\ D_u m + m\nabla\cdot u + \nabla p + g\rho e_3 &= 0,\\ D_u T + RT\nabla\cdot u &= q, \end{aligned} \tag{179.1}$$

where $m = \rho u$ is momentum, $p = R\rho T$ is pressure, $R = c_p - c_v$ with c_v and c_p specific heats under constant volume and pressure, and $D_u v = \dot v + u\cdot\nabla v$ is the material time derivative with respect to the velocity u with $\dot v = \frac{\partial v}{\partial t}$ the partial derivative with respect to time t, $e_3 = (0,0,1)$ is the upward direction, g gravitational acceleration and q is a heat source. For air $c_p = 1$ and $\frac{c_p}{c_v} = 1.4$. The Euler equations are complemented by initial values for

ρ, m and T at $t = 0$, and the boundary condition $u \cdot n = 0$ on the boundary of Ω where n is normal to the boundary.

We assume that the heat source q adds heat energy at lower latitudes/altitudes and subtracts heat at higher latitudes/altitudes (radiation to outer space) including evaporation (subtraction of heat) at low altitudes and condensation (addition of heat) at higher altitudes.

We thus consider the full 3D (three-dimensional) Euler/Navier-Stokes equations without any simplification of the vertical flow as in 2D geostrophic flow or in hydrostatic approximation of vertical momentum balance, as a required feature of the next generation of climate models [9] not present in the current generation [2]. This is important because the heat transport involves both horizontal and vertical flow, roughly speaking ascending air at low latitudes and descending air at high latitudes, combined with high altitude poleward flow and low altitude flow towards the Equator.

179.6 The 1st and 2nd Laws of Thermodynamics

> ...no one knows what entropy is, so if you in a debate use this concept, you will always have an advantage. (von Neumann to Shannon)

We recall the 2nd Law of Thermodynamics as stated in [5]:

$$\dot{K} + \dot{P} = W - D, \quad \dot{E} = -W + D + Q, \tag{179.2}$$

where

$$\begin{aligned} K(t) &= \frac{1}{2}\int_\Omega \rho u \cdot u(x,t)\,dx, \quad P(t) = \int_0^t \int_\Omega g\rho u(x,s)\,dxds, \\ E(t) &= \int_\Omega c_v \rho T(x,t)\,dx, \quad W(t) = \int_\Omega p\nabla \cdot u(x,t)\,dx, \\ Q(t) &= \int_\Omega q(x,t)\,dx, \end{aligned} \tag{179.3}$$

is momentary total kinetic energy $K(t)$, potential energy $P(t)$, internal energy $E(t)$ and work rate $W(t)$, and $D(t) \geq 0$ is rate of turbulent dissipation and $Q(t)$ rate of supplied heat or heat forcing. The work W is positive in expansion with $\nabla \cdot u$ positive, and negative in compression with $\nabla \cdot u$ negative (since the pressure p is positive).

Adding the two equations of the 2nd Law, we find that the change of total energy $(K + P + E)$ is balanced by the heat forcing:

$$\frac{d}{dt}(K + P + E) = Q, \tag{179.4}$$

which can be viewed to express the 1st Law of Thermodynamics as conservation of total energy.

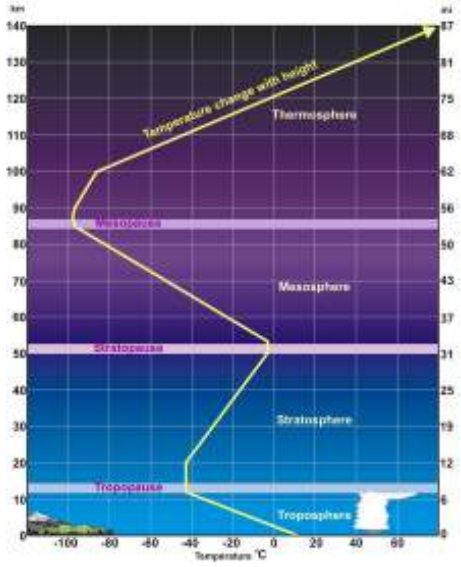


FIGURE 179.3. Temperature profile of the atmosphere, with constant lapse rate in the troposphere of $6.5\,C/km$ (NOAA).

Thermodynamics essentially concerns transformations between heat energy E and the sum $K+P$ of kinetic and potential energies with the transfer being $\pm(W-D)$: whatever $K+P$ gains is lost by E and vice versa. The 2nd Law sets the following limits for these transformations:

- heat energy E can be transformed to kinetic/potential energy $K+P$ only under expansion with $W>0$,
- turbulent dissipation D can transform kinetic/potential energy $K+P$ into heat energy E,
- turbulent dissipation D cannot transform heat energy to kinetic/potential energy, because $D\geq 0$.

179.7 Basic Isothermal and Isentropic Solutions

> As anyone who has taken a course in thermodynamics is well aware, the mathematics used in proving Clausius' theorem (the 2nd Law) is of a very special kind, having only the most tenuous relation to that known to mathematicians. (Mathematician S. Brush)

We identify the following hydrostatic equilibrium base solutions, here fitted to an observed Earth surface temperature of $288\,K$, assuming $Q=0$:

$$\bar{u} = 0,\ \bar{T} = 288 - gx_3,\ \bar{\rho} = \alpha(288 - gx_3)^{\frac{1}{\gamma}},\ \bar{p} = R\alpha(288 - gx_3)^{\frac{1}{\gamma}+1},$$
$$\bar{u} = 0,\ \bar{T} = 288(K),\ \bar{\rho} = \alpha\exp(-gx_3),\ \bar{p} = R\,288\alpha\exp(-gx_3), \tag{179.5}$$

where $\gamma = \frac{R}{c_v}$ $(= 0.4)$ and thus $R(\frac{1}{\gamma} + 1) = c_p = 1$, we scale x_3 in km and α denotes a positive constant to be determined by data.

The first solution is non-turbulent (or isentropic) with $D = 0$ in the 2nd Law:

$$\dot{E} + W = 0, \tag{179.6}$$

or in conventional notation

$$c_v\,dT + p\,dV = 0, \tag{179.7}$$

which combined with hydrostatic balance $\frac{\partial p}{\partial x_3} = -g\rho$ and the differentiated form $pdV + Vdp = RdT$ of the gas law, gives

$$(c_v + R)\frac{\partial T}{\partial x_3} = -g. \tag{179.8}$$

With $c_v + R = c_p = 1000\,J/K\,kg$ the heat capacity of dry air we obtain an isentropic *dry adiabatic lapse rate* of $10\,C/km$. With the double heat capacity of saturated moist air we obtain an isentropic *moist adiabatic lapse* of $5\,C/km$.

The second solution has constant temperature and exponential drop of density and pressure, and can be associated with lots of turbulent dissipation (with $D = W$) effectively equilibrating the temperature.

We summarize the properties of the above base solutions (with $Q = 0$):

- isothermal: maximal turbulent dissipation: $D = W$,
- isentropic: minimal turbulent dissipation: $D = 0$.

We find real solutions between these extreme cases, with roughly $D = \frac{W}{2}$ and $\bar{\rho} \sim (288 - gx_3)^5$, $\bar{p} \sim (288 - gx_3)^6$, with a quicker drop with height than for the isentropic solution with $\bar{\rho} \sim (288 - gx_3)^{2.5}$ and $\bar{p} \sim (288 - gx_3)^{3.5}$, or turned the other way, with a smaller lapse rate of $6.5\,C/km$).

179.8 Basic Thermodynamics

> ...thermodynamics is a dismal swamp of obscurity... a prime example to show that physicists are not exempt from the madness of crowds... Clausius' verbal statement of the second law makes no sense...All that remains is a Mosaic prohibition; a century of philosophers and journalists have acclaimed this commandment; a century

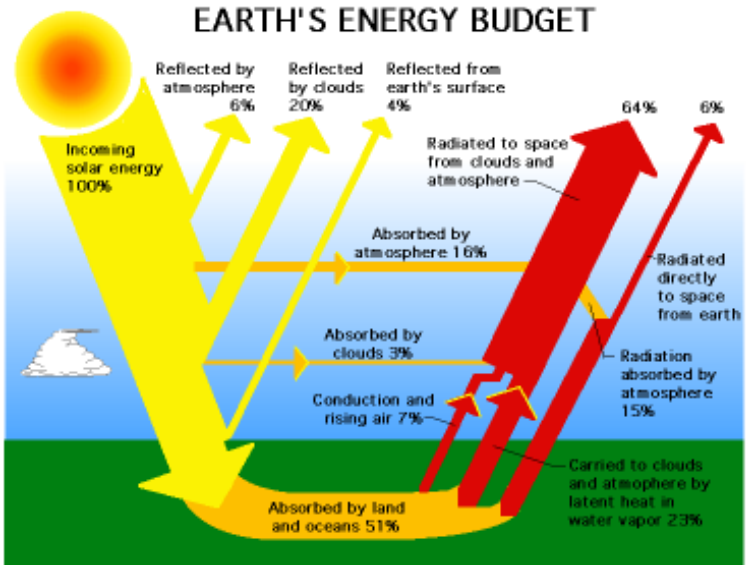


FIGURE 179.4. Earth energy budget (NASA Atmospheric Science Data Center).

> of mathematicians have shuddered and averted their eyes from the unclean...Seven times in the past thirty years have I tried to follow the argument Clausius offers and seven times has it blanked and gravelled me. I cannot explain what I cannot understand. (Physicist C. Truesdell)

We have formulated a basic model of the atmosphere acting as an air conditioner/refrigerator by transporting heat energy from the Earth surface to the top of the atmosphere in a thermodynamic cyclic process with radiation/gravitation forcing, consisting of

- ascending/expanding/cooling air heated by low altitude/latitude radiative forcing,

- descending/compressing/warming air cooled by high altitude/latitude outgoing radiation,

combined with low altitude evaporation and high altitude condensation.

The model is compatible with observation and suggests that the lapse rate/surface temperature is mainly determined by thermodynamics and not by radiation.

The thermodynamics of a standard refrigerator requires a compressor, which in the case of an atmosphere is taken over by gravitation causing compression of descending air.

179.9 Basic Data

> You can fool all the people some time, and some of the people all the time, but you cannot fool all the people all the time. (Abraham Lincoln)

We collect the following observed data, for the first half of the above cycle:

- average upward velocity $= 0.01\,m/s$,
- average density $= 0.6\,kg/m^3$,
- average altitude of TOA $= 5000\,m$,
- $c_p = 1000\,J/K\,kg$
- $Q \approx 180\,W/m^2$ absorbed by the Earth surface with $60\,W$ allocated to radiation, and $120\,W$ to thermodynamics with $100\,W$ to evaporation and $20\,W$ to convection.
- observed lapse rate ≈ -6.5 C/km,
- evaporation $\approx 4\,cm/day$,
- heat of vaporization of water $2200\,kJ/kg$,
- turbulent dissipation rate: $0.002\,W/kg$,

For the upward motion of a column of air over a square meter of surface, we have :

- $\dot{P} \approx 0.01 \times 0.7 \times 5000 \times g = 350\,W$,
- $\dot{E} \approx -0.01 \times 0.7 \times 1000 \times 5000 \times \frac{6.5}{1000} \approx -230\,W$,
- phase change: $2.2 \times 10^6 \times 10^2 \times \frac{0.04}{24\times 3600} \approx 100\,W$,

which is compatible with $W - D = \dot{P} = 350\,W$ and $\dot{E} = -W + D + Q = -230\,W$.

The observed lapse rate of $6.5\,C/km$ can be viewed as being obtained by moderating the dry adiabatic rate of $10\,C/km$ by a combined process of phase change and turbulent dissipation effectively reducing the drop of temperature with altitude. The energy transfer in this process ($\approx \frac{3.5}{6.5} \times 230 = 120\,W$ with $100 - 110\,W$ for evaporation and $20 = 0.002 \times 5000 \approx 10 - 20\,W$ for turbulence) is roughly equal to the heat forcing allocated to thermodynamics ($= 120\,W$). Increasing heat transfer then corresponds to non-increasing lapse rate and non warming; the main message of our analysis.

The observed lapse rate of $6.5\,C/km$ is bigger than the moist adiabatic rate of $5\,C/km$, which causes unstable overturning of rising warm air and turbulent dissipation.

179.10 Lapse Rate vs Radiative Forcing

If the lapse rate is L then $\dot{P} + \dot{E} = Q$ combined with $\dot{E}/\dot{P} = \frac{L}{10}$ according to the above computation, gives $L = 10(1 - Q/\dot{P})$. If Q is increased then L will decrease if $\dot{P}$ stays constant, but if $\dot{P}$ increases quicker than Q, then L may increase. Increasing Q may be expected to give an increase of $\dot{P}$ by increasing the vertical convection velocity, but a decrease by increasing phase change evaporation/condensation. Which effect will dominate: convection or phase change? Computations with an answer are under way... until then we notice that out of $120\,W/m^2$ of radiative heat forcing, a major part of say 100 can be allocated to phase change, which gives phase change a good chance to compete with convection...

179.11 Summary: Atmosphere as Air Conditioner

> A good many times I have been present at gatherings of people who, by the standards of the traditional culture, are thought highly educated and who have with considerable gusto been expressing their incredulity at the illiteracy of scientists. Once or twice I have been provoked and have asked the company how many of them could describe the Second Law of Thermodynamics. The response was cold: it was also negative. (C. P. Snow in 1959 Rede Lecture entitled The Two Cultures and the Scientific Revolution).

Let us now sum up the experience from our analysis. We have seen that the atmosphere acts as a thermodynamic air conditioner transporting heat energy from the Earth surface to a TOA under radiative heat forcing. We start from an isentropic stable equilibrium state with lapse rate $9.8\,C/km$ with zero heat forcing and discover the following scenario for the response of the air conditioner under increasing heat forcing:

1. increased heat forcing of the Ocean surface at low latitudes is balanced by increased vaporization,
2. increased vaporization increases the heat capacity which decreases the moist adiabatic lapse rate,
3. if the actual lapse rate is bigger than the actual moist adiabatic rate, then unstable convective overturning is triggered,
4. unstable overturning causes turbulent convection with increased heat transfer.

The atmospheric air conditioner thus may respond to increased heat forcing by (i) increased vaporization decreasing the moist adiabatic lapse rate combined with (ii) increased turbulent convection if the actual lapse rate

is bigger than the moist adiabatic lapse rate. This is how a boiling pot of water reacts to increased heating.

> If someone points out to you that your pet theory of the universe is in disagreement with Maxwell's equations, then so much the worse for Maxwell's equations. If it is found to be contradicted by observation, well, these experimentalists do bungle things sometimes. But if your theory is found to be against the second law of thermodynamics, I can give you no hope; there is nothing for it but to collapse in deepest humiliation (Sir Arthur Stanley Eddington in The Nature of the Physical World, 1915)

References

[1] S. Arrhenius, On the Influence of Carbonic Acid in the Air upon the Temperature of the Ground, Philosophical Magazine and Journal of Science Series 5, Volume 41, April 1896, pages 237-276.

[2] D. Frierson, Climate Models, http://courses.washington.edu/pcc587/notes/587_9.pdf.

[3] J. Fourier, General remarks on the temperature of the earth and outer space. American Journal of Science. 32, 1-20 (1837) by Ebeneser Burgess. English translation of "Remarques générales sur les températures du globe terrestre et des espaces planétaires." Annales de Chimie et de Physique. (Paris) 2nd ser., 27, 136-67 (1824), by Jean-Baptiste Joseph Fourier.

[4] J. Hoffman and C. Johnson, *Computational Turbulent Incompressible Flow*, Springer, 2007.

[5] J. Hoffman and C. Johnson, *Computational Thermodynamics*, http://www.nada.kth.se/ cgjoh/ambsthermo.pdf.

[6] C. Johnson, Mathematical Simulation Technology, http://www.nada.kth.se/ cgjoh/preview/bodysoul.pdf.

[7] C. Johnson, Computational Blackbody Radiation, this book.

[8] H. Osawa, A. Ohmura, R. D. Lorenz, T. Pujol, The Second Law of Thermodynamics and the Gobal Climate System: A Review of the Maximum Entropy Production Principle, Reviews of Geophysics, 41, 4 1018, 2003.

[9] J. Slingo et al, Developing the next generation climate systems models: challenges and achievements, Phil. Trans. R. Soc. A 2009 367, 815-831, doi: 10.1098/rsta.2008.0207.

[10] J. Tyndall, On the Absorption and Radiation of Heat by Gases and Vapours, and on the Physical Connexion of Radiation, Absorption, Conduction.-The Bakerian Lecture, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 4, Vol. 22, pp. 169-194, 273-285, 1861.

180
Cosmology

180.1 To Watch

- From 20 million to 14 billion
- The Millennium Simulation
- How the Milky Wave will end
- How large is the Universe?

180.2 Simulator

Simulate the dynamics of a compressible gas subject to gravity forces by the compressible Navier-Stokes equations.

180.3 Investigation

Study how if fluctuations in an initial smooth mass distribution can develop into non-smooth mass concentrations representing sparsely distributed galaxies.

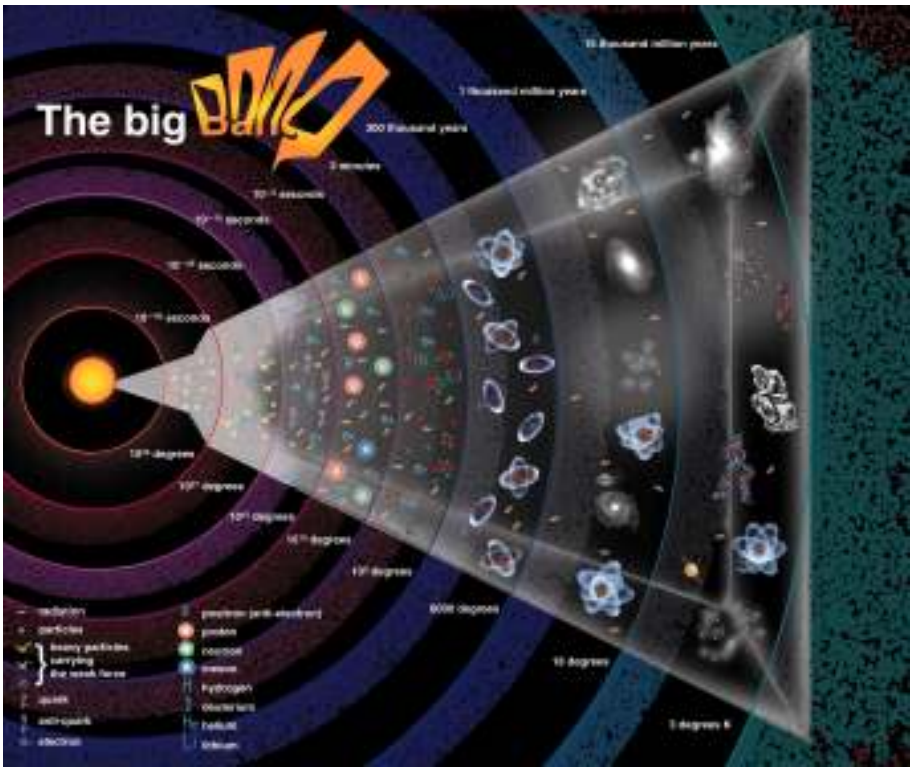


FIGURE 180.1. Big Bang.

181
Quantum Mechanics

> There is a computer disease that anybody who works with computers knows about. It's a very serious disease and it interferes completely with the work. The trouble with computers is that you 'play' with them! (Richard Feynman)

181.1 To Read

- Many-Minds Quantum Mechanics

181.2 Simulator

Construct a simulator based on Many-Minds solutions of Schrödinger's equation.

181.3 Investigation

Determine the ground states of Helium, Lithium and Beryllium, and check with the literature and experiments.

182
Digital Photography

182.1 Digital Images

Photography is being revolutionized by digital technology for digital imaging. Digital photography is one of several forms of digital imaging. Digital images are also created by non-photographic equipment such as computer tomography scanners and radio telescopes. Digital images can also be made by scanning conventional photographic images.

182.2 Digital Image Processing

Digital image processing is the use of mathematical computer algorithms to perform image processing on digital images, such as

- Compression-wavelets
- Linear filtering
- Principal components analysis
- PDE-Anisotropic diffusion: softening, sharpening, despeckling,...

182.3 To Read

- Tutorial for Image Processing

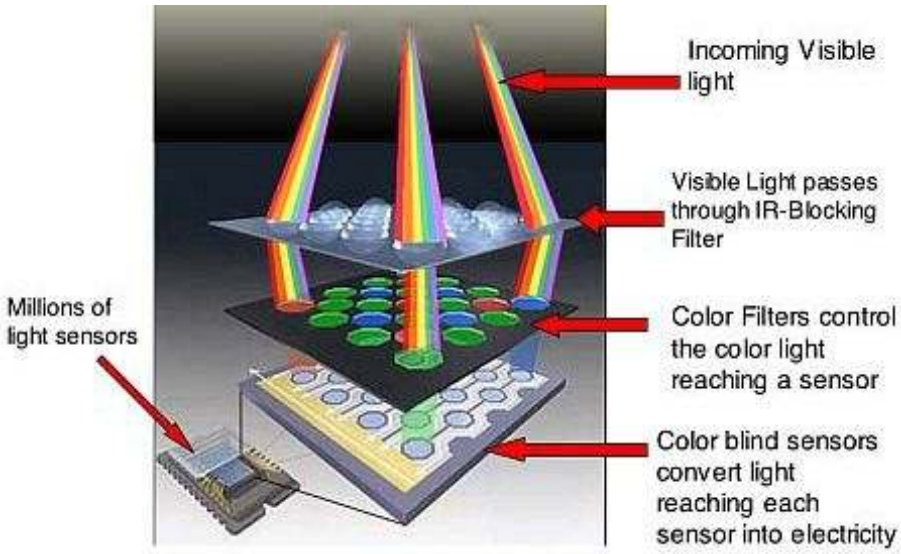


FIGURE 182.1. CCD Charge-Coupled Device sensor.

- Video Lectures
- CCD Sensors

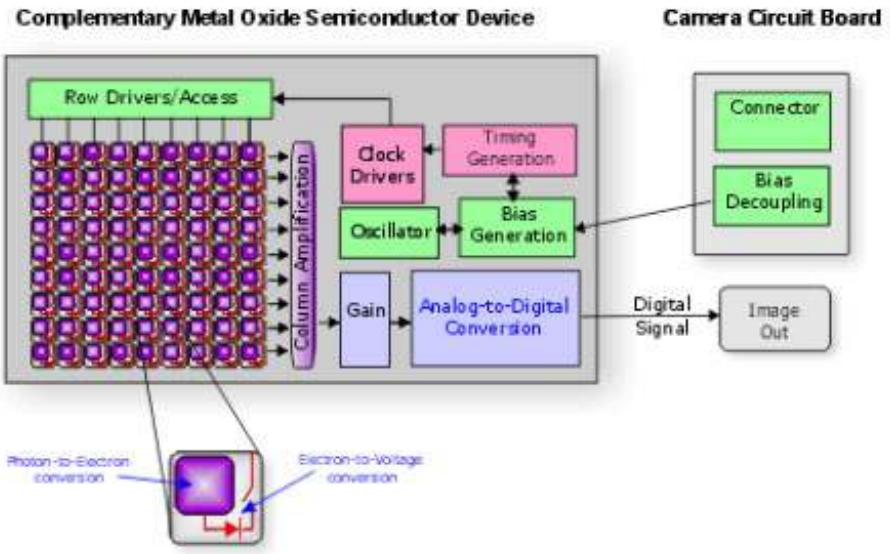


FIGURE 182.2. CCD sensor.

Part XII

1D Calculus

(Partly from Applied Mathematics Body and Soul, Vol 1-2, Springer 2003, coauthored with Kenneth Eriksson and Don Estep).

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

$$u(x_N) - u(x_0) = \int_{x_0}^{x_N} u'(x)\,dx$$
$$\approx \sum_{j=1}^{N} u'(x_{j-1})(x_j - x_{j-1})$$

183
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)

183.1 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.

183.2 The Natural Numbers

The *natural numbers* such as $1, 2, 3, 4, ...$, 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*

$$m \times n = n \times m \tag{183.1}$$

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. 183.1.

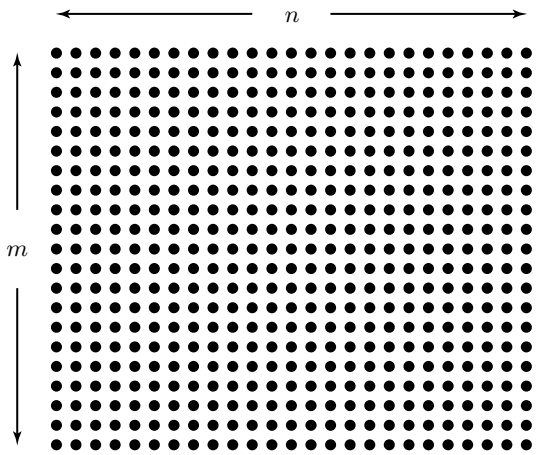


FIGURE 183.1. Illustration of the commutative rule for multiplication $m \times n = n \times m$. We get the same sum if first add up the dots by counting across the rows or down the columns.

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

$$m \times (n \times p) = (m \times n) \times p \tag{183.2}$$

and the *distributive rule* combining addition and multiplication,

$$m \times (n + p) = m \times n + m \times p, \tag{183.3}$$

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 (183.3) (and (183.1)) we obtain the useful formula

$$(m + n)(p + q) = (m + n)p + (m + n)q = mp + np + mq + nq. \tag{183.4}$$

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

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

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

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

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

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

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. 183.2. 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.

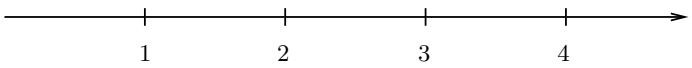


FIGURE 183.2. The natural number line.

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

$$0 + n = n + 0 = n \tag{183.5}$$

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

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

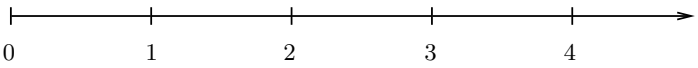


FIGURE 183.3. The extended natural number line, including 0.

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

$$d \times 10^p \tag{183.6}$$

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

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

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

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

$$1001 = 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0, \tag{183.7}$$

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

183.3 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.

183.4 The Set $\mathbb{N}$ of All Natural Numbers

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

We can think of $\mathbb{N}$ as the *set of possible natural numbers* and it is clear that this set is always under construction and can never actually be completed. It is like a 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.

183.5 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. 183.4.

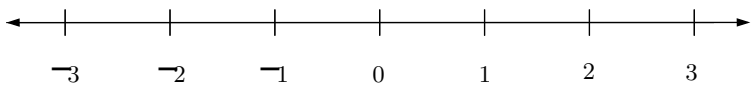


FIGURE 183.4. The integer number line.

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

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

$n - m = -m + n = n + (-m)$, which we refer to as *subtracting* m from n. We see that subtracting m from n is the same as adding $-m$ to n.

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

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

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

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

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

183.6 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 183.1. 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

$$|x-79| \leq 5. \tag{183.8}$$

There are two possible cases: $x < 79$ and $x \geq 79$. If $x \geq 79$ then $|x-79| = x-79$ and (183.8) becomes $x-79 \leq 5$ or $x \leq 84$. If $x < 79$ then $|x-79| = -(x-79)$ and (183.8) 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 183.2. $|x-79| \le 5$ means that

$$\begin{array}{ccccc} -5 & \le & x-79 & \le & 5 \\ 74 & \le & x & \le & 84 \end{array}$$

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

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

EXAMPLE 183.3. To solve the following inequality in x:

$$|x-79| \ge 5. \tag{183.9}$$

we first assume that $x \ge 79$, in which case (183.9) becomes $x-79 \ge 5$ or $x \ge 84$. Next, if $x \le 79$ then (183.9) becomes $-(x-79) \ge 5$ or $(x-79) \le -5$ or $x \le -74$. The answer is thus all x with $x \ge 84$ or $x \le -74$.

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

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

183.7 Division with Remainder

We define *division with remainder* of a natural number n by another natural number m, as the process of computing nonnegative integers p and $r < m$ such that $n = pm+r$. The existence of unique p and r follows by considering the sequence of natural numbers m, $2m$, $3m$,..., and noting that there must be a unique p such that $pm \le n < (p+1)m$, see Fig. 183.5.

$$m = 5 \text{ and } n = pm + r \text{ with } r = 2 < m$$

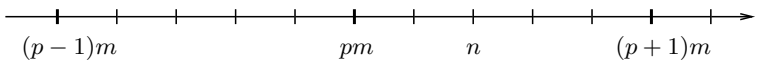


FIGURE 183.5. Illustration of $pm \le n < (p+1)m$.

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

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

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

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

A more systematic procedure for division with remainder is the *long division* algorithm, which is taught in school. We give two examples ($63 = 4 \times 15 + 3$ again, and $2418610 = 19044 \times 127 + 22$) in Fig. 183.5.

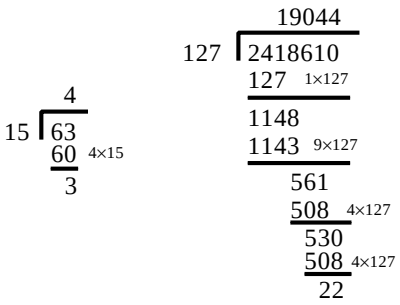


FIGURE 183.6. Two examples of long division.

183.8 Factorization into Prime Factors

A *factor* of a natural number n is a natural number m that divides into n without leaving a remainder, that is, $n = pm$ for some natural number p. For example, 2 and 3 are both factors of 6. A natural number n always has factors 1 and n since $1 \times n = n$. A natural number n is called a *prime number* if the only factors of n are 1 and n. The first few prime numbers (excluding 1 since such factors are not of much interest) are $\{2, 3, 5, 7, 11, \cdots\}$. The only even prime number is 2. Suppose that we take the natural number n and try to find two factors $n = pq$. Now there are two possibilities: either the only two factors are 1 and n, i.e. n is prime, or we find two factors p and q, neither of which are 1 or n. By the way, it is easy to write a program to search for all the factors of a given natural number n by systematically dividing by all the natural numbers up to n. Now in the second case, both p and q must be less than n. In fact $p \leq n/2$ and $q \leq n/2$ since the smallest possible factor not equal to 1 is 2. Now we repeat by factoring p and q separately. In each case, we either find the number is

prime or we factor it into a product of smaller natural numbers. Then we continue with the smaller factors. Eventually this process must stop since n is finite and the factors at any stage are no larger than half the size of the factors of the previous stage. When the process has stopped, we have *factored* n into a product of prime numbers. This factorization is unique except for order. One consequence of the factorization into prime numbers is the following fact. Suppose that we know that 2 is a factor of n. If $n = pq$ is any factorization of n, it follows that at least one of the factors p and q must have a factor of 2. The same is true for prime number factors 3, 5, 7 etc., that is for any prime number factor.

183.9 Computer Representation of Integers

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

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

Chapter 183 Problems

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

183.2. 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:

183.3. 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 $(m \times n) \times p = m \times (n \times p)$..

183.4. 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.

183.5. 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 $p = 0$?

183.6. Verify using (183.4) that for integers n and m,

$$\begin{aligned} (n+m)^2 &= n^2 + 2nm + m^2 \\ (n+m)^3 &= n^3 + 3n^2m + 3nm^2 + m^3 \\ (n+m)(n-m) &= n^2 - m^2. \end{aligned} \tag{183.10}$$

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

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

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

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

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

183.11. (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.

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

183.13. 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?

183.14. 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.

183.15. Solve the following inequalities:

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

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

183.16. 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) $|a+b| \le |a| + |b|$ (e) $|a-b| \le |a|+|b|$ (f) $|a+b-c| \le |a|+|b|+|c|$ (g) $|a| \le |a-b|+|b|$ (h) $||a|-|b|| \le |a-b|$

183.17. Show that the inequalities (e)-(h) of Problem 183.17 follow once you have (d) and the fact that $|a| = |-a|$ for any integer a.

183.18. 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.

184
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)

184.1 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.

184.2 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

$$xy = \frac{p}{q} \times \frac{r}{s} = \frac{pr}{qs} \quad \text{or} \quad (p, q) \times (r, s) = (pr, qs), \tag{184.1}$$

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

$$x + y = \frac{p}{q} + \frac{r}{s} = \frac{ps + qr}{qs} \quad \text{or} \quad (p, q) + (r, s) = (ps + qr, qs). \tag{184.2}$$

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 183.5). Since the two equations $qx = p$ and $sqx = sp$ have the same solution x,

$$\frac{p}{q} = x = \frac{sp}{sq} \quad \text{or} \quad (p, q) = (sp, sq). \tag{184.3}$$

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 (184.3), we *define* $(p, q) = (sp, sq)$ for $s \neq 0$, thus considering (p, q) and (sp, sq) to be (two representatives of) one and the same rational number. For example, $\frac{6}{4} = \frac{3}{2}$.

We next *define* the operations of multiplication $\times$ and addition $+$ of rational numbers by (184.1) and (184.2). 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

$$|r| = \begin{cases} r & \text{if } r \geq 0, \\ -r & \text{if } r < 0. \end{cases} \tag{184.4}$$

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.

184.3 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.

184.4 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

$$\pm\, p_m p_{m-1} \cdots p_2 p_1 p_0 . q_1 q_2 \cdots q_n, \tag{184.5}$$

where the *digits* p_m, p_{m-1}, .., p_0, q_0, .., q_n are each equal to one of the natural numbers $\{0, 1, \cdots, 9\}$ while m and n are natural numbers. The decimal expansion (184.5) is a shorthand notation for the number

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

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 (184.5) 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. 184.1.

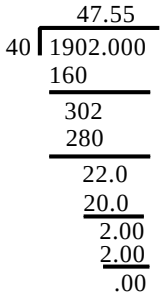


FIGURE 184.1. Using long division to obtain a decimal expansion.

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 $432.576 = 432576/10^3$.

184.5 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. 184.2. The word "infinite" is here to indicate that the decimal expansion continues without ever stopping. We can find many examples of infinite decimal

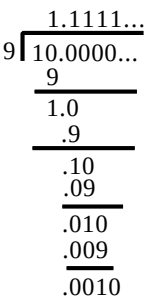


FIGURE 184.2. The decimal expansion of 10/9 never stops.

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:

$$\frac{1043}{439} = 2.37585421412300683371298405466970387243735763097$$
$$94988610478359908883826879271070615034168564920$$
$$27334851936218678815489749430523917995444191343$$
$$96355353075170842824601366742596810933940774487$$
$$47152619589977220956719817767 \quad 37585421412300$$
$$68337129840546697038724373576309794988610478359$$
$$90888382687927107061503416856492027334851936218$$
$$67881548974943052391799544419134396355353075170$$
$$84282460136674259681093394077448747152619589977$$
$$22095671981776765 \cdots$$

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 $1.1\cdots1_n$ (that is with n decimals equal to 1 after the point), we have

$$1.11\cdots11_n = 1 + 10^{-1} + 10^{-2} + \cdots + 10^{-n+1} + 10^{-n}.$$

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

$$1.11\cdots11_n = \frac{1-10^{-n-1}}{1-0.1} = \frac{10}{9}\,(1-10^{-n-1}), \tag{184.6}$$

and thus

$$\frac{10}{9} = 1.11\cdots11_n + \frac{10^{-n}}{9}. \tag{184.7}$$

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 $1.11\cdots11_n$ 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 $1.111\cdots1_n$ as close as we like to $10/9$ by taking n large. In particular, we would have

$$|\frac{10}{9} - 1.11\cdots11_n| \leq 10^{-n}.$$

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

$$\begin{aligned}
.1818\cdots18_m &= \frac{18}{100} + \frac{18}{10000} + \frac{18}{1000000} + \cdots + \frac{18}{10^{2m}} \\
&= \frac{18}{100}\left(1 + \frac{1}{100} + \frac{1}{100^2} + \cdots + \frac{1}{100^{m-1}}\right) \\
&= \frac{18}{100}\,\frac{1-(100^{-1})^m}{1-100^{-1}} = \frac{18}{100}\,\frac{100}{99}(1-100^{-m}) \\
&= \frac{2}{11}(1-100^{-m}).
\end{aligned}$$

that is

$$\frac{2}{11} = 0.1818\cdots18_m + \frac{2}{11}100^{-m},$$

so that

$$|\frac{2}{11} - 0.1818\cdots18_m| \leq 100^{-m}.$$

We thus interpret $2/11 = .1818181818\cdots$ as meaning that we can make the numbers $.1818\cdots18_m$ 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 (**??**), 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

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

We conclude that we may interpret

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

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

EXAMPLE 184.1. $0.123123123\cdots$ 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 184.1 *The decimal expansion of a rational number is periodic. A periodic decimal expansion is equal to a rational number.*

184.6 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. 184.3

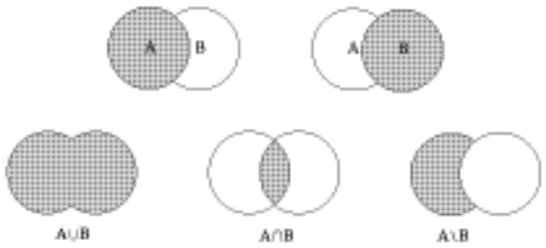


FIGURE 184.3. The sets $A \cup B$, $A \cap B$ and $A \backslash B$.

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 184.2. 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 $A \times B = \{(1, 3), (1, 4), (2, 3), (2, 4), (3, 3), (3, 4)\}$.

184.7 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.

184.8 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. 184.4.

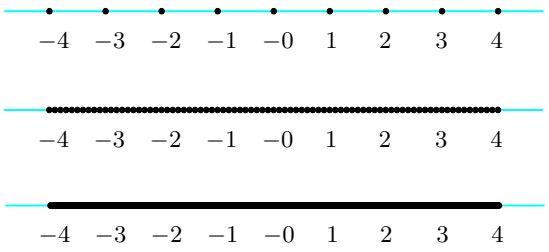


FIGURE 184.4. Filling in the rational number line between -4 and 4 starting with integers, rationals with one digit, and rationals with two digits, and so on.

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

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

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

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

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

We write the set of numbers less than a in a similar way. We also represent intervals graphically by marking the points on the rational line segment, as we show in Fig. 184.5. Note how we use an open circle or a closed circle to mark the endpoints of open and closed intervals.

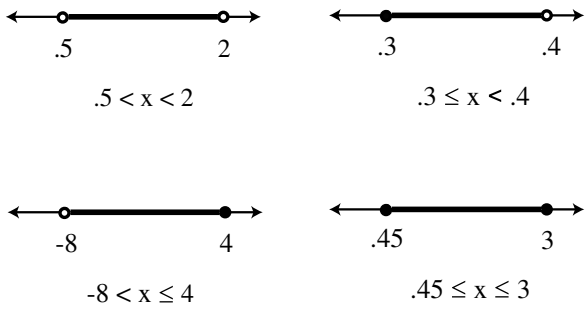


FIGURE 184.5. Various rational line intervals.

184.9 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

$$P_n = 2P_{n-1} \tag{184.8}$$

for $n \geq 1$. This model is similar to the model (**??**) we used to describe the insect population in Model **??**. 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 (184.8) 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

$$P_n = \frac{2}{1 + P_{n-1}/K}\, P_{n-1}. \tag{184.9}$$

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 (184.9) 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 (184.9) implies (check this!) that

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

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

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

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

$$P_n = \frac{1}{Q_n} = \frac{1}{\frac{1}{2^n}Q_0 + \frac{1}{K}\left(1 - \frac{1}{2^n}\right)}. \tag{184.10}$$

184.10 Chemical Equilibrium

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

$$\mathrm{A}_x\,\mathrm{B}_y \rightleftharpoons x\,\mathrm{A}^{y+} + y\,\mathrm{B}^{x-} \tag{184.11}$$

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

$$K_{sp} = [\,\mathrm{A}^{y+}]^x[\,\mathrm{B}^{x-}]^y. \tag{184.12}$$

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 (184.12), we get the equation

$$S\,(.02 + 2S)^2 = 1.57 \times 10^{-9}. \tag{184.13}$$

Chapter 184 Problems

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

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

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

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

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

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

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

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

$$|s-t| \leq |s| + |t|, \tag{184.14}$$

$$|s-t| \leq |s-u| + |t-u|, \tag{184.15}$$

and

$$|st| = |s|\,|t|. \tag{184.16}$$

184.7. 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.

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

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

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

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

184.12. 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

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

184.13. 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.

185
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)

185.1 Introduction

The concept of a *function* is fundamental in mathematics. We already met this concept in the context of the Dinner Soup model, where the total cost was $15x$ (dollars) if the amount of beef was x (pounds). For every amount of beef x, there is a corresponding total cost $15x$. We say that the total cost $15x$ is a function of, or depends on, the amount of beef x.

The term function and the mathematical notation we use today was introduced by Leibniz (1646-1716), who said that $f(x)$, which reads "f of x", is a *function* of x if for each value of x in some prescribed set of values over which x can vary, there is assigned a unique value $f(x)$. In the Dinner Soup model $f(x) = 15x$. It is helpful to think of x as the *input*, while $f(x)$ is the corresponding *output*, so that as the value of x varies, the value of $f(x)$ varies according to the assignment. Correspondingly, we often write $x \to f(x)$ to signify that x is mapped onto $f(x)$. We also think of the function f as a "machine" that transforms x into $f(x)$:

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

see also Fig. 185.1.

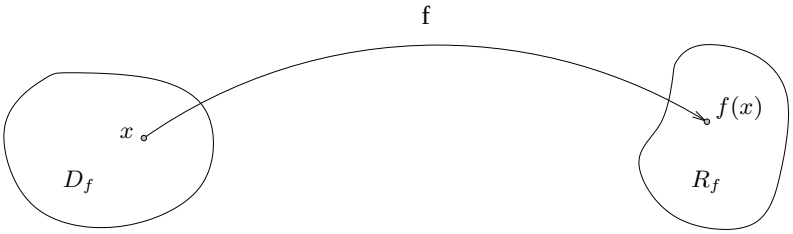


FIGURE 185.1. Illustration of $f : D_f \to R_f$.

We refer to x as a *variable* since x can take different values, and x is also called the *argument* of the function. The prescribed set of values over which x can vary is called the *domain* of the function f and is denoted by $D(f)$. The set of values $f(x)$ corresponding to the values of x in the domain $D(f)$, is called the *range* $R(f)$ of $f(x)$. As x varies over the domain $D(f)$, the corresponding function value $f(x)$ varies over $R(f)$. We often write this symbolically as $f : D(f) \to R(f)$ indicating that for each $x \in D(f)$ there is a value $f(x) \in R(f)$ assigned.

In the context of the Dinner Soup model with $f(x) = 15x$, we may choose $D(f) = [0, 1]$, if we decide that the amount of beef x can vary in the interval $[0, 1]$, in which case $R(f) = [0, 15]$. For each amount x of beef in the interval $[0, 1]$, there is a corresponding total cost $f(x) = 15x$ in the interval $[0, 15]$. Again: the total cost $15x$ is a function of the amount of beef x. We may also choose the domain $D(f)$ to be some other set of possible values of the amount of beef x such as $D(f) = [a, b]$, where a and b are positive rational numbers, with the corresponding range $R(f) = [15a, 15b]$, or $D(f) = \mathbb{Q}^+$ with the corresponding range $R(f) = \mathbb{Q}^+$, where $\mathbb{Q}^+$ is the set of positive rational numbers. We may even consider the function $x \to f(x) = 15x$ with $D(f) = \mathbb{Q}$ and the corresponding range $R(f) = \mathbb{Q}$, which would lead

outside the Dinner Soup model since there x is non-negative. For a given assignment $x \to f(x)$, that is, a given function $f(x)$, we may thus associate different domains $D(f)$ and corresponding ranges $R(f)$ depending on the setting.

It is common to assign a variable name to the output of a function, for example we may write $y = f(x)$. Thus, the value of the variable y is given by the value $f(x)$ assigned to the variable x. We therefore call x the *independent variable* and y the *dependent variable*. The independent variable x takes on values in the domain $D(f)$, while the dependent variable y takes on values in the range $R(f)$.

Note that the names we use for the independent variable and the dependent variable for a given function f can be changed. The names x and y are common, but there is nothing special about these letters. For example, $z = f(u)$ denotes the same function if we do not change f, i.e. the function $y = 15x$ can just as well be written $z = 15u$. In both cases, to a given number x or u the function f assigns that number multiplied by 15, that is $15x$ or $15u$. Thus we refer to "the function $f(x)$" while in fact it would be more correct to just say "the function f", because f is the "name" of the function, while $f(x)$ is more like a description or definition of the function. Nevertheless we will often use the somewhat sloppy language "the function $f(x)$" because it identifies both the name of the function and its definition/description.

EXAMPLE 185.1. The function $x \to f(x) = x^2$, or in short the function $f(x) = x^2$, may be considered with domain $D(f) = \mathbb{Q}^+$ and range $R(f) = \mathbb{Q}^+$, but also with domain $D(f) = \mathbb{Q}$ and again $R(f) = \mathbb{Q}^+$, or with $D(f) = \mathbb{Z}$, and $R(f) = \{0, \pm 1, \pm 2, \pm 4, ...\}$. We illustrate in Fig. 185.2.

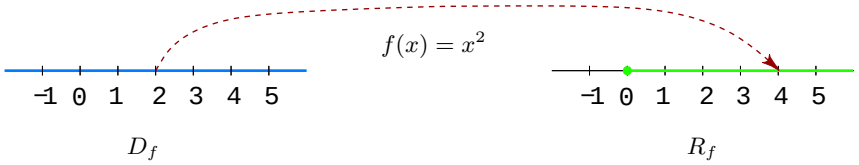


FIGURE 185.2. Illustration of $f : \mathbb{Q} \to \mathbb{Q}^+$ with $f(x) = x^2$.

EXAMPLE 185.2. For the function $f(z) = z + 3$ we may choose, for example, $D(f) = \mathbb{N}$ and $R(f) = \{4, 5, 6, ...\}$, or $D(f) = \mathbb{Z}$ and $R(f) = \mathbb{Z}$.

EXAMPLE 185.3. We may consider the function $f(n) = 2^{-n}$ with $D(f) = \mathbb{N}$ and $R(f) = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, ...\}$.

EXAMPLE 185.4. 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 185.5. 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. 185.2.

EXAMPLE 185.6. 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 $f : D(f) \to \mathbb{Q}$.

185.2 Functions in Daily Life

In daily life, we stumble over functions right and left. A car dealer assigns a price $f(x)$, which is a number, to each car x in his lot. Here $D(f)$ may

be a set of numbers if each car is identified by a number, or $D(f)$ may be some other listing of the cars such as {Chevy85blue, Olds93pink,...}, and the range $R(f)$ is the set of all different prices of the cars in $D(f)$. When the government makes out our tax bill, it is assigning one number $f(x)$, representing the amount we owe, to another number x, representing our salary. Both the domain $D(f)$ and the range $R(f)$ in this example change a lot depending on the political winds.

Any quantity which varies over time may be viewed as a function of time. The daily maximum temperature in degrees Celsius in Stockholm during 1999 is a certain function $f(x)$ of the day x of the year, with $D(f) = \{1, 2, ..., 365\}$ and $R(f)$ normally a subset of $[-30, 30]$. The price $f(x)$ of a stock during one day of trade at the Stockholm Stock Exchange is a function of the time x of the day, with $D(f) = [10.00, 17.00]$ and $R(f)$ the range of variation of the stock price during the day. The length of women's skirts varies over the years around the level of the knee, and is supposed to be a good indicator of the variation of the economical climate. The length of a human being varies over the life time, and the thickness of the ozone layer over years.

We may also simultaneously consider several quantities depending on time, like for example the temperature $t(x)$ in degrees Celsius and wind velocity $w(x)$ in meter per second in Chicago as functions of time x, where x ranges over the month of January, and we may combine the two values $t(x)$ and $w(x)$ into a pair of numbers "$t(x)$ and $w(x)$", which we may write as $f(x) =$ "$t(x)$ and $w(x)$" or in short-hand $f(x) = (t(x), w(x))$ with the parenthesis enclosing the pair. For example writing, $f(10) = (-30, 20)$, would give the information that the 10th of January was a tough day with temperature $-30°C$ and wind 20 meter per second. From this information we could compute the adjusted temperature -50$°C$ that day taking the wind factor into account.

Likewise the input variable x could represent a pair of numbers, like a temperature and a wind speed and the output could be the adjusted temperature with the wind factor taken into account (Find the formula!).

We conclude that the input x of a function $f(x)$ may be of many different types, single numbers, pairs of numbers, triples of numbers, et cetera, as well as the output $f(x)$.

EXAMPLE 185.7. 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 185.8. A movie consists of a sequence of pictures that are displayed at the rate of 16 pictures per second. We usually watch a movie from the first to the last picture. Afterwards we might talk about different scenes in the movie, which corresponds to subsets of the totality of pictures. A very few people, like the film editor and director,

might consider the movie on the level of the individual elements in the domain, that is the pictures on the film. When editing the movie, they number the picture frames $1, 2, 3, \cdots, N$ where $1, 2, ...16$ are the numbers of the pictures displayed sequentially during the first second, and N is the number of the last picture. We may then consider the movie as a function $f(n)$ with $D(f) = \{1, 2, \cdots, N\}$, which to each number n in $D(f)$ associates the picture frame with number n.

EXAMPLE 185.9. 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 185.10. 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. 185.3. 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 $x \to f(x)$ with a certain domain $D(f)$ and range $R(f)$, namely, the *ordering* of $D(f)$. The pages of a book, and pictures of a film are numbered consecutively, and the domain of a directory is also ordered alphabetically. In the case of a book or film the ordering helps to make sense out of the material, and a dictionary without any order is almost useless. Of course, swapping through films has become a part of the life-style of to-day, but the risk of a loss of understanding is obvious. To be able to catch the main

FIGURE 185.3. Hermann Hollerith, inventor of the punched card machine: "My friend Dr. Billings one night at the pub suggested to me that there ought to be some mechanical way of doing the census, something on the principle of the Jacquard loom, whereby holes in a card regulate the pattern to be woven".

idea or plot of a book or film *as a whole* it is necessary to read the pages or view the pictures in order. The ordering helps us to get an overall meaning.

Similarly, it is useful to be able to catch the main properties of a function $f(x)$, and this can sometimes be done by graphing or visualizing the function using some suitable ordering of $D(f)$. We now go into the topic of graphing functions $f(x)$ with $D(f)$ and $R(f)$ subsets of $\mathbb{Q}$, of course with the usual ordering of $\mathbb{Q}$, and with the purpose of trying to grasp the nature of a given function "as a whole".

185.3 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. 185.4. 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. 185.4. This is called the *integer coordinate plane*. Each number line is called an *axis* of

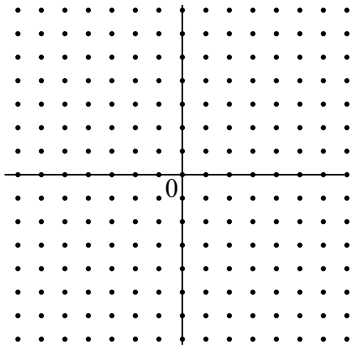


FIGURE 185.4. The integer coordinate plane.

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

As we saw, a function $f : \mathbb{Z} \to \mathbb{Z}$ can be represented by making a list with the inputs placed side-by-side with the corresponding outputs. We show such a table for $f(n) = n^2$ in Fig. 185.5. 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. 185.5.

EXAMPLE 185.11. In Fig. 185.6, we plot n, n^2, and 2^n along the vertical axis with $n = 1, 2, 3, .., 6$ along the horizontal axis. The plot suggests 2^n grows more quickly than both n and n^2 as n increases. In

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

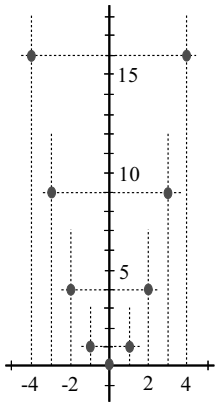


FIGURE 185.5. A tabular listing of $f(n) = n^2$ and a graph of the points associated with the function $f(n) = n^2$ with domain equal to the integers.

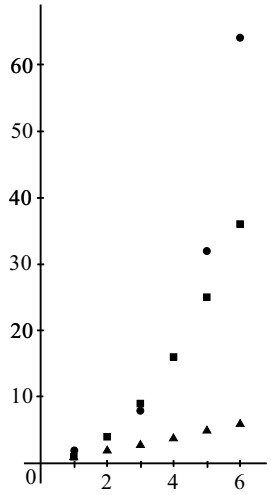


FIGURE 185.6. Plots of the functions ▲ $f(n) = n$, ■ $f(n) = n^2$, and • $f(n) = 2^n$ with $D(f) = \mathbb{N}$.

Fig. 185.7, 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. 185.6.

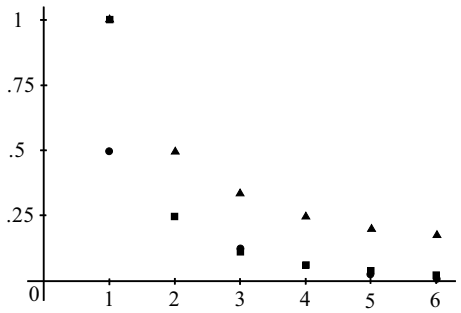


FIGURE 185.7. Plots of the functions ▲ $f(n) = n^{-1}$, ■ $f(n) = n^{-2}$, and ● $f(n) = 2^{-n}$ with $D(f) = \mathbb{N}$.

Instead of using a table to list the points for a function, we can represent a point on the integer plane mathematically by means of an *ordered pair* of numbers. To the point in the plane located at the intersection of the vertical line passing through n on the horizontal axis and the horizontal line passing through m on the vertical axis, we associate the pair of numbers (n, m). These are the *coordinates* of the point. Using this notation, we can describe the function $f(n) = n^2$ as the set of ordered pairs

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

Note that we always associate the first number in the ordered pair with the horizontal location of the point and the second number with the vertical location. This is an arbitrary choice.

We can illustrate the idea of a function giving a transformation of its domain into its range nicely using its graph. Consider Fig. 185.5. 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.

185.4 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. 184.4, 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 185.12. 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. 185.8.

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. 185.8. Therefore, to graph a function accurately, we would need to evaluate it at many more points than we have used in Fig. 185.8 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

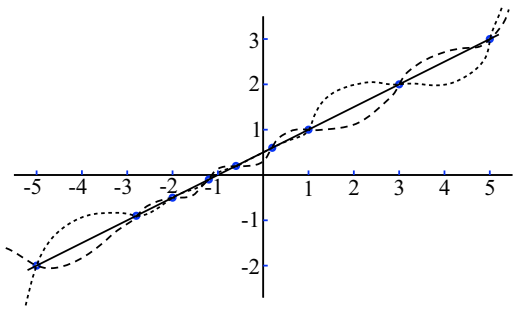


FIGURE 185.8. A plot of the function values of $f(x) = \frac{1}{2}x + \frac{1}{2}$, and several functions taking on the same values at the sample points.

the 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 185.13. We list some values of the function $f(x) = x^2$ defined on the rational numbers:

x	x^2	x	x^2	x	x^2
-4	16	$-.8$	.64	2.3	5.29
-3.5	12.25	$-.4$	.16	2.4	5.76
-3.1	9.618	0	0	3	9
-2	4	.2	.04	3.1	9.61
-1.8	3.24	1.2	1.44	3.6	12.96
-1.4	1.96	1.5	2.25	3.7	13.69
-1	1	2.21	4.8841	4	16

and then plot the function values in Fig. 185.9.

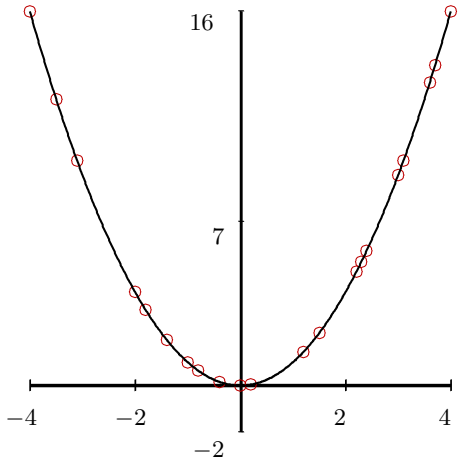


FIGURE 185.9. A plot of some of the points given by $f(x) = x^2$ and a smooth curve that passes through the points.

185.5 A Function of Two Variables

We give an example of a function of two variables. The total cost in the Dinner Soup/Ice Cream model was

$$15x + 3y,$$

where x was the amount of beef and y that of ice cream. We may view the total cost $15x + 3y$ as a function $f(x, y) = 15x + 3y$ of the two variables x and y. For each value of x and y there is a corresponding function value $f(x, y) = 15x + 3y$ representing the total cost. We think here of both x and y as independent variables which may vary freely, corresponding to any combination of beef and ice cream, and the function value $z = f(x, y)$ as a dependent variable. For each pair of values of x and y there is assigned a value of $z = f(x, y) = 15x + 3y$. We may write $(x, y) \to f(x, y) = 15x + 3y$, 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 185.14. 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 185.15. 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 185.16. 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.

185.6 Functions of Several Variables

Of course we may go further and consider functions depending on several independent variables.

EXAMPLE 185.17. 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 185.18. 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 185 Problems

185.1. Identify four functions you encounter in your daily life and determine the domain and range for each.

185.2. For the function $f(x) = 4x - 2$, determine the range corresponding to: (a) $D(f) = (-2, 4]$, (b) $D(f) = (3, \infty)$, (c) $D(f) = \{-3, 2, 6, 8\}$.

185.3. Given that $f(x) = 2 - 13x$, find the domain $D(f)$ corresponding to the range $R(f) = [-1, 1] \cup (2, \infty)$.

185.4. 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.

185.5. 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 $t = 0$.

185.6. Find the range of the function $f(n) = 1/n^2$ defined on $D(f) = \{n \in \mathbb{N} : n \geq 1\}$.

185.7. Find the domain and a set B containing the range of the function $f(x) = 1/(1 + x^2)$.

185.8. Find the domain of the functions

$$\text{(a) } \frac{2-x}{(x+2)x(x-4)(x-5)} \quad \text{(b) } \frac{x}{4-x^2} \quad \text{(c) } \frac{1}{2x+1} + \frac{x^2}{x-8}$$

185.9. *(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 $R(f)$.

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

185.11. 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) $f(n) = (n + 1)^3$.

185.12. Draw three different curves that pass through the points

$(-2, -1)$, $(-1, -.5)$, $(0, .25)$, $(1, 1.5)$, $(3, 4)$.

185.13. Plot the functions; (a) 2^{-n}, (b) 5^{-n}, and (c) 10^{-n}; defined on the natural numbers n. Compare the plots.

185.14. Plot the function $f(n) = \frac{10}{9}(1-10^{-n-1})$ defined on the natural numbers.

185.15. Plot the function $f : \mathbb{Q} \to \mathbb{Q}$ with $f(x) = x^3$ after making a table of values.

185.16. Write a *MATLAB*© function that takes two rational arguments x and y and returns their sum $x + y$.

185.17. 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.

186
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)

186.1 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

$$f(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n, \tag{186.1}$$

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 184.10.

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.

186.2 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. 186.1, 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. 186.1. This means that such points lie on a line that passes through the origin as indicated.

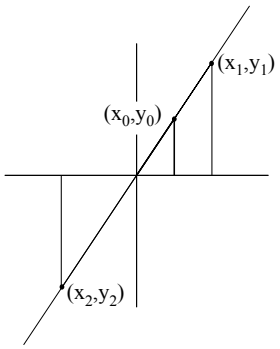


FIGURE 186.1. Points satisfying $y = mx$ form similar triangles. In this figure, $m = 3/2$.

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. 186.2. 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. 186.3. 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

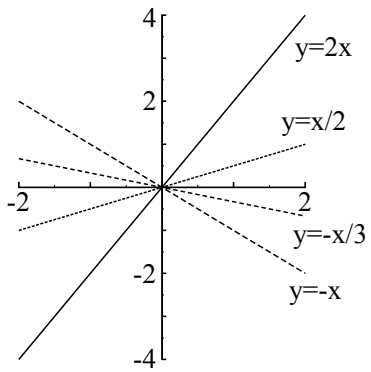


FIGURE 186.2. Examples of lines.

slope of the line through the points (x_0, y_0) and (x_1, y_1) is

$$m = \frac{y_1 - y_0}{x_1 - x_0}.$$

If (x, y) is any other point on the line, then we know that

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

or

$$(y - y_0) = m(x - x_0). \tag{186.2}$$

This is called the *point-slope* equation for a line.

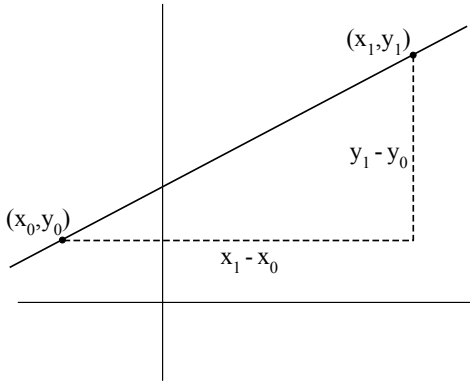


FIGURE 186.3. The slope of any line is determined by the amount of rise over the amount of run between any two points on the line.

EXAMPLE 186.1. 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 (186.2) to resemble (186.1) by multiplying out the terms in (186.2) and solving for y. This yields the *slope-intercept* form:

$$y = mx + b, \tag{186.3}$$

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. 186.4. When $b > 0$ we move the line up and when $b < 0$ we move

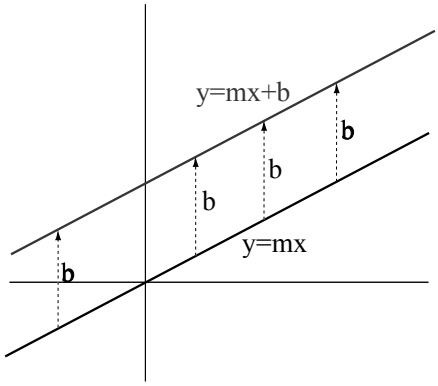


FIGURE 186.4. The graph of $y = mx + b$ is found by translating the graph of $y = mx$ vertically by an amount b. In this case $b > 0$.

the

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

EXAMPLE 186.2. 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.

186.3 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 186.3. 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 186.4. 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 $y = -7x - 4 = -7(\frac{-7}{9}) - 4 = \frac{13}{9}$.

186.4 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. 186.5. The lines

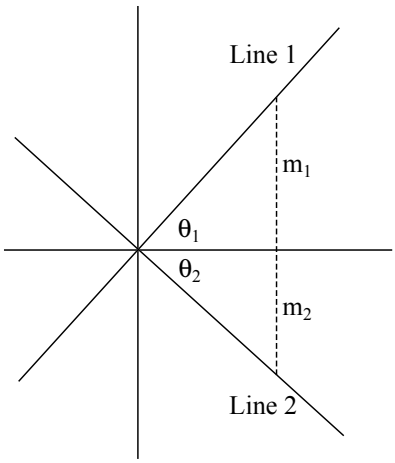


FIGURE 186.5. Similar triangles defined by perpendicular lines with slope m_1 and m_2. The angles θ_1 and θ_2 add up to 90°.

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. This means that $1/|m_1| = 1/|m_2|$ or $|m_1|\,|m_2| = 1$. This shows the result since m_1 and m_2 have opposite signs or $m_1m_2 < 0$.

Finally, assuming that $m_1m_2 = -1$ shows that the two triangles are similar and the orthogonality follows.

EXAMPLE 186.5. 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.

186.5 Quadratic Polynomials

The general quadratic polynomial has the form

$$f(x) = a_2x^2 + a_1x + a_0$$

for constants a_2, a_1, and a_0, where we assume $a_2 \neq 0$ (otherwise we go back to the linear case).

We show how to plot such a function by using the idea of plotting lines in the previous section starting with the simplest example of a quadratic function

$$y = f(x) = x^2.$$

The domain of f is the set of rational numbers while the range contains some of the nonnegative rational numbers. We list some of the values here:

x	x^2
-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

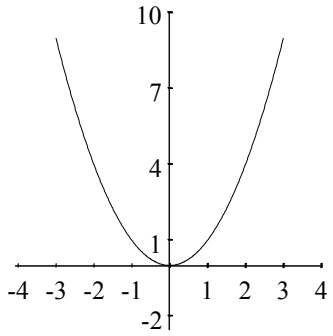


FIGURE 186.6. Plot of $f(x) = x^2$. The function is decreasing for $x < 0$ and increasing for $x > 0$.

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. 186.6 for 601 equally spaced points between $x = -3$ and $x = 3$.

To draw the graph of a general quadratic function, we follow the idea behind computing the graphs of lines by using translation. We start with $f(x) = x^2$ and then change that graph to get the graph of any other quadratic. There are two kinds of changes we make.

The first change is called *scaling*. Consider the plots of the quadratic functions in Fig. 186.7. Each of these functions has the form $y = f(x) = a_2x^2$ for a constant a_2. Their plots all have the same basic shape as $y = x^2$. However the heights of the points on $y = a_2x^2$ are a factor of $|a_2|$ higher or lower than the height of the corresponding point on $y = x^2$: higher if $|a_2| > 1$ and lower if $|a_2| < 1$. If $a_2 < 0$ then the plot is also "flipped" or *reflected* through the x-axis.

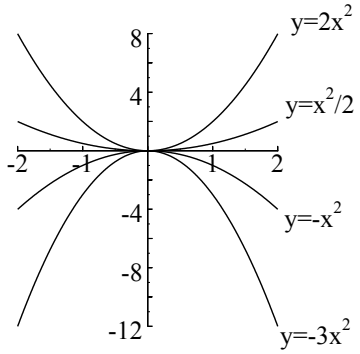


FIGURE 186.7. Plots of $y = x^2$ scaled four different ways.

The second change we consider is translation. The two possibilities are to translate horizontally, or sideways, and vertically. We show examples of both in Fig. 186.8. 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.

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

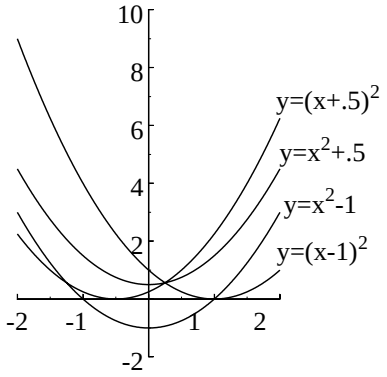


FIGURE 186.8. Plots of $y = x^2$ translated four different ways.

$y = x^2$. 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 186.6. We plot $y = -2(x+1)^2+3$ in Fig. 186.9 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).

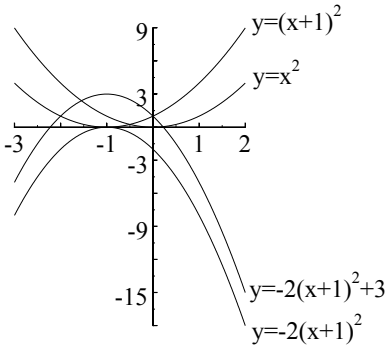


FIGURE 186.9. Plotting $y = -2(x+1)^2+3$ in a systematic way.

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 186.7. 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 186.8. 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 186.9. We complete the square on $y = \frac{1}{2}x^2 - 2x + 3$.

$$\begin{aligned}\frac{1}{2}x^2 - 2x + 3 &= \frac{1}{2}(x^2 - 4x) + 3 \\ &= \frac{1}{2}(x^2 - 4x + 4 - 4) + 3 \\ &= \frac{1}{2}(x-2)^2 - 2 + 3 \\ &= \frac{1}{2}(x-2)^2 + 1.\end{aligned}$$

186.6 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 186.10. 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 186.11. 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 (186.1) 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 186.12. 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

$$f(x) = a_0 + a_1 x^1 + a_2 x^2 + \cdots + a_n x^n$$

and

$$g(x) = b_0 + b_1 x^1 + b_2 x^2 + \cdots + b_n x^n$$

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

$$(f + g)(x) = \sum_{i=0}^{n} (a_i + b_i)x^i = \sum_{i=0}^{n} a_i x^i + \sum_{i=0}^{n} b_i x^i = f(x) + g(x).$$

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 186.13. 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

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

of degree n (assuming that $a_n \neq 0$) and the polynomial

$$g(x) = \sum_{i=0}^{m} b_i x^i$$

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 186.14.

$$\sum_{i=0}^{15}(i+1)x^i + \sum_{i=0}^{30} x^i = \sum_{i=0}^{30} a_i x^i$$

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

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

and a number $c \in \mathbb{Q}$ we define a new polynomial denoted by $(cf)(x)$, and referred to as the *product* of $f(x)$ *by the number* c, as follows:

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

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

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

EXAMPLE 186.15.

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

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 186.16. $f(x) = x^2 - 4$ 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

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

is a new polynomial called the *linear combination* of the polynomials f_1, $\cdots$, f_n with *coefficients* c_1, $\cdots$, c_n.

EXAMPLE 186.17. 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. 186.11 below, with the coefficients a_0, a_1,...,a_n. To make the notation consistent, we set $x^0 = 1$ for all x.

We sum up:

Theorem 186.1 *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

$$f + g = g + f, \tag{186.4}$$

$$(f + g) + h = f + (g + h), \tag{186.5}$$

$$c(f + g) = cf + cg, \tag{186.6}$$

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

$$\begin{aligned}(fg)(x) = x^j g(x) &= a_0 x^j + a_1 x^j \times x + a_2 x^j \times x^2 + \cdots + a_n x^j \times x^n \\ &= a_0 x^j + a_1 x^{1+j} + a_2 x^{2+j} + \cdots + a_n x^{n+j} \\ &= \sum_{i=0}^{n} a_i x^{i+j},\end{aligned}$$

which is a polynomial of degree $n + j$.

Example 186.18.

$$x^3(2 - 3x + x^4 + 19x^8) = 2x^3 - 3x^4 + x^7 + 19x^{11}.$$

Finally, for two general polynomials $f(x) = \sum_{i=0}^{n} a_i x^i$ and $g(x) = \sum_{j=0}^{m} b_j x^j$, we have

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

which is a polynomial of degree $n + m$. We consider an example

EXAMPLE 186.19.

$$\begin{aligned}(1+2x+3x^2)(x-x^5) &= 1(x-x^5)+2x(x-x^5)+3x^2(x-x^5)\\ &= x-x^5+2x^2-2x^6+3x^3-3x^7\\ &= x+2x^2+3x^3-x^5-2x^6-3x^7\end{aligned}$$

We sum up:

Theorem 186.2 *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:

$$fg = gf, \tag{186.7}$$

$$(fg)h = f(gh), \tag{186.8}$$

$$(f+g)h = fh+gh, \tag{186.9}$$

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:

$$(x+a)^2 = (x+a)(x+a) = x^2+2ax+a^2$$
$$(x+a)(x-a) = x^2-a^2$$
$$(x+a)^3 = x^3+3ax^2+3a^2x+a^3$$

186.7 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. 186.10. 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. 186.10 is 987940.8 at $x = 3$.

On the other hand, we can plot the monomials rather easily. It turns out that once the degree $n \geq 2$, the plots of the monomials with even degree n all have a similar shape, as do the plots of all the monomials with odd degree. We show some samples in Fig. 186.11. One of the most obvious feature of the graphs of the monomials are the symmetry in the plots. When the degree is even, the plots are symmetric across the y-axis, see Fig. 186.12. This means that the value of the monomial is the same for

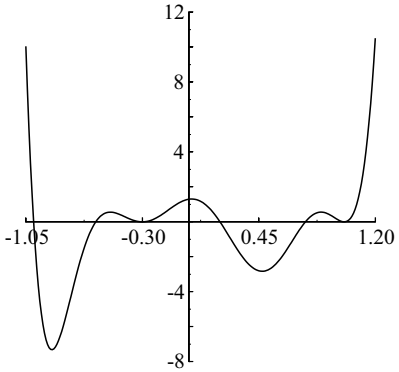


FIGURE 186.10. A plot of $y = 1.296 + 1.296x - 35.496x^2 - 57.384x^3 + 177.457x^4 + 203.889x^5 - 368.554x^6 - 211.266x^7 + 313.197x^8 + 70.965x^9 - 97.9x^{10} - 7.5x^{11} + 10x^{12}$.

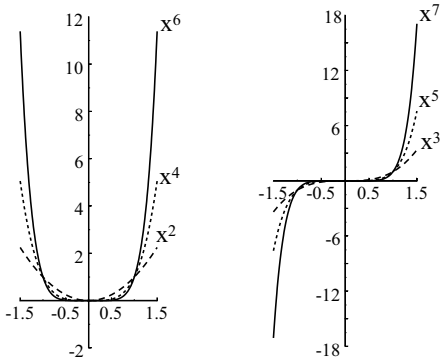


FIGURE 186.11. Plots of some monomials.

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.

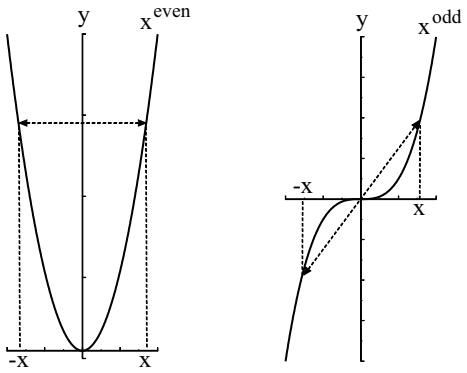


FIGURE 186.12. The symmetries of the monomial functions of even and odd degree.

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

EXAMPLE 186.20. We plot $y = -.5(x - 1)^3 - 6$ in Fig. 186.13 by systematically using translations and scaling. Luckily, however, there is no procedure like completing the square for monomials of higher degree.

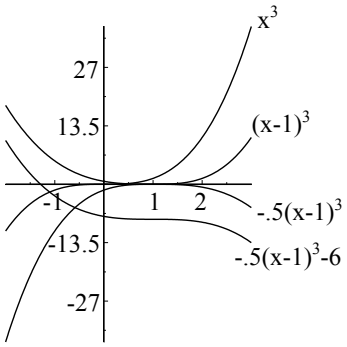


FIGURE 186.13. The procedure for plotting $y = -.5(x - 1)^3 - 6$.

186.8 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. 186.14. The most interesting thing to note about the graph of $|x|$

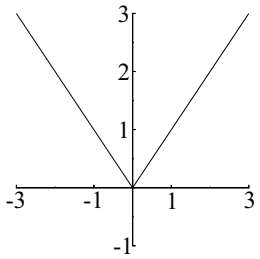


FIGURE 186.14. Plot of $y = |x|$.

is the sharp corner at $x = 0$, which occurs right at the transition point of this piecewise polynomial.

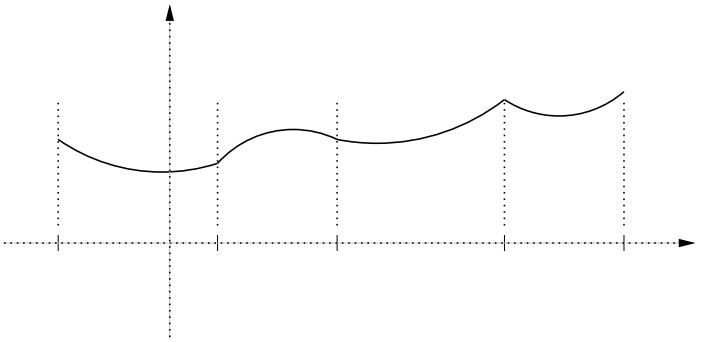


FIGURE 186.15. Plot of a piecewise (quadratic) polynomial function.

Chapter 186 Problems

186.1. 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) $(2, -1)$ & $(4, -7)$.

186.2. 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) $(-1, -1)$ & $(-4, 7)$.

186.3. Find a formula for the x-intercept of a line given in the form $y = mx + b$ in terms of m and b.

186.4. 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.

186.5. Are the lines $2 - y = 7(4 - x)$ and $y = 7x - 13$ parallel?

186.6. Are the lines $y = \frac{3}{11}x - 4$ and $y = 13 - \frac{11}{3}x$ perpendicular?

186.7. 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 $y + 3 = -4(x - 9)$.

186.8. Find the lines that are (a) parallel and (b) perpendicular to the line through $(9, 4)$ and $(-1, 3)$ and passing through the point $(3, 0)$.

186.9. Find the lines that are (a) parallel and (b) perpendicular to the line through $(-2, 7)$ and $(8, 8)$ and passing through the point $(1, 2)$.

186.10. Show that $f(x) = x^2$ is decreasing for $x < 0$.

186.11. Plot the following quadratic functions for $-2 \le x \le 2$: (a) $6x^2$, (b) $-\frac{1}{4}x^2$, (c) $\frac{4}{3}x^2$.

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

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

186.14. 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) $\frac{1}{3}(x-3)^2 - 1$.

186.15. 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) $-\frac{1}{3}x^2 + 2x - 1$.

186.16. 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) $1 + x^2 + x^4 + x^6 + \cdots + x^{2n}$.

186.17. 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 $n + 7$.

186.18. Given $f_1(x) = -4+6x+7x^3$, $f_2(x) = 2x^2-x^3+4x^5$ and $f_3(x) = 2-x^4$, compute the following polynomials: (a) $f_1 - 4f_2$, (b) $3f_2 - 12f_1$, (c) $f_2 + f_1 + f_3$, (d) $f_2 f_1$, (e) $f_1 f_3$, (f) $f_2 f_3$, (g) $f_1 f_3 - f_2$, (h) $(f_1 + f_2)f_3$, (i) $f_1 f_2 f_3$.

186.19. For a equal to a constant, compute (a) $(x+a)^2$, (b) $(x+a)^3$, (c) $(x-a)^3$, (d) $(x+a)^4$.

186.20. Compute $f_1 f_2$ where $f_1(x) = \sum_{i=0}^{8} i^2 x^i$ and $f_2(x) = \sum_{j=0}^{11} \frac{1}{j+1} x^j$.

186.21. 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.

186.22. (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 $x > 0$.

186.23. Plot the following monomial functions for $-3 \le x \le 3$: (a) x^3, (b) x^4, (c) x^5.

186.24. 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) $(x+1)^5 - 1$.

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

187
Combinations of functions

And he gave a deep sigh, and tried very hard to listen to what Owl was saying. (Winnie-the Pooh)

187.1 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.

187.2 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 187.1. 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 $\mathbb{Z}$.

187.3 Linear Combinations of Functions

Given n functions $f_1 : D_{f_1} \to \mathbb{Q}$, ..., $f_n : D_{f_n} \to \mathbb{Q}$, and numbers c_1,...,c_n, we define the *linear combination* of $f_1, ..., f_n$ with coefficients $c_1, ..., c_n$, denoted by $(c_1 f + \cdots + c_n f_n)(x)$, as follows

$$(c_1 f + \cdots + c_n f_n)(x) = c_1 f_1(x) + \cdots + c_n f_n(x)$$

The domain $D_{c_1 f + \cdots + c_n f_n}$ of the linear combination $c_1 f + \cdots + c_n f_n$ is the intersection of the domains $D_{f_1}, \cdots, D_{f_n}$.

EXAMPLE 187.2. 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 187.3. 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 $\{x \text{ in } \mathbb{Q} : x \neq 0\}$.

187.4 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 187.4. The function

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

with $D_f = \{x \in \mathbb{Q} : x \neq 0\}$ is the product of the functions $f_1(x) = (x^2 - 3)^3$ and $f_2(x) = x^6 - 1/x - 3$. The function $f(x) = x^2\, 2^x$ is the product of x^2 and 2^x.

EXAMPLE 187.5. 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 $\{x \text{ in } \mathbb{Q} : x \neq -3, 5/2\}$.

187.5 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 187.6. 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 187.7. We show a couple of examples of polynomial division. In Fig. 187.1, we give an example where the remainder is zero. We

$$\begin{array}{r|l}
x-1 & x^2 + 5x + 3 \\
\hline
x^3 + 4x^2 - 2x - 3 & \\
\underline{x^3 - x^2} & \\
5x^2 - 2x & \\
\underline{5x^2 - 5x} & \\
3x^2 - 3x & \\
\underline{3x^2 - 3x} & \\
0 &
\end{array}$$

FIGURE 187.1. An example of polynomial division with no remainder.

conclude that

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

In Fig. 187.2, 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 numerator is "missing" a term so we fill in the missing term with a zero coefficient to make the division easier. We conclude

$$
\begin{array}{r}
\;2x^2 - 2x + 15 \\
x^2+x-3 \;\big|\; 2x^4 + 0x^3 + 7x^2 - 8x + 3 \\
\underline{\;2x^4 + 2x^3 - 6x^2\;} \\
\;-2x^3+13x^2 - 8x \\
\;\underline{-2x^3 - 2x^2 + 6x\;} \\
\;15x^2 - 14x + 3 \\
\;\underline{15x^2 + 15x - 45\;} \\
\;-29x + 48
\end{array}
$$

FIGURE 187.2. An example of polynomial division with a remainder.

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

$$f(x) = (x - \bar{x})\, g(x) + r(x), \tag{187.1}$$

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

$$f(x) = (x - \bar{x})\, g(x) + r(x) \tag{187.2}$$

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 187.1 *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 $x - \bar{x}$.*

187.6 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. 187.3.

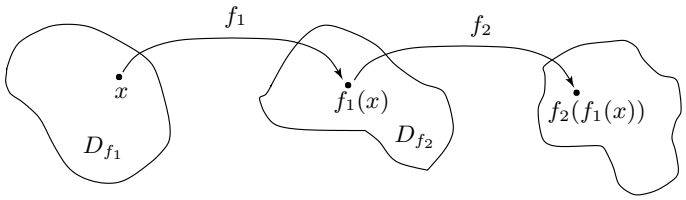


FIGURE 187.3. Illustration of the composition $f_2 \circ f_1$.

EXAMPLE 187.8. 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 187.9. 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 187 Problems

187.1. Determine the domains of the following functions

(a) $3(x-4)^3 + 2x^2 + \dfrac{4x}{3x-1} + \dfrac{6}{(x-1)^2}$

(b) $2 + \dfrac{4}{x} - \dfrac{6x+4}{(x-2)(2x+1)}$

(c) $x^3\left(1 + \dfrac{1}{x}\right)$

(d) $\dfrac{(2x-3)^{\frac{2}{x}}}{4x+6}$

(e) $\dfrac{6x-1}{(2-3x)(4+x)}$

(f) $\dfrac{4}{x+2} + \dfrac{6}{x^2+3x+2}$

187.2. 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) $\dfrac{2}{x-1} + \dfrac{4}{x-2} + \dfrac{8}{x-3} + \cdots + \dfrac{8192}{x-13}$

187.3. (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.

187.4. 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 $\mathbb{N}$

187.5. 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) $f_1 \circ f_2 \circ f_3$

187.6. With $f_1(x) = 4x + 2$ and $f_2(x) = x/x^2$, show that $f_1 \circ f_2 \neq f_2 \circ f_1$.

187.7. 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.

187.8. 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 $f_2(x) = \dfrac{x+1}{x}$

188
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)

188.1 Introduction

When we graph a function $f(x)$ of a rational variable x, we make a leap of faith and assume that the function values $f(x)$ vary "smoothly" or "continuously" between the sample points x, so that we can draw the graph of the function without lifting the pen. In particular, we assume that the function value $f(x)$ does not make unknown sudden jumps for some values of x. We thus assume that the function value $f(x)$ changes by a small amount if we change x by a small amount. A basic problem in Calculus is to measure how much the function values $f(x)$ may change when x changes, that is, to measure the "degree of continuity" of a function. In this chapter, we approach this basic problem using the concept of *Lipschitz continuity*, which plays a basic role in the version of Calculus presented in this book.

There will be a lot of inequalities ($<$ and $\leq$) and absolute values ($|\cdot|$) in this chapter, so it might be a good idea before you start to review the rules for operating with these symbols from Chapter *Rational numbers*.

FIGURE 188.1. Rudolph Lipschitz (1832-1903), Inventor of Lipschitz continuity: "Indeed, I have found a very nice way of expressing continuity....".

188.2 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

$$|f(x_2) - f(x_1)| = |(mx_2 + b) - (mx_1 + b)| = |m(x_2 - x_1)| = |m||x_2 - x_1|. \tag{188.1}$$

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 188.1. 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. 188.2.

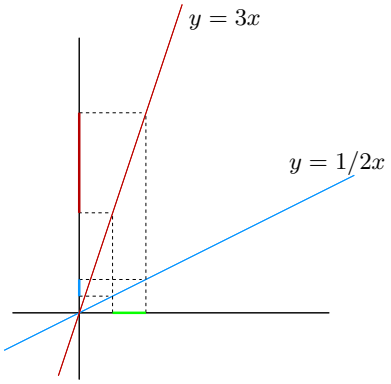


FIGURE 188.2. These two linear functions which change a different amount for a given change in input.

EXAMPLE 188.2. 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 $.01/100 = .0001$.

188.3 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

$$|f(x_1) - f(x_2)| \leq L_f |x_1 - x_2| \quad \text{for all } x_1, x_2 \in I. \tag{188.2}$$

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 188.3. 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 188.4. 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 186.6, we get

$$|f(x_2) - f(x_1)| = |x_2 + x_1|\,|x_2 - x_1|. \tag{188.3}$$

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 (188.1) 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 (188.2) and we have to be circumspect about drawing conclusions when the Lipschitz constant is large. The reason is that (188.2) 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 188.5. From Example 188.4, 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

$$|f(x_2) - f(x_1)| \leq 4|x_2 - x_1| \leq 121|x_2 - x_1|.$$

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 188.6. We show that $f(x) = x^2$ is Lipschitz continuous on the interval $I = [2, 4]$, with Lipschitz constant $L_f = 8$. Starting with (188.3), 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. 188.3, where we show the change in f corresponding to equal changes in x near $x = 2$ and $x = 4$. Because $f(x) = x^2$ is steeper near $x = 4$, f changes more near $x = 4$ for a given change in input.

EXAMPLE 188.7. $f(x) = x^2$ 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 188.8. The set of rational numbers $I = [-1, 500]$ is bounded but the set of even integers is not bounded.

While linear functions are Lipschitz continuous on the unbounded set $\mathbb{Q}$, functions that are not linear are usually only Lipschitz continuous on bounded sets.

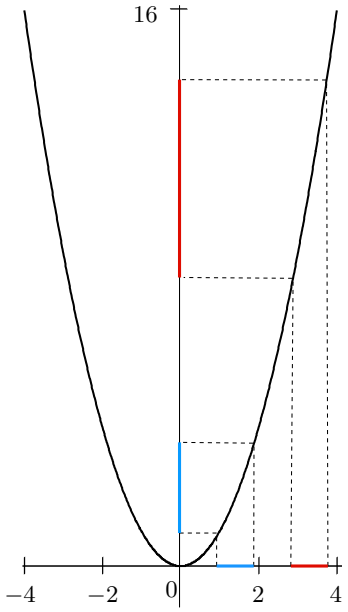


FIGURE 188.3. The change in $f(x) = x^2$ for equal changes in x near $x = 2$ and $x = 4$.

EXAMPLE 188.9. The function $f(x) = x^2$ is **not** Lipschitz continuous on the set $\mathbb{Q}$ of rational numbers. This follows from (188.3) 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.

188.4 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 188.10. 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. 186.12.

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 188.11. The function $f(x) = x^m$ is Lipschitz continuous on any interval $I = [-a, a]$, where a is a positive rational number, with Lipschitz constant $L_f = ma^{m-1}$. Given x_1 and x_2 in I, we want to estimate

$$|f(x_2) - f(x_1)| = |x_2^m - x_1^m|$$

in terms of $|x_2 - x_1|$. We can do this using the fact that

$$\begin{aligned} x_2^m - x_1^m &= (x_2 - x_1)(x_2^{m-1} + x_2^{m-2}x_1 + \cdots + x_2x_1^{m-2} + x_1^{m-1}) \\ &= (x_2 - x_1) \sum_{i=0}^{m-1} x_2^{m-1-i}x_1^i. \end{aligned}$$

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

$$\begin{aligned} &(x_2 - x_1) \sum_{i=0}^{m-1} x_2^{m-1-i}x_1^i \\ &\qquad = x_2^m + \sum_{i=1}^{m-1} x_2^{m-i}x_1^i - \sum_{i=1}^{m-1} x_2^{m-i}x_1^i - x_1^m = x_2^m - x_1^m. \end{aligned}$$

This is tedious, but it is good practice to go through the details and make sure this argument is correct.

This means that

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

We have the desired difference $|x_2 - x_1|$ on the right and we just have to bound the factor

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

By the triangle inequality

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

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

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

and

$$|f(x_2) - f(x_1)| \leq m a^{m-1} |x_2 - x_1|.$$

188.5 Linear Combinations of Functions

Now that we have seen that the monomials are Lipschitz continuous on a given bounded interval, it is a short step to show that any polynomial is Lipschitz continuous on a given bounded interval. But rather than just do this for polynomials, we show that a linear combination of arbitrary Lipschitz continuous functions is Lipschitz continuous

Suppose that f_1 is Lipschitz continuous with constant L_1 and f_2 is Lipschitz continuous with constant L_2 on the interval I. Note that here (and below) we condense the notation and write e.g. L_1 instead of L_{f_1}. Then $f_1 + f_2$ is Lipschitz continuous with constant $L_1 + L_2$ on I, because if we choose two points x and y in I, then

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

by the triangle inequality. The same argument shows that $f_2 - f_1$ is Lipschitz continuous with constant $L_1 + L_2$ as well (not $L_1 - L_2$ of course!). It is even easier to show that if $f(x)$ is Lipschitz continuous on an interval I with Lipschitz constant L then $cf(x)$ is Lipschitz continuous on I with Lipschitz constant $|c|L$.

From these two facts, it is a short step to extend the result to any linear combination of Lipschitz continuous functions. Suppose that $f_1, \cdots, f_n$ are Lipschitz continuous on I with Lipschitz constants $L_1, \cdots, L_n$ respectively. We use induction, so we begin by considering the linear combination of two functions. From the remarks above, it follows that $c_1 f_1 + c_2 f_2$ is Lipschitz continuous with constant $|c_1|L_1 + |c_2|L_2$. Next given $i \leq n$, we assume that

$c_1f_1 + \cdots + c_{i-1}f_{i-1}$ is Lipschitz continuous with constant $|c_1|L_1 + \cdots + |c_{i-1}|L_{i-1}$. To prove the result for i, we write

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

But the assumption on $\big(c_1f_1 + \cdots + c_{i-1}f_{i-1}\big)$ means that we have written $c_1f_1 + \cdots + c_if_i$ as the sum of two Lipschitz continuous functions, namely $\big(c_1f_1 + \cdots + c_{i-1}f_{i-1}\big)$ and c_nf_n. The result follows by the result for the linear combination of two functions. By induction, we have proved

Theorem 188.1 *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 188.2 *A polynomial is Lipschitz continuous on any bounded interval.*

EXAMPLE 188.12. We show that the function $f(x) = x^4 - 3x^2$ is Lipschitz continuous on $[-2, 2]$, with constant $L_f = 44$. For x_1 and x_2 in $[-2, 2]$, we have to estimate

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

From Example 188.11, 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

$$|f(x_2) - f(x_1)| \le 32|x_2 - x_1| + 3 \times 4|x_2 - x_1| = 44|x_2 - x_1|.$$

188.6 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. 188.4

$$|f(x)| \le M \text{ for all } x \text{ in } I.$$

In fact if we think about the estimates we have made to verify the definition of Lipschitz continuity (188.2), we see that in every case these involved showing that some function is bounded on the given interval.

EXAMPLE 188.13. To show that $f(x) = x^2$ is Lipschitz continuous on $[-2, 2]$ in Example 188.4, we proved that $|x_1 + x_2| \le 4$ for x_1 and x_2 in $[-2, 2]$.

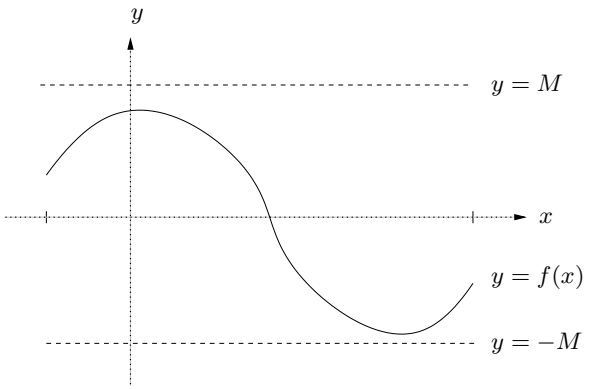


FIGURE 188.4. A bounded function on I.

It turns out that a function that is Lipschitz continuous on a bounded domain is automatically bounded on that domain. To be more precise, suppose that a function f is Lipschitz continuous with Lipschitz constant L_f on a bounded set I with size a and choose a point y in I. Then for any other point x in I

$$|f(x) - f(y)| \leq L_f|x - y|.$$

First we know that $|x - y| \leq |x| + |y| \leq 2a$. Also, since $|b + c| \leq |d|$ means that $|b| \leq |d| + |c|$ for any numbers a, b, c, we get

$$|f(x)| \leq |f(y)| + L_f|x - y| \leq |f(y)| + 2L_f a.$$

Even though we don't know $|f(y)|$, we do know that it is finite. This shows that $|f(x)|$ is bounded by the constant $M = |f(y)| + 2L_f a$ for any x in Q. We express this by saying that $f(x)$ is *bounded on* I. We have thus proved

Theorem 188.3 *A Lipschitz continuous function on a bounded set I is bounded on I.*

EXAMPLE 188.14. In Example 188.12, 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.

188.7 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

$$\begin{aligned}|f_1(y)f_2(y) - f_1(x)f_2(x)| &\\ &= |f_1(y)f_2(y) - f_1(y)f_2(x) + f_1(y)f_2(x) - f_1(x)f_2(x)|\\ &\leq |f_1(y)f_2(y) - f_1(y)f_2(x)| + |f_1(y)f_2(x) - f_1(x)f_2(x)|\\ &= |f_1(y)|\,|f_2(y) - ff_2(x)| + |f_2(x)|\,|f_1(y) - f_1(x)|\end{aligned}$$

Now Theorem 188.3, 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 188.4 *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 188.15. 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 188.4.

188.8 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 188.16. We show that $f(x) = 1/x^2$ is Lipschitz continuous on the interval $[1/2, 2]$, with Lipschitz constant $L = 64$. We choose two points x_1 and x_2 in Q and we estimate the change

$$|f(x_2) - f(x_1)| = \left|\frac{1}{x_2^2} - \frac{1}{x_1^2}\right|$$

by first doing some algebra

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

This means that

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

Now we have the good difference on the right, we just have to bound the factor. The numerator of the factor is the same as in Example 188.4, 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 188.5 *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 188.17. The function $1/x^2$ does not satisfy the assumptions of Theorem 188.5 on the interval $[0, 2]$ and we know that it is not Lipschitz continuous on that interval.

188.9 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 188.6 *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 188.18. 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 188.10, 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 188.19. 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 $x = \pm 2$.

188.10 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 188.20. 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 188.21. 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 $x_1, x_2 \in [0, 1]$

$$\begin{aligned} |x_1 x_2 - y_1 y_2| &= |x_1 x_2 - y_1 x_2 + y_1 x_2 - y_1 y_2| \\ &\leq |x_1 - y_1| x_2 + y_1 |x_2 - y_2| \leq 2(|x_1 - y_1| + |x_2 - y_2|). \end{aligned}$$

188.11 Functions of Several Rational Variables

The concept of a function also extends to several variables, i.e. we consider functions $f(x_1, ..., x_d)$ of d rational variables. We write $f : \mathbb{R}^d \to \mathbb{Q}$ if for given rational numbers $x_1, \cdots, x_d$, a rational number denoted by $f(x_1, ..., x_d)$ is given.

The definition of Lipschitz continuity also directly extends. We say that $f : \mathbb{Q}^d \to \mathbb{Q}$ is Lipschitz continuous with Lipschitz constant L_f if for all $x_1, \cdots, x_d \in \mathbb{Q}$ and $y_1, \cdots, y_d \in \mathbb{Q}$,

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

EXAMPLE 188.22. The function $f : \mathbb{R}^d \to \mathbb{Q}$ defined by $f(x_1, ..., x_d) = x_1 + x_2 + \cdots x_d$ is Lipschitz continuous with Lipschitz constant $L_f = 1$.

Chapter 188 Problems

188.1. Verify the claims in Example 188.7.

188.2. Show that $f(x) = x^2$ is Lipschitz continuous on $[10, 13]$ directly and compute a Lipschitz constant.

188.3. Show that $f(x) = 4x - 2x^2$ is Lipschitz continuous on $[-2, 2]$ directly and compute a Lipschitz constant.

188.4. Show that $f(x) = x^3$ is Lipschitz continuous on $[-2, 2]$ directly and compute a Lipschitz constant.

188.5. Show that $f(x) = |x|$ is Lipschitz continuous on $\mathbb{Q}$ directly and compute a Lipschitz constant.

188.6. In Example 188.10, 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.

188.7. Show that $f(x) = 1/x^2$ is Lipschitz continuous on $[1, 2]$ directly and compute the Lipschitz constant.

188.8. Show that $f(x) = 1/(x^2 + 1)$ is Lipschitz continuous on $[-2, 2]$ directly and compute a Lipschitz constant.

188.9. Compute the Lipschitz constant of $f(x) = 1/x$ on the intervals (a) $[.1, 1]$, (b) $[.01, 1]$, and $[.001, 1]$.

188.10. Find the Lipschitz constant of the function $f(x) = \sqrt{x}$ with $D(f) = (\delta, \infty)$ for given $\delta > 0$.

188.11. Explain why $f(x) = 1/x$ is not Lipschitz continuous on $(0, 1]$.

188.12. (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 $[1, 4]$?

188.13. 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.

188.14. 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.

188.15. Show that the Lipschitz constant of a polynomial $f(x) = \sum_{i=0}^{n} a_i x^i$ on the interval $[-c, c]$ is

$$L = \sum_{i=1}^{n} |a_i| i c^{i-1} = |a_1| + 2c|a_2| + \cdots + nc^{n-1}|a_n|.$$

188.16. Explain why $f(x) = 1/x$ is not bounded on $[-1, 0]$.

188.17. Prove Theorem 188.5.

188.18. 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 $[1, 2]$

188.19. 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 $[0, 1]$.

189
Sequences and limits

He sat down and thought, in the most thoughtful way he could think.
(Winnie-the-Pooh)

189.1 A First Encounter with Sequences and Limits

The decimal expansions of rational numbers discussed in chapter Rational Numbers leads into the concepts *sequence*, *converging sequence* and *limit* of a sequence, which play a fundamental role in mathematics. The development of calculus has largely been a struggle to come to grips with certain evasive aspects of these concepts. We will try to uncover the mysteries by being as concrete and down-to-earth as possible.

We begin recalling the decimal expansion 1.11... of $\frac{10}{9}$, and that by (184.7)

$$\frac{10}{9} = 1.11\cdots 11_n + \frac{1}{9}10^{-n}. \tag{189.1}$$

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 $10/9$,

$$\left|\frac{10}{9} - 1.11\cdots 11_n\right| \leq 10^{-n}. \tag{189.2}$$

This estimate shows that we may consider $1.11\cdots11_n$ as an approximation of $10/9$, which becomes increasingly accurate as the number of decimal places n increases. In other words, the *error* $|10/9 - 1.11\cdots11_n|$ can be made as small as we please by taking n sufficiently large. If we want the error to be smaller than or equal to 10^{-10}, then we simply choose $n \geq 10$.

We may view the successive approximations 1.1, 1.11, 1.111, $1.11...11_n$, and so on, as a *sequence* of numbers a_n, with $n = 1, 2, 3, ...$, where $a_1 = 1.1$, $a_2 = 1.11$,..., $a_n = 1.11...11_n$,..., are called the *elements of the sequence*. More generally, a sequence a_1, a_2, a_3,..., is a never-ending list of elements a_1, a_2, a_3,..., where the index takes successively the values of the natural numbers $1, 2, 3, ...$ A *sequence of rational numbers* is a list a_1, a_2, a_3,..., where each element a_n is a rational number. We will denote a sequence by

$$\{a_n\}_{n=1}^{\infty}$$

which thus means the never ending list a_1, a_2, a_3,..., of elements a_n, with the index n going through the natural numbers $n = 1, 2, 3, ...$ The symbol ∞, called "infinity", indicates that the list continues for ever in the same sense that the natural numbers $1, 2, 3, ...$, continues for ever without coming to an end.

We now return to the sequence of rational numbers $\{a_n\}_{n=1}^{\infty}$, where $a_n = 1.11..11_n$, that is the sequence $\{1.11..11_n\}_{n=1}^{\infty}$. The accuracy of element $a_n = 1.11...11_n$, as an approximation of $\frac{10}{9}$, increases as the number of decimals n increases. Each number in the sequence in turn is a better approximation to $10/9$ than the preceding number and as we move from left to right the numbers become ever closer to $10/9$. An advantage of considering the sequence $\{1.11..11_n\}_{n=1}^{\infty}$ or 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 $1.1, ..., 1.11..11_n, ...$, thus can be viewed as a collection of successively more accurate approximations of $\frac{10}{9}$, where we can satisfy any desired accuracy.

We say that the sequence $\{1.11..11_n\}_{n=1}^{\infty}$ *converges* to the value $\frac{10}{9}$, since the difference between $\frac{10}{9}$ and $1.11..11_n$ becomes smaller than any given positive number if only we take n large enough, as follows from (189.2). We say that $\frac{10}{9}$ is the *limit* of the sequence $\{a_n\}_{n=1}^{\infty}$=$\{1, 11..11_n\}_{n=1}^{\infty}$. We will express the convergence of the sequence $\{a_n\}_{n=1}^{\infty}$ with elements $a_n = 1.11...11_n$, as follows:

$$\lim_{n\to\infty} a_n = \frac{10}{9} \quad \text{or} \quad \lim_{n\to\infty} 1.11..11_n = \frac{10}{9}.$$

The limit $\frac{10}{9}$ does not have a finite decimal expansion. The elements $1.11..11_n$ of the converging sequence $\{1.11..11_n\}_{n=1}^{\infty}$ are finite decimal approximations of the limit $\frac{10}{9}$, with an error which is smaller than any given positive number if we only take n large enough.

Suppose that we restrict ourselves to work with finite decimal expansions, which is what a computer usually does. In this case we cannot exactly express the value $\frac{10}{9}$ with the available resources, because $\frac{10}{9}$ does not have a finite decimal expansion. As a substitute or approximation we may choose for example $1.11..11_{10}$, but there is limit to the accuracy with this single element. It would not be entirely correct to say that $\frac{10}{9} = 1.11..11_{10}$. If we instead have the whole sequence $\{1.11..11_n\}_{n=1}^{\infty}$ at hand, then we can meet any accuracy by choosing the element $1.11..11_n$ with n large enough. Choosing more and more decimals, we could increase the accuracy to any desired degree.

The sequence $\{1.11..11_n\}_{n=1}^{\infty}$ includes finite decimal approximations of $\frac{10}{9}$ satisfying any given positive tolerance or accuracy requirement. This is sometimes expressed as

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

where the three little dots are there to indicate that any precision could be attained by taking sufficiently many decimals (all equal to 1). Another way of writing this, would be

$$\lim_{n\to\infty} 1.11..11_n = \frac{10}{9},$$

avoiding the possible ambiguity using the three little dots.

189.2 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. 189.1.

An *amateur mechanic* would have one socket, of say dimension 0.7. A *pro mechanic* would perhaps have 10 sockets of dimensions 0.7, 0.67, 0.667,....,$0.66\cdots667_{10}$. Both the amateur and pro would get stuck under sufficiently tough conditions because the socket would be too large to do the job.

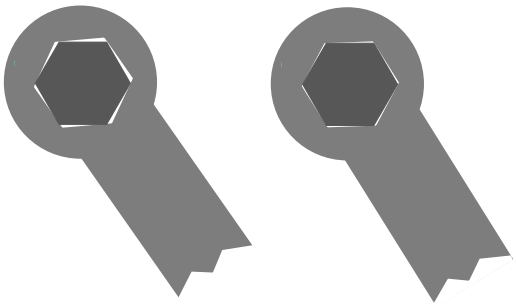


FIGURE 189.1. Two socket wrenches with different tolerances.

A *ideal expert mechanic* would have the whole sequence $\{0.66\cdots67_n\}_{n=1}^{\infty}$ at his/her disposal with the error of wrench number n being estimated by

$$|0.66\cdots67_n - \frac{2}{3}| \leq 10^{-n}.$$

The ideal expert can thus reach into the tool chest and pull out a wrench that meets any accuracy requirement, and would thus be able to turn the bolt under arbitrarily tough conditions, or meet any crank torque specified by a bicycle manufacturer. More precisely, the ideal expert could be thought of as being able to *construct* a socket himself to meet any given tolerance or accuracy. If necessary, the ideal mechanic could *construct* a wrench of for example the dimension $0.66\cdots67_{20}$, unless he already has such a wrench in his (big) tool chest. The amateur and pro mechanic would not have this capability of constructing their own wrenches, but would have to be content with their ready-made wrench sets (which they could buy in the hard-ware store). We expect the cost to construct a wrench of dimension $0.6\cdots67_n$ to increase (rapidly) with n, since the precision in the construction process has to improve.

As a general point, computing the numbers $0.66\cdots67_n$ by long division of $\frac{2}{3}$, requires more work as n increases. What we gain from doing more work is better accuracy in using $0.66\cdots67_n$ as an approximation to $\frac{2}{3}$. Trading work for accuracy is the central idea behind solving equations using computation, especially on a computer. An estimate like (189.2) 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 $\frac{2}{3}$.

189.3 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. 189.2. In principle the adjustable wrench is an analog device which fits a bolt of any size within a certain range. Every mechanic knows that an adjustable wrench may fail in cases when a properly chosen fixed size wrench does not, because the size of the adjustable wrench is not completely stable under increasing torque. .

FIGURE 189.2. The Swedish inventor J.P.Johansson with two adjustable wrenches of different design.

189.4 The Power of Language: From Infinitely Many to One

The decimal expansion 0.6666.... of $\frac{2}{3}$ contains infinitely many decimals. The sequence $\{0.66\cdots667_n\}_{n=1}^{\infty}$ contains infinitely many elements, which are increasingly accurate approximations of $\frac{2}{3}$. Talking or thinking of infinitely many decimals or infinitely many elements, presents a serious difficulty, which is handled by introducing the concept of a sequence. A sequence has *infinitely many elements*, but the sequence itself is just *one entity*. We thus group the infinitely many elements together to form one

sequence, and thus pass from infinity to one. After this semantic construction, we are thus able to speak about *one* sequence and may momentarily forget that the sequence in fact has infinitely many elements.

This would be like speaking about the expert mechanics tool chest containing the sequence $\{0.66\cdots667_n\}_{n=1}^{\infty}$ of infinitely many wrenches as one entity. One tool chest with infinitely many wrenches. To call a tool chest a wrench seems strange initially, but we could get there by first calling the tool chest something like a "super-wrench", and then later omit the "super".

Analogously, we could say that 0.6666.... is a "super-number" because it has infinitely many decimals, and then forget the "super" and say that 0.6666...., is a number. In fact, this makes complete sense since we identify 0.6666.... with $\frac{2}{3}$, which is a number. Below, we shall meet non-periodic infinite decimal expansions that do not correspond to rational numbers. Initially, we may think of these as some kind of "super-number", and then later will refer such numbers as "real numbers".

The discussion illustrates the usefulness of the concept of *one* set or sequence with *infinitely* many elements. Of course, we should be aware of the risk involved using the language to hide real facts. Political language is often used this way, which is one reason for the eroding credibility of politicians. As mathematicians, there is no reason that we should try to be as honest as possible, and use the language as clearly as possible.

189.5 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. 189.3 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

$$|a_n - A| \leq \epsilon \quad \text{for all} \quad n \geq N$$

For example, we know that the value of $10/9$ is approximated by the element $1.11\cdots1_n$ from the sequence $\{1.11\cdots1_n\}$ to any specified accuracy (bigger

than zero) by taking n sufficiently large. We know from (189.2) that

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

and thus

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

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

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

if $n \geq N$, where $10^{-N} \leq \epsilon$. If $\epsilon = .p_1p_2\cdots$, where $p_1 = p_2 = \cdots = p_m = 0$, while $p_{m+1} \neq 0$, then we may choose any N such that $N \geq m$. We see that choosing ϵ smaller, requires N to be bigger, and thus N depends on ϵ.

We emphasize that the $\epsilon - N$ definition of convergence is a fancy way of saying that the difference $|A - a_n|$ can be made smaller than any given positive number if only n is taken large enough.

There is a risk (and temptation) in using the $\epsilon - N$ definition of convergence, instead of the more pedestrian "as small as we please if only n is large enough". The statement "$|A-a_n|$ can be made smaller than any given positive number if only n is large enough" is a very qualitative statement. Nothing is said about *how large* n has to be to reach a certain accuracy. A very qualitative statement is necessarily a bit vague. On the other hand, the statement "for any $\epsilon > 0$ there is an N such that $|A - a_n| \leq \epsilon$ if $n \geq N$" has the form of a very exact and precise statement, while in fact it may be as qualitative as the first statement, unless the dependence of N on ϵ is made clear. The risk is thus that using the $\epsilon - N$-jargon, we may get confused and believe that something vague, in fact is very precise. Of course there is also a temptation in this, which relates to the general idea of mathematics as something being extremely precise. So be cautious and don't get fooled by simple tricks: the $\epsilon - N$ limit definition is vague to the extent the dependence of N on ϵ is vague.

The concept of a limit of a sequence of numbers is central to calculus. It is closely connected to never-ending decimal expansions, that is decimal expansions with infinitely many non-zero decimals. The elements in the sequence with this connection are obtained by successively taking more and more decimals into account. In fact, the fundamental reason for looking at sequences comes form this connection. However, as happens, the idea of a sequence and limit has taken on a life of its own, which has been plaguing many students of calculus. We will try to refrain from excesses in this direction and keep a strong connection with the original motivation for introducing the concepts of sequences and limits, namely describing successively better and better approximations of solutions of equations.

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

FIGURE 189.3. 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".

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 189.1. 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

$$\left|\frac{1}{n} - 0\right| \leq \epsilon \tag{189.3}$$

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 (189.3) 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 189.2. 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

$$\lim_{n\to\infty} \frac{n}{n+1} = 1. \tag{189.4}$$

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 189.3. 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 $s_n = 1.11\cdots11_n$. 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

$$\lim_{n\to\infty} s_n = \lim_{n\to\infty} (1 + r + r^2 + \cdots + r^n) = \frac{1}{1 - r}, \tag{189.5}$$

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 (189.5). We need to show that for any $\epsilon > 0$, there is an N such that

$$\left| \frac{1 - r^{n+1}}{1 - r} - \frac{1}{1 - r} \right| = \left| \frac{r^{n+1}}{1 - r} \right| \leq \epsilon$$

for all $n \geq N$. To this end it is sufficient, since $|r| < 1$, to find N such that

$$|r|^{N+1} \leq \epsilon \, |1 - r| \tag{189.6}$$

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 (189.6) 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 (189.6).

189.6 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 184.15,

$$|a - b| \leq |a - c| + |c - b| \tag{189.7}$$

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 (189.7) 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 (189.7), 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 $\lim_{n\to\infty} a_n$.

189.7 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

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

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

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

Since $\{a_n\}$ converges to A, the right-hand side can be made smaller than any given positive number by taking m large enough, and thus we can also make the left hand side smaller than any positive number by choosing m large enough, which shows the desired result.

Note that since $\lim_{n\to\infty} a_n$ is a rational number, the function value $f(\lim_{n\to\infty} a_n)$ is well defined since we assume that $f : \mathbb{Q} \to \mathbb{Q}$.

We see that it is sufficient that $f(x)$ is Lipschitz continuous on an interval I containing all the elements a_n as well as $\lim_{n\to\infty} a_n$. We have thus proved the following fundamental result.

Theorem 189.1 *Let $\{a_n\}$ be a sequence with rational limit $\lim_{n\to\infty} a_n$. Let $f : I \to \mathbb{Q}$ be a Lipschitz continuous function, and assume that $a_n \in I$ for all n and $\lim_{n\to\infty} a_n \in I$. Then,*

$$\lim_{n\to\infty} f(a_n) = f\left(\lim_{n\to\infty} a_n\right). \tag{189.8}$$

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 189.4. In the growth of bacteria model of Chapter Rational Numbers, we need to compute

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

The sequence $\{P_n\}$ is obtained by applying the function

$$f(x) = \frac{1}{Q_0 x + \frac{1}{K}(1-x)}$$

to the terms in the sequence $\{\frac{1}{2^n}\}$. Since $\lim_{n\to\infty} 1/2^n = 0$, we have $\lim_{n\to\infty} P_n = f(0) = K$, since f is Lipschitz continuous on for example $[0, 1/2]$. The Lipschitz continuity follows from the fact that $f(x)$ is the composition of the function $f_2(x) = Q_0 x + \frac{1}{K}(1-x)$ and the function $f_2(y) = 1/y$.

EXAMPLE 189.5. 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 189.6. By Theorem 189.1, 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 189.7. By Theorem 189.1,

$$\lim_{n\to\infty} (2^{-n})^7 + 14(2^{-n})^4 - 3(2^{-n}) + 2 = 0^7 + 14\times 0^4 - 3\times 0 + 2 = 2.$$

189.8 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

$$|f(x_1, x_2) - f(\bar{x}_1, \bar{x}_2)| \le L(|x_1 - \bar{x}_1| + |x_2 - \bar{x}_2|)$$

for $x_1, \bar{x}_1 \in I$ and $x_2, \bar{x}_2 \in J$.

Let now $\{a_n\}$ and $\{b_n\}$ be two converging sequences of rational numbers with $a_n \in I$ and $b_n \in J$. Then

$$f(\lim_{n\to\infty} a_n, \lim_{n\to\infty} b_n)) = \lim_{n\to\infty} f(a_n, b_n). \tag{189.9}$$

The proof is immediate:

$$|f(\lim_{n\to\infty} a_n, \lim_{n\to\infty} b_n)) - f(a_m, b_m)| \le L(|\lim_{n\to\infty} a_n - a_m| + |\lim_{n\to\infty} b_n - b_m|) \tag{189.10}$$

where the right hand side can be made arbitrarily small by choosing m large enough.

We give a first application of this result with $f(x_1, x_2) = x_1 + x_2$, which is Lipschitz continuous on $\mathbb{Q} \times \mathbb{Q}$ with Lipschitz constant $L = 1$. We conclude from (189.9) the natural formula

$$\lim_{n\to\infty} (a_n + b_n) = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n \tag{189.11}$$

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 (189.9) 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.

189.9 Computing Limits

We now apply the above rules to compute some limits.

EXAMPLE 189.8. Consider $\{2 + 3n^{-4} + (-1)^n n^{-1}\}_{n=1}^{\infty}$.

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

To do this example, we use (189.11) and the fact that

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

for any natural number p.

Another useful fact is

$$\lim_{n\to\infty} r^n = \begin{cases} 0 & \text{if } |r| < 1, \\ 1 & \text{if } r = 1, \\ \text{diverges to } \infty & \text{if } r > 1, \\ \text{diverges} & \text{otherwise.} \end{cases}$$

We showed the case when $r = 1/2$ in Example 189.4 and you will show the general result later as an exercise.

EXAMPLE 189.9. Using 189.3, we can solve for the limiting behavior of the population of bacteria described in Example 189.4. We have

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

In words, the population of the bacteria growing under the limited resources as modeled by the Verhulst model tends to a constant population.

EXAMPLE 189.10. Consider

$$\left\{ 4\,\frac{1+n^{-3}}{3+n^{-2}} \right\}_{n=1}^{\infty}.$$

We compute the limit using the different rules:

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

EXAMPLE 189.11. 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 189.12.

$$\begin{aligned}\lim_{n\to\infty} \frac{n^3-20n^2+1}{n^8+2n} &= \lim_{n\to\infty} \frac{(n^3-20n^2+1)n^{-3}}{(n^8+2n)n^{-3}} \\ &= \lim_{n\to\infty} \frac{1-20n^{-1}+n^{-3}}{n^5+2n^{-2}}.\end{aligned}$$

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 189.13.

$$\begin{aligned}\lim_{n\to\infty} \frac{-n^6+n+10}{80n^4+7} &= \lim_{n\to\infty} \frac{(-n^6+n+10)n^{-4}}{(80n^4+7)n^{-4}} \\ &= \lim_{n\to\infty} \frac{-n^2+n^{-3}+10n^{-4}}{80+7n^{-4}}.\end{aligned}$$

From this we see that the numerator grows in the negative direction without bound while the denominator tends towards 80. Therefore

$$\left\{\frac{-n^6+n+10}{80n^4+7}\right\}_{n=1}^{\infty} \quad \text{diverges to } -\infty.$$

189.10 Computer Representation of Rational Numbers

The decimal expansion $\pm p_m p_{m-1} \cdots p_1 p_0 . q_1 q_2 \cdots q_n$ uses the base 10 and consequently each of the digits p_i and q_j may take on one of the 10 values $0, 1, 2, ...9$. Of course, it is possible to use bases other than ten. For example, the Babylonians used the base sixty and thus their digits range between 0 and 59. The computer operates with the base 2 and the two digits 0 and 1. A base 2 number has the form

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

which we again may write in short hand

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

where again n and m are natural numbers, and now each p_i and q_j take the value 0 or 1. For example, in the base two

$$11.101 = 1 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 1 \cdot 2^{-3}.$$

In the floating point arithmetic of a computer using the standard 32 bits, numbers are represented in the form

$$\pm r 2^N,$$

where $1 \leq r \leq 2$ is the *mantissa* and the *exponent* N is an integer. Out of the 32 bits, 23 bits are used to store the mantissa, 8 bits are used to store the exponent and finally one bit is used to store the sign. Since $2^{10} \approx 10^{-3}$ this gives 6 to 7 decimal digits for the mantissa while the exponent N may range from -126 to 127, implying that the absolute value of numbers stored on the computer may range from approximately 10^{-40} to 10^{40}. Numbers outside these ranges cannot be stored by a computer using 32 bits. Some languages permit the use of double precision variables using 64 bits for storage with 11 bits used to store the exponent, giving a range of $-1022 \leq n \leq 1023$, 52 bits used to store the the mantissa, giving about 15 decimal places.

We point out that the finite storage capability of a computer has two effects when storing rational numbers. The first effect is similar to the effect of finite storage on integers, namely only rational numbers within a finite range can be stored. The second effect is more subtle but actually has more serious consequences. This is the fact that numbers are stored only up to a specified number of digits. Any rational number that requires more than the finite number of digits in its decimal expansions, which included all rational numbers with infinite periodic expansions for example, are therefore stored

on a computer with an error. So for example 2/11 is stored as .1818181 or .1818182 depending on whether the computer rounds or not.

But this is not the end of the story. Introduction of an error in the 7'th or 15'th digit would not be so serious except for the fact that such *round-off* errors accumulate when arithmetic operations are performed. In other words, if we add two numbers with a small error, the result may have a larger error being the sum of the individual errors (unless the errors have opposite sign or even cancel).

We give below in Chapter Series an example showing a startling consequence of working with finite decimal representations with round off errors.

189.11 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. 189.4. Her mentor was Gösta Mittag-Leffler (1846-1927), famous Swedish mathematician and founder of the prestigous journal Acta Mathematica, see Fig. 90.34.

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.

FIGURE 189.4. 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 189 Problems

189.1. Plot the functions; (a) 2^{-n}, (b) 5^{-n}, and (c) 10^{-n}; defined on the natural numbers n. Compare the plots.

189.2. Plot the function $f(n) = \frac{10}{9}(1-10^{-n-1})$ defined on the natural numbers.

189.3. 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) $\{125, 25, 5, 1, \frac{1}{5}, \frac{1}{25}, \frac{1}{125}, \cdots\}$.

189.4. Show the following limits hold using the formal definition of the limit:

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

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

189.6. 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.

189.7. Show the following hold using the formal definition for divergence to infinity:

$$\text{(a) } \lim_{n\to\infty} -4n+1 = -\infty \qquad \text{(b) } \lim_{n\to\infty} n^3 + n^2 = \infty.$$

189.8. Show that $\lim_{n\to\infty} r^n = \infty$ for any r with $|r| \geq 2$.

189.9. Find the values of

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

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

(c) $5^{-2} + 5^{-3} + 5^{-4} + \cdots$

189.10. 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) $1 - r + r^2 - r^3 + r^4 - r^5 + \cdots$

189.11. 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) $\left\{\dfrac{2^{\,\mathrm{car}}}{4+(-1)^{\,\mathrm{car}}}\right\}_{\mathrm{car}\,=1}^{\infty}$ (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) $\left\{8\dfrac{2^n}{4+(-1)^{n+3}}\right\}_{n=-2}^{\infty}$.

189.12. 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 $-\infty$.

189.13. Show that (184.14) holds by considering the different cases: $a < 0$, $b < 0$, $a < 0$, $b > 0$, $a > 0$, $b < 0$, $a > 0$, $b > 0$. Show that (189.7) holds using (184.14) and the fact that $a - c + c - b = a - b$.

189.14. 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 $A - B$.

189.15. *(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!

189.16. Compute the limits of the sequences $\{a_n\}_{n=1}^{\infty}$ with the indicated terms or show they diverge.

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

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

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

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

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

189.17. Compute the following limits

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

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

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

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

189.19. 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.

189.20. Determine the number of digits used to store rational numbers in the programming language that you use and whether the language truncates or rounds.

189.21. 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...

190
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)

190.1 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:

$$\begin{aligned} x = 1.&41421356237309504880168872420969807856967187537694807317667973799073247846210703885038753432\\ &76415727350138462309122970249248360558507372126441214970999358314132226659275055927557999505\\ &01152782060571470109559971605970274534596862014728517418640889198609552329230484308714321450\\ &83976260362799525140798968725339654633180882964062061525835239505474575028775996172983557522\\ &03375318570113543746034084988471603868999706 99 \end{aligned}$$

Computing x^2 again using *MAPLE*© , we find that

$$\begin{aligned} x^2 = 1.&99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999999999999999999999999999999999999999999\\ &99999999998638103700279039354754492148156752071936433672239224862717918909878701580996023264059726131264076040569129995030929574783188859695007088740560583365016522715738094455933206900458172642221739359695332425151587602336042729948891418035989710382049561848123333216251601609728313712306449949794365347969862977651606097283137123064499497943653479698629776\\ &68333406657702403185133060024272321251752730435477674866080899878079357977747596458770825031700686 87058548601 0 \end{aligned}$$

The number $x = 1.4142\cdots699$ 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.

190.2 $\sqrt{2}$ Is Not a Rational Number!

In this section, we show that $\sqrt{2}$ cannot be a rational number of the form p/q with p and q natural numbers, and thus the decimal expansion of $\sqrt{2}$ cannot be periodic. In the proof we use the fact that a natural number can be uniquely factored into prime factors. We showed this in chapter Natural Numbers and Integers. One consequence of the factorization into prime numbers is the following fact: Suppose that we know that 2 is a factor of n. If $n = pq$ is a factorization of n into integers p and q, if follows that at least one of the factors p and q must have a factor of 2.

We argue by contradiction. Thus we shall show that assuming that $\sqrt{2}$ is rational leads to a contradiction, and thus $\sqrt{2}$ cannot be rational. We thus assume that $\sqrt{2} = p/q$, where all common factors in the natural numbers p and q have been divided out. For example if p and q both have the factor 3, then we replace p by $p/3$ and q by $q/3$, which does not change the quotient p/q. We write this as $\sqrt{2}q = p$ where p and q have no common factors, or squaring both sides, $2q^2 = p^2$. Since the left hand side contains the factor 2, the right hand side p^2 must contain the factor 2, which means that p must contain the factor 2. Thus we can write $p = 2 \times \bar{p}$ with $\bar{p}$ a natural number. We conclude that $2q^2 = 4 \times \bar{p}^2$, that is $q^2 = 2 \times \bar{p}^2$. But the same argument implies that q must also contain a factor of 2. Thus both p and q contain the factor 2 which contradicts the original assumption that p and q had no common factors. Assuming $\sqrt{2}$ to be rational number thus leads to a contradiction and therefore $\sqrt{2}$ cannot be a rational number.

The argument just given was known to the Pythagoreans, who thus knew that $\sqrt{2}$ is not a rational number. This knowledge caused a lot of trouble. On one hand, $\sqrt{2}$ represents the diagonal of a square of side one, so it seemed that $\sqrt{2}$ had to exist. On the other hand, the Pythagorean school of philosophy was based on the principle that everything could be described in terms of natural numbers. The discovery that $\sqrt{2}$ was not a rational number, that is that $\sqrt{2}$ could not be viewed as a pair of natural numbers, came as a shock! Legend says that the person who discovered the proof was punished by the gods for revealing an imperfection in the universe. The Pythagoreans tried to keep the discovery secret by teaching it only to a select few, but eventually the discovery was revealed and after that the Pythagorean school quickly fell apart. At the same time, the Euclidean school, which was based on geometry instead of numbers, became more influential. Considered from the point of view of geometry, the difficulty with $\sqrt{2}$ seems to "disappear", because no one would question that a square

of side length 1 will have a diagonal of a certain length, and we could then simply define $\sqrt{2}$ to be that length. The Euclidean geometric school took over and ruled all through the Dark Ages until the time of Descartes in the 17th century who resurrected the Pythagorean school based on numbers, in the form of analytical geometry. Since the digital computer of today is based on natural numbers, or rather sequences of $0s$ and $1s$, we may say that Pythagoras ideas are very much alive today: everything can be described in terms of natural numbers. Other Pythagorean dogmas like "never eat beans" and "never pick up anything that has fallen down" have not survived equally well.

190.3 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

$$f(x) = 0 \tag{190.1}$$

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 (190.1) 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 (190.1) 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 (190.1) 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 (190.1) 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 (190.1). 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. 190.1 with $x_0 = 1$ and $X_0 = 2$.

i	x_i	X_i
0	1.00000000000000	2.00000000000000
1	1.00000000000000	1.50000000000000
2	1.25000000000000	1.50000000000000
3	1.37500000000000	1.50000000000000
4	1.37500000000000	1.43750000000000
5	1.40625000000000	1.43750000000000
6	1.40625000000000	1.42187500000000
7	1.41406250000000	1.42187500000000
8	1.41406250000000	1.41796875000000
9	1.41406250000000	1.41601562500000
10	1.41406250000000	1.41503906250000
11	1.41406250000000	1.41455078125000
12	1.41406250000000	1.41430664062500
13	1.41418457031250	1.41430664062500
14	1.41418457031250	1.41424560546875
15	1.41418457031250	1.41421508789062
16	1.41419982910156	1.41421508789062
17	1.41420745849609	1.41421508789062
18	1.41421127319336	1.41421508789062
19	1.41421318054199	1.41421508789062
20	1.41421318054199	1.41421413421631

FIGURE 190.1. 20 steps of the Bisection algorithm.

190.4 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,

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

In other words, the terms x_i either increase or stay constant while the X_i always decrease or remain constant as i increases, and any x_i is smaller than any X_j. Moreover, the choice of the midpoint means that the distance between X_i and x_i is always strictly decreasing as i increases. In fact,

$$0 \le X_i - x_i \le 2^{-i} \quad \text{for } i = 0, 1, 2, \cdots, \tag{190.2}$$

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. 190.1.

The estimate (190.2) on the difference of $X_i - x_i$ also implies that the terms in the sequence $\{x_i\}_{i=0}^{\infty}$ become closer as the index increase. This follows because $x_i \le x_j < X_j \le X_i$ if $j > i$ so (190.2) implies

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

that is

$$|x_i - x_j| \le 2^{-i} \quad \text{if } j \ge i. \tag{190.3}$$

We illustrate in Fig. 190.2. In particular, this means that when $2^{-i} \le 10^{-N-1}$, the first N decimals of x_j are the same as the first N decimals in x_i for any $j \ge i$.

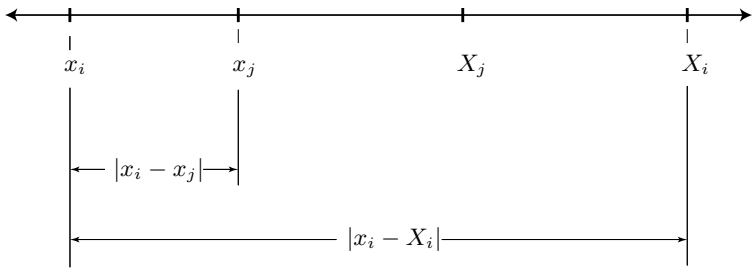


FIGURE 190.2. $|x_i - x_j| \le |X_i - x_i|$.

In other words, as we compute more and more numbers x_i, more and more leading decimals of the numbers x_i agree. We conclude that the sequence $\{x_i\}_{i=0}^{\infty}$ determines a specific (infinite) decimal expansion. To get the first N digits of this expansion, we simply take the first N digits of any number x_j in the sequence with $2^{-j} \le 10^{-N-1}$. By the inequality (190.3), all such x_j agree in the first N digits.

If this infinite decimal expansion was the decimal expansion of a rational number $\bar{x}$, then we would of course have

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

However, we showed above that the decimal expansion defined by the sequence $\{x_i\}_{i=0}^{\infty}$ cannot be periodic. So there is no rational number $\bar{x}$ that can be the limit of the sequence $\{x_i\}$.

We have now come to the point where the Pythagoreans got stuck 2.500 years ago. The sequence $\{x_i\}$ "tries to converge" to a limit, but the limit is not a number of the type we already know, that is a rational number. To avoid the fate of the Pythagoreans, we have to find a way out of this dilemma. The limit appears to be a number of a new kind and thus it appears that we have to somehow extend the rational numbers. The extension will be accomplished by viewing any infinite decimal expansion, periodic or not, as some kind of number, more precisely as a *real number*. In this way, we will clearly get an extension of the set of rational numbers since the rational numbers correspond to periodic decimal expansions. We will refer to non-periodic decimal expansions as *irrational numbers*.

For the extension from rational to real numbers to make sense, we must show that we can compute with irrational numbers in pretty much the same way as with rational numbers. We shall see this is indeed possible and we shall see that the basic idea when computing with irrational numbers is the natural one: compute with truncated decimal expansions! We give the details in the next chapter devoted to a study of real numbers.

Let us now summarize and see where we stand: the Bisection algorithm applied to the equation $x^2 - 2 = 0$ generates a sequence $\{x_i\}_{i=1}^{\infty}$ satisfying

$$|x_i - x_j| \leq 2^{-i} \quad \text{if } j \geq i. \tag{190.4}$$

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

$$\sqrt{2}\sqrt{2} = 2. \tag{190.5}$$

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 (190.5) 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!

190.5 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

$$|x_i - x_j| \leq 2^{-i} \quad \text{if } j \geq i, \tag{190.6}$$

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 (190.6) 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 (190.6) 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 $\lim_{i\to\infty} x_i = 1$.

190.6 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*© m-file implementation of this algorithm in Fig. 190.3.

i	x_i	X_i
0	1.00000000000000	2.00000000000000
1	1.40000000000000	1.50000000000000
2	1.41000000000000	1.42000000000000
3	1.41400000000000	1.41500000000000
4	1.41420000000000	1.41430000000000
5	1.41421000000000	1.41422000000000
6	1.41421300000000	1.41421400000000
7	1.41421350000000	1.41421360000000
8	1.41421356000000	1.41421357000000
9	1.41421356200000	1.41421356300000
10	1.41421356230000	1.41421356240000
11	1.41421356237000	1.41421356238000
12	1.41421356237300	1.41421356237400
13	1.41421356237300	1.41421356237310
14	1.41421356237309	1.41421356237310

FIGURE 190.3. 14 steps of the deca-section algorithm.

By construction

$$|x_i - X_i| \leq 10^{-i},$$

and also

$$|x_i - x_j| \leq 10^{-i} \quad \text{for } j \geq i. \tag{190.7}$$

The inequality (190.7) 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 190 Problems

190.1. Use the *evalf* function in *MAPLE*© to compute $\sqrt{2}$ to 1000 places and then square the result and compare to 2.

190.2. (a) Show that $\sqrt{3}$ (see Problem **??**) 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.

190.3. Specify three different irrational numbers using the digits 3 and 4.

190.4. 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.

190.5. (a) Use the program in Problem 190.4 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.)

190.6. Using the results in Problem 190.4(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.

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

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

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

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

190.11. Compute the cost of the tridiagonal algorithm from Problem 190.10 and compare to the costs of the Bisection and Deca-section methods.

190.12. Use the Bisection code from Problem 190.4 to compute $\sqrt{3}$ (recall Problem **??**). Hint: $1 < \sqrt{3} < 2$.

191
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)

191.1 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

$$\pm p_m \cdots p_0.q_1q_2q_3\cdots$$

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

1.4142135623.... of $\sqrt{2}$ above. The corresponding sequence $\{x_i\}_{i=1}^{\infty}$ of truncated decimal expansions is given by the rational numbers

$$x_i = \pm p_m \cdots p_0.q_1 \cdots q_i = \pm(p_m 10^m + \cdots + q_i 10^{-i}).$$

We have for $j > i$,

$$|x_i - x_j| = |0.0 \cdots 0 q_{i+1} \cdots q_j| \leq 10^{-i}. \tag{191.1}$$

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

$$\lim_{i\to\infty} x_i = x, \tag{191.2}$$

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 (191.2) 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, (191.2) serves to denote the limit $\lim_{i\to\infty} x_i$ by $\pm p_m \cdots p_0.q_1 q_2 q_3 \cdots$.

For the sequence $\{x_i\}_{i=1}^{\infty}$ generated by the Bisection algorithm applied to the equation $x^2 - 2 = 0$, we decided to write $\sqrt{2} = \lim_{i\to\infty} x_i$, and thus

we may write $\sqrt{2} = 1.412...$, with $1.412...$, denoting the infinite decimal expansion given by the Bisection algorithm.

We shall now specify how to compute with real numbers defined in this way. In particular, we shall specify how to add, subtract, multiply and divide real numbers. Of course we will do this so that it extends our experience in computing with rational numbers. This will complete our process of extending the natural numbers to obtain first the integers and then the rational numbers, and finally the real numbers.

We denote by $\mathbb{R}$ the set of all possible real numbers, that is the set of all possible infinite decimal expansions. We discuss this definition in Chapter Do Mathematicians Quarrel? below.

191.2 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

$$\bar{x} = \frac{1043}{439} = 2.37585421412300683 37\cdots,$$

we compute $y_i = x_i + \bar{x}_i$ for $i = 1, 2, ...$, which defines the decimal expansion of $x + \bar{x}$, see Fig. 191.1. We may notice that occasionally adding two digits

i	$x_i + \bar{x}_i$
1	3
2	3.7
3	3.78
4	3.789
5	3.7900^*
6	3.79006
7	3.790067
8	3.7900677
9	3.79006777
10	3.790067776
11	3.7900677764
12	3.79006777649
13	3.790067776496
14	3.7900677764960
15	3.79006777649609
16	3.790067776496101^*
17	3.7900677764961018
18	3.79006777649610187
19	3.790067776496101881^*
20	3.7900677764961018825^*
⋮	⋮

FIGURE 191.1. Computing the decimal expansion of $\sqrt{2} + 1043/439$ by using the truncated decimal sequences. Note the changes in the digits marked by the $*$ where adding the new digits affects previous digits.

affects the digits to the left, as in $0.9999 + 0.0001 = 1.000$.

Similarly, the difference $x - \bar{x}$ of two real numbers $x = \lim_{i\to\infty} x_i$ and $\bar{x} = \lim_{i\to\infty} \bar{x}_i$ is of course defined by

$$x - \bar{x} = \lim_{i\to\infty} (x_i - \bar{x}_i).$$

191.3 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

$$|f(x_i,\bar{x}_i) - f(x_j,\bar{x}_j)| \le L(|x_i - x_j| + |\bar{x}_i - \bar{x}_j|)$$

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

$$\lim_{i\to\infty} x_i + \lim_{i\to\infty} \bar{x}_i = \lim_{i\to\infty} (x_i + \bar{x}_i). \tag{191.3}$$

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 $i = 1, 2, ...$

191.4 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

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

191.5 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 (191.1) we have $|x_j - x_i| \leq 10^{-i}$ for $j > i$, and thus taking the limit as j tends to infinity,

$$|x - x_i| \leq 10^{-i} \quad \text{for } i = 1, 2, ... \tag{191.4}$$

191.6 Comparing Two Real Numbers

Let $x = \lim_{i\to\infty} x_i$ and $\bar{x} = \lim_{i\to\infty} \bar{x}_i$ be two real numbers with corresponding sequences of truncated decimal expansions $\{x_i\}_{i=1}^{\infty}$ and $\{\bar{x}_i\}_{i=1}^{\infty}$. How can we tell if $x = \bar{x}$? Is it necessary that $x_i = \bar{x}_i$ for all i? Not quite. For example, consider the two numbers $x = 0.99999\cdots$ and $\bar{x} = 1.0000\cdots$. In fact, it is natural to give a little more freedom and say that $x = \bar{x}$ if and only if

$$|x_i - \bar{x}_i| \leq 10^{-i} \quad \text{for } i = 1, 2, ... \tag{191.5}$$

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 (191.5) 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 $|x_i - x_j| \le 10^{-i}$ 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

$$x - \bar{x} \ge \epsilon \quad \text{or} \quad x - \bar{x} \le -\epsilon.$$

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 $x > 0$.

191.7 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

$$x + \bar{x} = \lim_{i\to\infty} (x_i + \bar{x}_i) = \lim_{i\to\infty} (\bar{x}_i + x_i) = \bar{x} + x.$$

191.8 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

$$\sqrt{2} = \lim_{i\to\infty} x_i, \tag{191.6}$$

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

$$\sqrt{2}\sqrt{2} = \lim_{i\to\infty} x_i^2, \tag{191.7}$$

and we thus need to show that

$$\lim_{i\to\infty} x_i^2 = 2 \tag{191.8}$$

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

$$|(x_i)^2 - (X_i)^2| \le 4|x_i - X_i| \le 2^{-i+2}.$$

where we use the inequality $|x_i - X_i| \le 2^{-i}$. By construction $x_i^2 < 2 < X_i^2$, and thus in fact

$$|x_i^2 - 2| \le 2^{-i+2}$$

which shows that

$$\lim_{i\to\infty} (x_i)^2 = 2$$

and (191.8) 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 $x^2 = 2$.

191.9 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.

191.10 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

$$|x_i - x_j| \leq \epsilon \quad \text{for} \quad i, j \geq N. \tag{191.9}$$

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 191.1 *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 $|x_i - x_j| \leq 2^{-i}$ for $j \geq i$.

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 191.2 *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.

191.11 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

$$f(x) = \lim_{i\to\infty} f(x_i). \tag{191.10}$$

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:

$$f(\lim_{i\to\infty} x_i) = \lim_{i\to\infty} f(x_i), \tag{191.11}$$

for any convergent sequence $\{x_i\}$ in J (with automatically $\lim_{i\to\infty} x_i \in J$ because J is closed).

191.12 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 191.3 *A Lipschitz continuous function $f : I \to \mathbb{R}$, where $I = [a, b]$ is an interval of real numbers, satisfies:*

$$f(\lim_{i\to\infty} x_i) = \lim_{i\to\infty} f(x_i), \tag{191.12}$$

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 191.1. 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 191.2. 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 $n < 0$, $x \neq 0$.

191.13 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 $f : \mathbb{Q} \to \mathbb{Q}$.

191.14 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 191.3. Consider the quotient

$$\frac{x-1}{x-1}$$

with domain $\{x \in \mathbb{R} : x \neq 1\}$. Since

$$x - 1 = 1 \times (x - 1) \tag{191.13}$$

for all x, it is natural to "divide" the polynomials to get

$$\frac{x-1}{x-1} = 1. \tag{191.14}$$

However, the domain of the constant function 1 is $\mathbb{R}$ so the left- and right-hand sides of (191.14) have different domains and therefore must represent different functions. We plot the two functions in Fig. 191.2. We see that the two functions agree at every point except for the "miss-

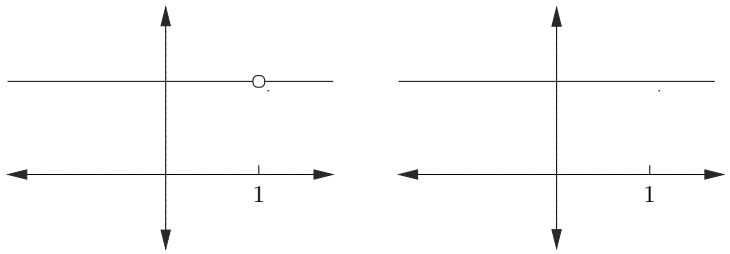


FIGURE 191.2. Plots of $(x-1)/(x-1)$ on the left and 1 on the right.

ing" point $x = 1$.

EXAMPLE 191.4. 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 191.5. 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 $x = 1$.

191.15 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. 191.3. The word "open" refers to

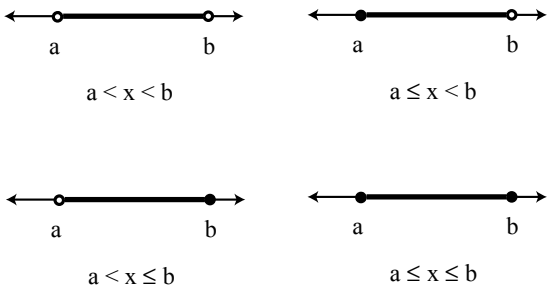


FIGURE 191.3. Intervals corresponding to the real numbers between two real numbers a and b. Note the use of a solid and closed circles in the four cases.

the strict inequality defining (a, b) and we use the curved parentheses "(" and the open circle on the number line to mark this. a and b are called the *endpoints* of the interval. An open interval does not contain its endpoints. The *closed interval* $[a, b]$ is the set $\{x : a \leq x \leq b\}$ and is denoted on the number line using solid circles. Note the use of square parentheses "[" when the inequalities are not strict. A closed interval does contain its endpoints. Finally, we can have *half-open intervals* with one end open and the other closed, such as $(a, b] = \{x : a < x \leq b\}$. See Fig. 191.3.

We also have "infinite" intervals such as $(-\infty, a) = \{x : x < a\}$ and $[b, \infty) = \{x : b \leq x\}$. We illustrate these in Fig. 191.4. With this notation,

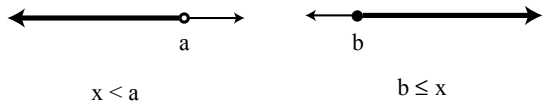


FIGURE 191.4. Infinite intervals $(-\infty, a)$ and $[b, \infty)$.

we denote the set of real numbers by $\mathbb{R} = (-\infty, \infty)$.

Clearly, we can now consider Lipschitz continuous functions $f : I \to \mathbb{R}$ defined on intervals I of $\mathbb{R}$.

191.16 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. 191.5. This means that the conventional idea of viewing $f(x)$ as a function of x comes into a new light. In the traditional way of thinking of a function $f(x)$, we think of x as given and then associating the value $f(x)$ to x. We may even write this as $x \to f(x)$ indicating that we go *from* x *to* $f(x)$.

However, we just noticed that when x is irrational, we cannot start from knowing all the decimals of x, and then determine $f(x)$. Instead, we determine successively the decimal expansions x_i and the corresponding function values $f(x_i)$, that is, we may write $x_i \to f(x_i)$ for $i = 1, 2, ...$, but not really $x \to f(x)$. We rather jump back and forth between approximations x_i of x and approximations $f(x_i)$ of $f(x)$. This means that we do not have exact knowledge of x when we compute $f(x)$. In order to make this process to be meaningful, we need the function $f(x)$ to be Lipschitz continuous. In this case, small changes in x cause small changes in $f(x)$, and the extension process is possible.

EXAMPLE 191.6. We evaluate $f(x) = .4x^3 - x$ for $x = \sqrt{2}$ using the truncated decimal sequence $\{x_i\}$ in Fig. 191.5.

i	x_i	$.4x_i^3 - x_i$
1	1	$-.6$
2	1.4	.0976
3	1.41	.1212884
4	1.414	.1308583776
5	1.4142	.1313383005152
6	1.41421	.1313623002245844
7	1.414213	.1313695002035846388
8	1.4142135	.13137070020305452415
9	1.41421356	.131370844203047931474406 4
10	1.414213562	.131370849003047922153528 1312
⋮	⋮	⋮

FIGURE 191.5. 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

$$f(x) = 0 \quad \text{if } x \quad \text{is rational,}$$
$$f(x) = 1 \quad \text{if } x \quad \text{is irrational.}$$

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" $f(x)$

$$f(x) = a \quad \text{if } x < \bar{x},$$
$$f(x) = b \quad \text{if } x \geq \bar{x},$$

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. 191.6.

It thus seems that we have to reject the very idea that a function $f(x)$ can be discontinuous. This is because we cannot assume that we know x exactly, and thus we can only handle a situation where small changes in x causes small changes in $f(x)$, which is the essence of Lipschitz continuity. Instead we are led to handle functions with jumps as combinations of Lipschitz continuous functions with two possible values at the jumps, one value from the right and another value from the left.

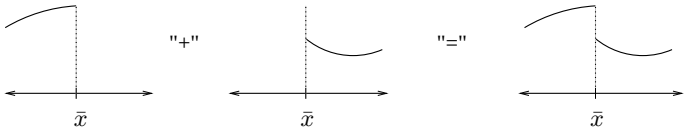


FIGURE 191.6. A "jump function" viewed as two functions

191.17 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 $f : I \to \mathbb{R}$:

$$f(\lim_{i\to\infty} x_i) = \lim_{i\to\infty} f(x_i), \tag{191.15}$$

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 (191.15), and reads as follows: The function $f : I \to \mathbb{R}$ is said to be *continuous* on I (according to the standard definition) if (191.15) 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 191 Problems

191.1. 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.

191.2. 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 (184.14) and (191.4), 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 (191.4) implies that for i sufficiently big, $|x_i| \leq |x| + 1$.

191.3. 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 $|xy - x_iy_j| \leq 10^{-6}$

191.4. 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. 191.1.

191.5. Let x be the limit of the sequence $\left\{\dfrac{i}{i+1}\right\}$. Is $x < 1$?. Give a reason for your answer.

191.6. 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.

191.7. 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 $-x > -y$.

191.8. Find the set of x that satisfies (a) $|\sqrt{2}x - 3| \leq 7$ and (b) $|3x - 6\sqrt{2}| > 2$.

191.9. Verify that the triangle inequality (184.14) extends to real numbers s and t.

191.10. *(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.

191.11. Show that the following sequences are Cauchy sequences.

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

191.12. Show that the sequence $\{i^2\}$ is **not** a Cauchy sequence.

191.13. Let $\{x_i\}$ denote the sequence of real numbers defined by

$$\begin{aligned} x_1 &= .373373337\cdots \\ x_2 &= .337733377333377\cdots \\ x_3 &= .333777333377733333777\cdots \\ x_4 &= .33337777333337777333337777\cdots \\ &\vdots \end{aligned}$$

(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.

191.14. Can a number of the form $sx+t$, with s and t rational and x irrational, be rational?

191.15. 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.

191.16. 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.

191.17. *(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

191.18. 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. 191.5 and use the *evalf* function of *MAPLE*© in order to determine all the digits.

191.19. 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 $\{f(x_i)\}$?

191.20. Show that $|x|$ is Lipschitz continuous on the real numbers $\mathbb{R}$.

191.21. 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 191.3. Hint: Use the identity

$$\begin{aligned} x_2^n - x_1^n &= (x_2 - x_1)\big(x_2^{n-1} + x_2^{n-2}x_1 + x_2^{n-3}x_1^2 \\ &\qquad + \cdots + x_2^2 x_1^{n-3} + x_2 x_1^{n-2} + x_1^{n-1}\big) \\ &= (x_2 - x_1)\sum_{j=0}^{n-1} x_2^{n-1-j}\, x_1^j \end{aligned}$$

after showing that it is true. Note there are n terms in the last sum, the Lipschitz constant definitely depends on n.

191.22. Show Theorem 191.3 is true.

191.23. 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) $\{x : x < 0\} \cup \{x : x > 1\}$

191.24. Produce an interval that contains all the points $3 - 10^{-j}$ for $j \geq 0$ but does not contain 3.

191.25. 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 $y = \sqrt{2} \times x$.

191.26. (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 $[a, b]$.

191.27. 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.

191.28. 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))

192
The Bisection Algorithm for $f(x) = 0$

Divide ut regnes (divide and conquer). (Machiavelli 1469-1527)

192.1 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

$$f(x) = 0 \tag{192.1}$$

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 192.1 (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 192.2 (The Intermediate Value Theorem) *If $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous then for any real number y in the interval between $f(a)$ and $f(b)$, there is a real number $x \in [a, b]$ such that $f(x) = y$.*

FIGURE 192.1. 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".

192.2 An Example

As an application of the Bisection algorithm, we compute the roots of the chemical equilibrium equation (184.13) in Chapter *Rational Numbers*,

$$S\,(.02+2S)^2 - 1.57\times 10^{-9} = 0. \tag{192.2}$$

We show a plot of the function involved in Fig. 192.2. Apparently there

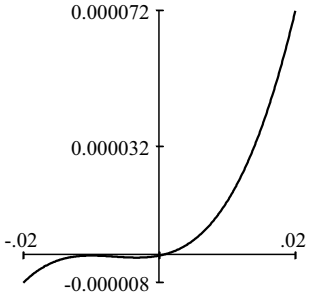


FIGURE 192.2. A plot of the function $S\,(.02+2S)^2 - 1.57\times 10^{-9}$

are roots near $-.01$ and 0, but to compute them it seems advisable to first rescale the equation. We then first multiply both sides of (192.2) by 10^9 to get

$$10^9 \times S\,(.02+2S)^2 - 1.57 = 0,$$

and write

$$10^9 \times S\,(.02 + 2S)^2 = 10^3 \times S \times 10^6 \times (.02 + 2S)^2$$
$$= 10^3 \times S \times \left(10^3\right)^2 \times (.02 + 2S)^2 = 10^3 \times S \times \left(10^3 \times (.02 + 2S)\right)^2$$
$$= 10^3 \times S \times \left(20 + 2 \times 10^3 \times S\right)^2.$$

If we name a new variable $x = 10^3 S$, then we obtain the following equation to solve

$$f(x) = x(20 + 2x)^2 - 1.57 = 0. \tag{192.3}$$

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. 192.3. It appears that $f(x) = 0$

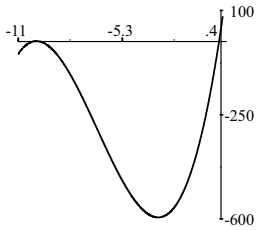


FIGURE 192.3. A plot of the function $f(x) = x(20 + 2x)^2 - 1.57$

might have one root near 0 and another root near -10.

To compute a positive root, we now choose $x_0 = -.1$ and $X_0 = .1$ and apply the Bisection algorithm for 20 steps. We show the results in Fig. 192.4. This suggests that the root of (192.3) is $x \approx .00392$ or $S \approx 3.92 \times 10^{-6}$.

192.3 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

$$N_b(C_m + C_f + C_\pm + 4C_s). \tag{192.4}$$

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

$$N_d(9C_m + 9C_f + 9C_{\pm} + 20C_s). \tag{192.5}$$

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 (192.5). This means that the Bisection algorithm is cheaper to use than the Deca-section algorithm.

i	x_i	X_i
0	-0.10000000000000	0.10000000000000
1	0.00000000000000	0.10000000000000
2	0.00000000000000	0.05000000000000
3	0.00000000000000	0.02500000000000
4	0.00000000000000	0.01250000000000
5	0.00000000000000	0.00625000000000
6	0.00312500000000	0.00625000000000
7	0.00312500000000	0.00468750000000
8	0.00390625000000	0.00468750000000
9	0.00390625000000	0.00429687500000
10	0.00390625000000	0.00410156250000
11	0.00390625000000	0.00400390625000
12	0.00390625000000	0.00395507812500
13	0.00390625000000	0.00393066406250
14	0.00391845703125	0.00393066406250
15	0.00391845703125	0.00392456054688
16	0.00392150878906	0.00392456054688
17	0.00392150878906	0.00392303466797
18	0.00392150878906	0.00392227172852
19	0.00392189025879	0.00392227172852
20	0.00392189025879	0.00392208099365

FIGURE 192.4. 20 steps of the Bisection algorithm applied to (192.3) using $x_0 = -.1$ and $X_0 = .1$.

193
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)

193.1 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, see Fig. 193.2. Turing laid out plans for an electronic computer named Analytical Computing Engine ACE, with reference to Babbages' Analytical Engine, at the same time as the ENIAC was designed in the US.

FIGURE 193.1. Kurt Gödel (with Einstein 1950): "Every formal system is incomplete"

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.

FIGURE 193.2. Alan Turing: "I wonder if my machine will come to a halt?".

193.2 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.

193.3 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

FIGURE 193.3. Bertrand Russell: "I am protesting"

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 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. 193.4.

FIGURE 193.4. "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. **??**.

193.4 The Constructivists

The *intuitionist/constructivist* view is to consider the basic concepts to have a meaning which may be directly "intuitively" understood by our brains and bodies through experience, without any further explanation. Furthermore, the intuitionists would like to use as concrete or "constructive" arguments as possible, in order for their mathematics always to have an intuitive "real" meaning and not just be a formality like a game.

An intuitionist may say that the natural numbers 1, 2, 3,...., are obtained by repeatedly adding 1 starting at 1. We took this standpoint when introducing the natural numbers. We know that from the constructivist point of view, the natural numbers are something in the state of being created in a process without end. Given a natural number n, there is always a next natural number $n+1$ and the process never stops. A constructivist would not speak of the set of all natural numbers as something having been completed and constituting an entity in itself, like the set of all natural numbers as a formalist or logicist would be willing to do. Gauss pointed out that "the set of natural numbers" rather would reflect a "mode of speaking" than existence as a set.

An intuitionist would not feel a need of "justification" or a proof of consistency through some extra arguments, but would say that the justification is built into the very process of developing mathematics using constructive processes. A constructivist would so to speak build a machine that could fly (an airplane) and that very constructive process would itself be a proof of the claim that building an airplane would be possible. A constructivist is thus in spirit close to a practicing engineer. A formalist would not actually build an airplane, rather make some model of an airplane, and would then need some type of argument to convince investors and passengers that his airplane would actually be able to fly, at least in principle. The leader of the intuitionist school was Brouwer (1881-1967),see Fig. 193.5. Hard-core constructivism makes life very difficult (like strong vegetarianism), and because the Brouwer school of constructivists were rather fundamentalist in their spirit, they were quickly marginalized and lost influence in the 1930s. The quote by Courant given above shows the strong feelings involved related to the fact that very fundamental dogmas were at stake, and the general lack of rational arguments to meet the criticism from the intuitionists, which was often replaced by ridicule and oppression.

Van der Waerden, mathematician who studied at Amsterdam from 1919 to 1923 wrote: "Brouwer came [to the university] to give his courses but

FIGURE 193.5. 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".

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".

193.5 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?

193.6 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.

193.7 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 $1, 2, 3, 4, ...$, would "exist" as an object in itself as *the set of all natural numbers* which could be grasped by human brains, while Kronecker would deny such a possibility and reserve it to a higher being. Of

course, Kronecker did not claim that there are only finitely many natural numbers or that there is a largest natural number, but he would (following Aristotle) say that the existence of arbitrarily large natural numbers is like a "potential" rather than an actual reality.

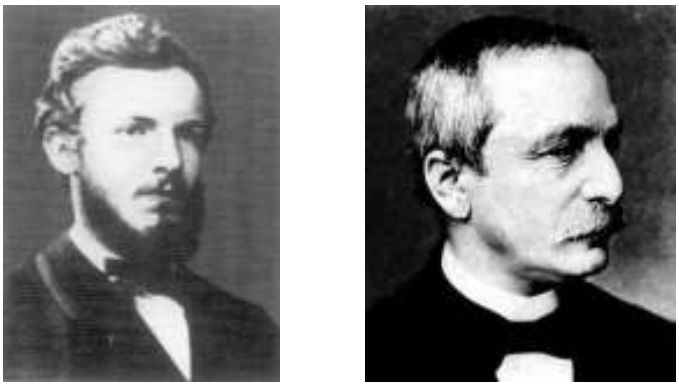


FIGURE 193.6. Cantor (left): "I realize that in this undertaking I place myself in a certain opposition to views widely held concerning the mathematical infinite and to opinions frequently defended on the nature of numbers". Kronecker (right): "God created the integers, all else is the work of man".

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.

193.8 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.

193.9 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.

193.10 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.

193.11 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.

193.12 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.

193.13 A Non-Constructive Proof

We now give an example of a proof with non-constructive aspects that plays an important role in many Calculus books. Although because of the non-constructive aspects, the proof is considered to be so difficult that it can only by appreciated by selected math majors.

The setting is the following: We consider a bounded increasing sequence $\{a_n\}_1^\infty$ of real numbers, that is $a_n \leq a_{n+1}$ for $n = 1, 2, ...$, and there is s constant C such that $a_n \leq C$ for $n = 1, 2, ...$ The claim is that the sequence $\{a_n\}_1^\infty$ converges to a limit A. The proof goes as follows: all the numbers a_n clearly belong to the interval $I = [a_1, C]$. For simplicity suppose $a_1 = 0$ and $C = 1$. Divide now the interval $[0, 1]$ into the two intervals $[0, 1/2]$ and $[1/2, 1]$. and make the following choice: if there is a real number a_n such that $a_n \in [1/2, 1]$, then choose the right interval $[1/2, 1]$ and if not choose the left interval $[0, 1/2]$. Then repeat the subdivision into a left and a right interval, choose one of the intervals following the same principle: if there is a real number a_n in the right interval, then choose this interval, and if not choose the left interval. We then get a nested sequence of intervals with length tending to zero defining a unique real number that is easily seen to be the limit of the sequence $\{a_n\}_1^\infty$. Are you convinced? If not, you must be a constructivist.

So where is the hook of non-constructiveness in this proof? Of course, it concerns the choice of interval: in order to choose the correct interval you must be able to check if there is some a_n that belongs to the right interval, that is you must check if a_n belongs to the right interval for all sufficiently large n. The question from a constructivist point of view is if we can perform each check in a finite number of steps. Well, this may depend on the particular sequence $a_n \leq a_{n+1}$ under consideration. Let's first consider a sequence which is so simple that we may say that we know everything of interest: for example the sequence $\{a_n\}_1^\infty$ with $a_n = 1 - 2^{-n}$, that is the sequence $1/2, 3/4, 7/8, 15/16, 31/32, ...$, which is a bounded increasing

sequence clearly converging to 1. For this sequence, we would be able to always choose the correct interval (the right one) because of its simplicity.

We now consider the sequence $\{a_n\}_1^\infty$ with $a_n = \sum_1^n \frac{1}{k^2}$, which is clearly an increasing sequence, and one can also quite easily show that the sequence is bounded. In this case the choice of interval is much more tricky, and it is not clear how to make the choice constructively without actually constructing the limit. So there we stand, and we may question the value of the non-constructive proof of existence of a limit, if we anyway have to construct the limit.

At any rate we sum up in the following result that we will use a a couple of times below.

Theorem 193.1 (non-constructive!) *A bounded increasing sequence converges.*

193.14 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 193 Problems

193.1. Can you figure out how the barber's paradox is constructed? Suppose the barber comes from another village. Does this resolve the paradox?

193.2. 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?

193.3. 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.

193.4. 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.

193.5. 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?

193.6. 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)

FIGURE 193.7. View of the river Cam at Cambridge 2003 with ACE in the fore-ground (and "UNTHINKABLE" in the background to the right)

194
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).

194.1 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 $x \in \mathbb{R}^+$,

$$(\sqrt{x})^2 = x. \tag{194.1}$$

We plot the function $\sqrt{x}$ in Fig. 194.1.

The function $y = \sqrt{x}$ is increasing: if $x > \bar{x}$, then $\sqrt{x} > \sqrt{\bar{x}}$. Further, if $\{x_i\}$ is a sequence of positive real numbers with $\lim_{i\to\infty} x_i = 0$, then

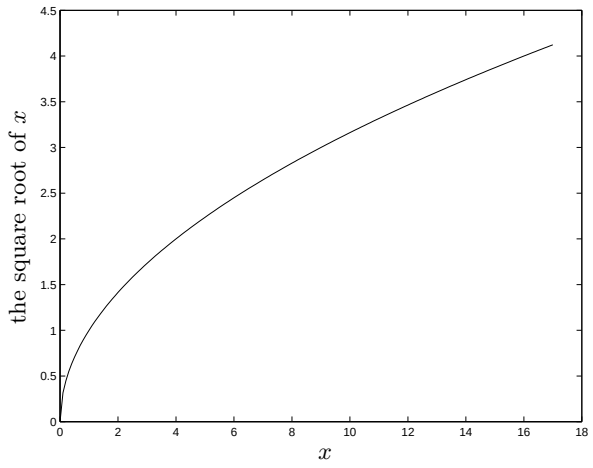


FIGURE 194.1. The function $\sqrt{x}$ of x.

obviously $\lim_{i\to\infty} \sqrt{x_i} = 0$, that is

$$\lim_{x\to 0^+} \sqrt{x} = 0. \tag{194.2}$$

194.2 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,

$$\sqrt{a}\sqrt{b} = \sqrt{ab}. \tag{194.3}$$

We find similarly that

$$\frac{\sqrt{a}}{\sqrt{b}} = \sqrt{\frac{a}{b}}. \tag{194.4}$$

194.3 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 $\frac{1}{2\delta}$.

194.4 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, $\sqrt{x} = x^{\frac{1}{2}}$.

194.5 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 $r, s \in \mathbb{Q}$,

$$x^r x^s = x^{r+s}, \quad \frac{x^r}{x^s} = x^{r-s}. \tag{194.5}$$

194.6 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

$$|\sqrt{x} - \sqrt{\bar{x}}| \le |x - \bar{x}|^{1/2}. \tag{194.6}$$

To prove this estimate, we assume that $x > \bar{x}$ and compute backwards, starting with $\bar{x} \le \sqrt{x}\sqrt{\bar{x}}$ to get $x + \bar{x} - 2\sqrt{x}\sqrt{\bar{x}} \le x - \bar{x} = |x - \bar{x}|$, which can be written

$$(\sqrt{x} - \sqrt{\bar{x}})^2 \le (|x - \bar{x}|^{1/2})^2$$

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!

194.7 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 194 Problems

194.1. 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) $(xy)^r = x^r y^s$

194.2. Is $f(x) = \sqrt[3]{x}$, Lipschitz continuous on $(0,\infty)$ in the generalized sense? If yes give then the Lipschitz constant and exponent.

194.3. A Lipschitz continuous function with a Lipschitz constant L with $0 \le L < 1$ 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) $f(x) = x^3$

194.4. Let $f(x) = 1$, for $x \le 0$, and $f(x) = \sqrt{1 + x^2}$, for $x > 0$. Is f a contraction mapping?

195

Fixed Points and Contraction Mappings

Give me one fixed point on which to stand, and I will move the Earth.
(Archimedes)

195.1 Introduction

A special case of the basic problem of solving an algebraic equation $f(x) = 0$ takes the form: find $\bar{x}$ such that

$$\bar{x} = g(\bar{x}), \tag{195.1}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a given Lipschitz continuous function. The equation (195.1) 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. 195.1.

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 (195.1) that is of central importance in mathematics. This is the *Fixed Point Iteration* algorithm, which takes the following form: Starting with

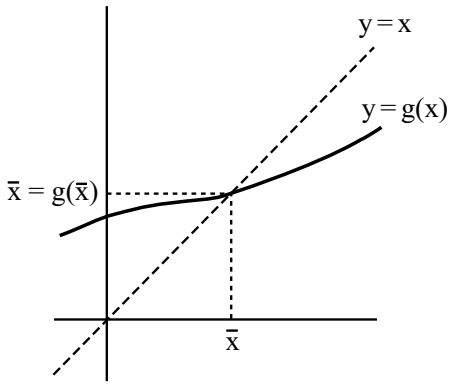


FIGURE 195.1. Illustration of a fixed point problem $g(\bar{x}) = \bar{x}$.

some x_0, for $i = 1, 2, ...$, compute

$$x_i = g(x_{i-1}) \quad \text{for } i = 1, 2, 3, ... \tag{195.2}$$

In words, we start with an initial approximation x_0 then compute $x_1 = g(x_0)$, $x_2 = g(x_1)$, $x_3 = g(x_2)$, and so on. Stepwise, given a current value x_{i-1}, we compute the corresponding output $g(x_{i-1})$, and then choose as new input $x_i = g(x_{i-1})$. Repeating this procedure, we will generate a sequence $\{x_i\}_{i=1}^{\infty}$.

We shall below study the following basic questions related to the sequence $\{x_i\}_{i=1}^{\infty}$ generated by Fixed Point Iteration:

- Does $\{x_i\}_{i=1}^{\infty}$ converge, that is does $\bar{x} = \lim_{i\to\infty} x_i$ exist?
- Is $\bar{x} = \lim_{i\to\infty} x_i$ a fixed point of $y = g(x)$, that is $\bar{x} = g(\bar{x})$?

We shall also investigate whether or not a fixed point $\bar{x}$ is uniquely determined.

195.2 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.

$$|g(x) - g(y)| \le L|x - y| \quad \text{for all } x, y \in \mathbb{R}, \tag{195.3}$$

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 (195.3) 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.

195.3 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 $x > y$,

$$g(x) - g(y) = x + \alpha f(x) - (y + \alpha f(y)) = x - y + \alpha(f(x) - f(y))$$
$$= (1 + \alpha\frac{f(x) - f(y)}{x - y})|x - y|,$$

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. 195.2.

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.

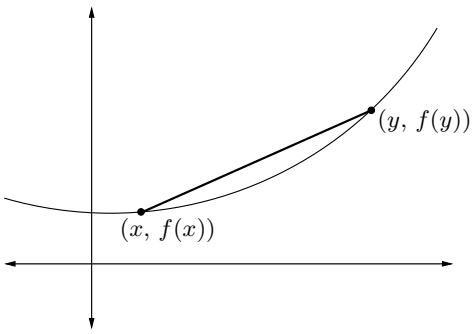


FIGURE 195.2. Corda connecting the points $(x, f(x))$ and $(y, f(y))$ in $\mathbb{R}^2$.

195.4 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

$$g(x) = \frac{1}{4} + \frac{1}{4}x \tag{195.4}$$

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

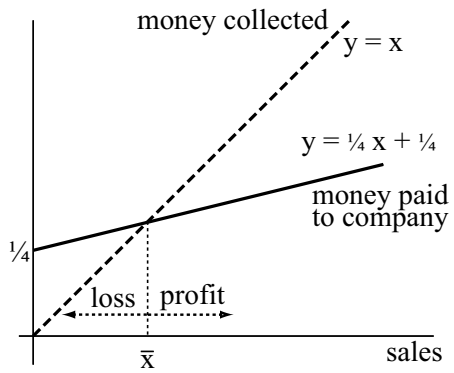


FIGURE 195.3. Illustration of the problem of determining the break-even point for selling greeting cards door-to-door. Sales above the break-even point $\bar{x}$ give a profit to the salesman, but sales below this point mean a loss.

195.5 Private Economy Model

Your roommate has formulated the following model for her/his private economy: denote the net income by x that is variable including contributions from family, fellowship and a temporary job at McDonalds. The spending consists a fixed amount of 1 unit (of say 500 dollars per month) for rent and insurance, the variable amount of $x/2$ units for good food, good books and intellectual movies, and a variable amount of $1/x$ units for junk food, cigarettes and bad movies. This model is based on the observation that the more money your roommate has, the more educated a life she/he will live. The total spending is thus

$$g(x) = \frac{x}{2} + 1 + \frac{1}{x}$$

and the pertinent question is to find a balance of income and spending, that is to find the income $\bar{x}$ such that $\bar{x} = g(\bar{x})$ where the spending is the same as the income. If the income is bigger than $\bar{x}$, then your roommate will not use up all the money, which is against her/his nature, and if the income is less than $\bar{x}$, then your roommate's father will get upset, because he will have to pay the resulting debt.

Also in this case, we can directly find the fixed point $\bar{x}$ by solving the equation

$$\bar{x} = \frac{\bar{x}}{2} + 1 + \frac{1}{\bar{x}}$$

analytically and we then find that $\bar{x} = 1 + \sqrt{3} \approx 2.73$.

If we don't have enough motivation to go through the details of this calculation, we could instead try the Fixed Point Iteration. We would then start with an income $x_0 = 1$ say and compute the spending $g(1) = 2.5$,

then choose the new income $x_1 = 2.5$, and compute the spending $g(x_1) = g(2.5) = 2.65$, and then set $x_2 = 2.65$ and compute the spending $g(x_2) = ...$ and so on. Of course, we expect that $\lim_i x_i = \bar{x} = 1 + \sqrt{3}$. Below, we will prove that this is indeed true!

195.6 Fixed Point Iteration in the Card Sales Model

We now apply Fixed Point Iteration to the Card Sales Model. In Fig. 195.4, 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.

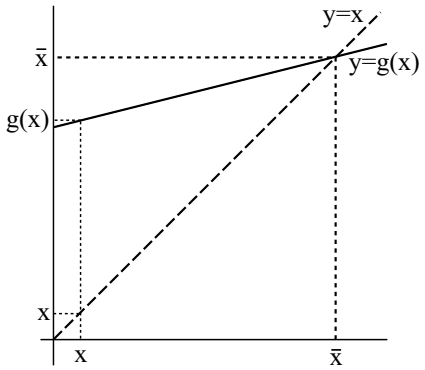


FIGURE 195.4. The first step of Fixed Point Iteration in Card Sales model: $g(x)$ is closer to $\bar{x}$ than x.

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,

$$|x_2 - \bar{x}| = \frac{1}{4}|x_1 - \bar{x}| = \frac{1}{16}|x_0 - \bar{x}|$$

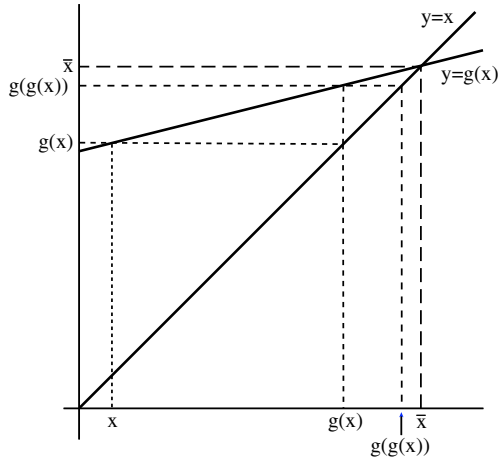


FIGURE 195.5. Two steps of the contraction map algorithm applied to the fixed point problem in Model 195.4. The distance of $g(g(x))$ to $\bar{x}$ is $1/4$ the distance from $g(x)$ to $\bar{x}$ and $1/16$ the distance from x to $\bar{x}$.

We illustrate this in Fig. 195.5. Generally, we have

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

and thus for $i = 1, 2, ...,$

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

Since 4^{-i} gets as small as we please if i is sufficiently large, this estimate shows that Fixed Point Iteration applied to the Card Sales model converges, that is $\lim_{i\to\infty} x_i = \bar{x}$.

We consider some more examples before getting into the question of convergence of Fixed Point Iteration in a more general case.

EXAMPLE 195.1. For the sake of comparison, we show the results for the fixed point problem in Model 195.4 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. 195.6. 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

i	Bisection Algorithm X_i	Fixed Point Iteration x_i
0	1.00000000000000	1.00000000000000
1	0.50000000000000	0.50000000000000
2	0.50000000000000	0.37500000000000
3	0.37500000000000	0.34375000000000
4	0.37500000000000	0.33593750000000
5	0.34375000000000	0.33398437500000
6	0.34375000000000	0.33349609375000
7	0.33593750000000	0.33337402343750
8	0.33593750000000	
9	0.33398437500000	
10	0.33398437500000	
11	0.33349609375000	
12	0.33349609375000	
13	0.33337402343750	

FIGURE 195.6. Results of the bisection algorithm and the fixed point iteration used to solve the fixed point problem in Model 195.4. The error of the fixed point iteration decreases more for each iteration.

additional sign check, the fixed point iteration costs less per iteration. We conclude that the fixed point iteration is truly "faster" than the bisection algorithm for this problem.

EXAMPLE 195.2. In solving for the solubility of $Ba(IO_3)_2$ in Model 184.10, we solved the root problem (192.3)

$$x(20 + 2x)^2 - 1.57 = 0$$

using the bisection algorithm. The results are in Fig. 192.4. In this example, we use the fixed point iteration to solve the equivalent fixed point problem

$$g(x) = \frac{1.57}{(20 + 2x)^2} = x. \tag{195.5}$$

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. 195.7

EXAMPLE 195.3. In the case of the fixed point iteration applied to the Card Sales model, we can compute the iterates explicitly:

$$x_1 = \frac{1}{4}x_0 + \frac{1}{4}$$

i	x_i
0	1.00000000000000
1	0.00484567901235
2	0.00392880662465
3	0.00392808593169
4	0.00392808536527
5	0.00392808536483

FIGURE 195.7. Results of the fixed point iteration applied to (195.5).

and

$$x_2 = \frac{1}{4}x_1 + \frac{1}{4} = \frac{1}{4}\left(\frac{1}{4}x_0 + \frac{1}{4}\right) + \frac{1}{4}$$
$$= \frac{1}{4^2}x_0 + \frac{1}{4^2} + \frac{1}{4}$$

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

$$x_n = \frac{1}{4^n}x_0 + \sum_{i=1}^{n}\frac{1}{4^i}. \tag{195.6}$$

The first term on the right-hand side of (195.6), $\frac{1}{4^n}x_0$ 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 (195.4), 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 (195.6) 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 (195.6) will not converge. We demonstrate this graphically in Fig. 195.8 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

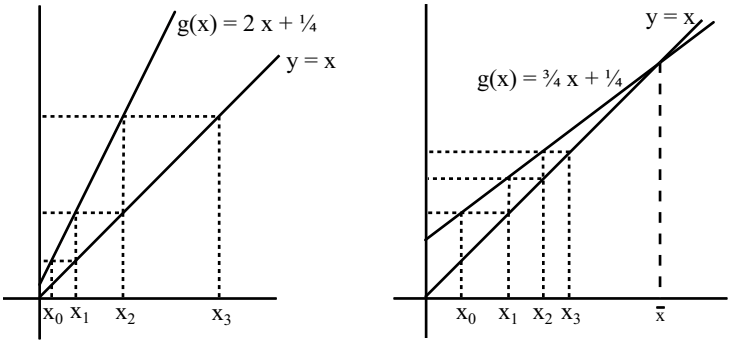


FIGURE 195.8. On the left, we plot the first three fixed point iterates for $g(x) = 2x + \frac{1}{4}$. The iterates increase without bound as the iteration proceeds. On the right, we plot the first three fixed point iterates for $g(x) = \frac{3}{4}x + \frac{1}{4}$. The iteration converges to the fixed point in this case.

the convergence for $g(x) = \frac{3}{4}x + \frac{1}{4}$ in Fig. 195.8. Thinking about (195.6), the reason is simply that the geometric series with factor L converges when $L < 1$.

195.7 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

$$|x_{k+1} - x_k| \leq L|x_k - x_{k-1}|. \tag{195.7}$$

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

$$|x_{k+1} - x_k| \leq L^k|x_1 - x_0|. \tag{195.8}$$

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 (195.8) 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

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

using the formula for the sum of a geometric series. We now use the assumption that $L < 1$, to conclude that $0 \leq 1 - L^{j-i} \leq 1$ and therefore for $j > i$,

$$|x_i - x_j| \leq \frac{L^i}{1 - L} \, |x_1 - x_0|.$$

Since $L < 1$, the factor L^i can be made as small as we please by taking i large enough, and thus $\{x_i\}_{i=1}^{\infty}$ is a Cauchy sequence and therefore converges to a limit $\bar{x} = \lim_{i \to \infty} x_i$.

Note that the idea of estimating $|x_i - x_j|$ for $j > i$ by estimating $|x_k - x_{k+1}|$ and using the formula for a geometric sum is fundamental and will be used repeatedly below.

Proof that $\bar{x} = \lim_i x_i$ is a Fixed Point

Since $g(x)$ is Lipschitz continuous, we have

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

By the nature of the Fixed Point Iteration with $x_i = g(x_{i-1})$, we have

$$\lim_{i\to\infty} g(x_{i-1}) = \lim_{i\to\infty} x_i = \bar{x}.$$

Since of course

$$\lim_{i\to\infty} g(x_{i-1}) = \lim_{i\to\infty} g(x_i),$$

we thus see that $g(\bar{x}) = \bar{x}$ as desired. We conclude that the limit $\lim_i x_i = \bar{x}$ is a fixed point.

Proof of Uniqueness

Suppose that x and y are two fixed points, that is $x = g(x)$ and $y = g(y)$. Since $g : \mathbb{R} \to \mathbb{R}$ is a contraction mapping,

$$|x - y| = |g(x) - g(y)| \leq L|x - y|$$

which is possible only if $x = y$ since $L < 1$. This completes the proof.

We have now proved that a contraction mapping $g : \mathbb{R} \to \mathbb{R}$ has a unique fixed point given by Fixed Point Iteration. We summarize in the following basic theorem.

Theorem 195.1 *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 $\bar{x}$.*

195.8 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 195.2 *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 195.4. 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 $x_0 = 9.9$.

195.9 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

$$|x_i - \bar{x}| = |g(x_{i-1}) - g(\bar{x})| \leq L|x_{i-1} - \bar{x}|, \tag{195.9}$$

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. 195.9. 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

FIGURE 195.9. 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.

195.10 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 (195.9) 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. 195.10. We see that Fixed Point Iteration converges

i	x_i for $\frac{1}{2}x$	x_i for $\frac{1}{2}x^2$
0	0.50000000000000	0.50000000000000
1	0.25000000000000	0.25000000000000
2	0.12500000000000	0.06250000000000
3	0.06250000000000	0.00390625000000
4	0.03125000000000	0.00001525878906
5	0.01562500000000	0.00000000023283
6	0.00781250000000	0.00000000000000

FIGURE 195.10. Results of the fixed point iterations for $\frac{1}{2}x$ and $\frac{1}{2}x^2$.

much more quickly for $\frac{1}{2}x^2$, reaching 15 places of accuracy after 7 iterations. The estimate (195.9) thus does not give the full picture.

We now take a closer look into the argument behind (195.9) 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,

$$\text{for } i = 1 \text{ the factor is } \tfrac{1}{2}|x_0|,$$

$$\text{for } i = 2 \text{ the factor is } \tfrac{1}{2}|x_1|,$$

$$\text{for } i = 3 \text{ the factor is } \tfrac{1}{2}|x_2|,$$

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 (195.9) 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 (195.9) 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.

195.11 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

$$x = g(x) = \frac{1}{x} + \frac{x}{2}. \tag{195.10}$$

To see this, it suffices to multiply the equation (195.10) by x. We now apply Fixed Point Iteration to (195.10) to compute $\sqrt{2}$ and show the result in Fig. 195.11. We note that it only takes 5 iterations to reach 15 places of

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

FIGURE 195.11. The fixed point iteration for (195.10).

accuracy. The convergence appears to be very quick.

To see how quick the convergence in fact is, we seek a relation between the error in two consecutive steps. Computing as in (195.9), we find that

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

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

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

to get

$$|x_i - \sqrt{2}| = \frac{(x_{i-1} - \sqrt{2})^2}{2x_{i-1}} \approx \frac{1}{2\sqrt{2}}(x_{i-1} - \sqrt{2})^2. \tag{195.11}$$

We conclude that the error in x_i is the square of the error of x_{i-1} up to the factor $\frac{1}{2\sqrt{2}}$. This is *quadratic* convergence, which is very quick. In each step of the iteration, the number of correct decimals doubles!

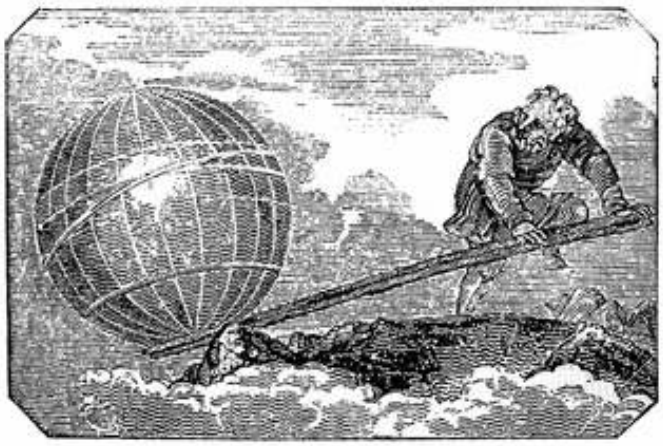


FIGURE 195.12. Archimedes moving the Earth with a lever and a fixed point.

Chapter 195 Problems

195.1. 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.

195.2. Rewrite the following fixed point problems as root problems three different ways each.

$$\text{(a) } \frac{x^3-1}{x+2} = x \qquad \text{(b) } x^5 - x^3 + 4 = x$$

195.3. Rewrite the following root problems as fixed point problems three different ways each.

$$\text{(a) } 7x^5 - 4x^3 + 2 = 0 \qquad \text{(b) } x^3 - \frac{2}{x} = 0$$

195.4. (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 $g(1) > 1$.

195.5. Write a program that implements Algorithm 195.2. 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. 195.9 that were computed using *MATLAB*© .

195.6. In Section 184.10, suppose that K_{sp} for $\mathrm{Ba(IO_3)_2}$ is 1.8×10^{-5}. Find the solubility S to 10 decimal places using the program from Proposition 195.5 after writing the problem as a suitable fixed point problem. Hint: $1.8 \times 10^{-5} = 18 \times 10^{-6}$ and $10^{-6} = 10^{-2} \times 10^{-4}$.

195.7. In Section 184.10, determine the solubility of $\mathrm{Ba(IO_3)_2}$ in a .037 mole/liter solution of $\mathrm{KIO_3}$ to 10 decimal places using the program from Proposition 195.5 after writing the problem as a suitable fixed point problem.

195.8. 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 195.5 after writing the problem as a fixed point problem.

195.9. 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 195.5 after writing the problem as a fixed point problem.

195.10. Verify that (195.6) is true.

195.11. (a) Find an explicit formula (similar to (195.6)) 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 ∞.

195.12. (a) Find an explicit formula (similar to (195.6)) 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.

195.13. (a) Find an explicit formula (similar to (195.6)) 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.

195.14. Draw a Lipschitz continuous function g that does *not* have the property that x in $[0, 1]$ means that $g(x)$ is in $[0, 1]$.

195.15. (a) If possible, find intervals suitable for application of the fixed point iteration to each of the three fixed point problems found in Problem 195.3(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 195.3(b). In each case, a suitable interval is one on which the function is a contraction map.

195.16. *Harder* Apply Theorem 195.2 to the function $g(x) = 1/(1 + x^2)$ to show that the fixed point iteration converges on any interval $[a, b]$.

195.17. 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 (195.8).

195.18. Verify the details of Example 195.4.

195.19. (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 195.5 to compute 6 fixed point iterations starting with $x_0 = .5$ and compare to the results in Fig. 195.10. (c) Show that the error of x_i is approximately the cube of the error of x_{i-1} for any i.

195.20. Verify that (195.11) is true.

195.21. (a) Show the root problem $f(x) = x^2 + x - 6$ can be written as the fixed point problem $g(x) = x$ with $g(x) = \dfrac{6}{x+1}$. Show that the error of x_i decreases at a linear rate to the fixed point $\bar{x} = 2$ when the fixed point iteration converges to 2 and estimate the convergence factor for x_i close to 2. (b) Show the root problem $f(x) = x^2 + x - 6$ can be written as the fixed point problem $g(x) = x$ with $g(x) = \dfrac{x^2+6}{2x+1}$. Show that the error of x_i decreases at a quadratic rate to the fixed point $\bar{x} = 2$ when the fixed point iteration converges to 2.

195.22. 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.

195.23. 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.

196
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)

196.1 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.

196.2 Paying Taxes

We return to the above example of describing our income tax as a function of income. Suppose we let x denote our total income next year and let $f(x)$ be the corresponding total income tax we would have to pay. The function $f(x)$ describes how our total income tax changes with our income x. For each given income x, there is a corresponding income tax $f(x)$ to pay. We plot a possible function $f(x)$ in the following figure:

The function $f(x)$ in this example is piecewise linear and Lipschitz continuous. The slope of $f(x)$ is zero up to the total income 5, the slope is 0.2

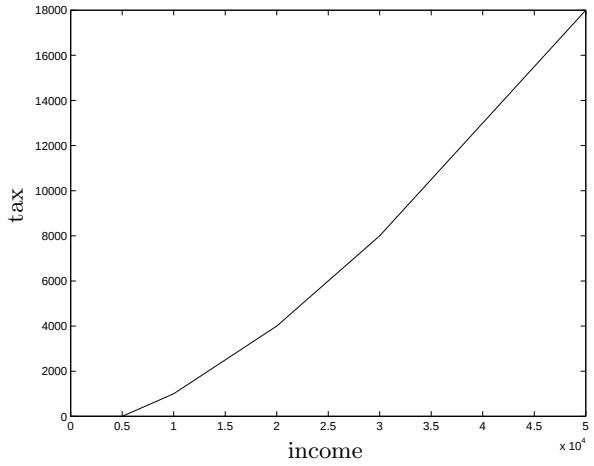


FIGURE 196.1. Income tax $f(x)$ varying with income x.

in the interval $[5\,000, 10\,000]$, 0.3 in $[10\,000, 20\,000]$, 0.4 in $[20\,000, 30\,000]$ and 0.5 in $[30\,000, \infty)$.

For a given $\bar{x}$, the slope of the straight line representing $f(x)$ close to $\bar{x}$, is the marginal tax. We denote the slope of $f(x)$ at $\bar{x}$ by $m(\bar{x})$. We see that the slope $m(\bar{x})$ varies with $\bar{x}$. For example, $m(\bar{x}) = 0.3$ for $\bar{x} \in (10\,000, 20\,000)$. If we add one extra dollar at the income $\bar{x} \in (10\,000, 20\,000)$, then our income tax will increase by 0.3 dollars.

The marginal tax is the same as the slope of the straight line representing the income tax $f(x)$ as a function of income x. Thus the marginal tax is $m(\bar{x})$ at the income $\bar{x}$. The marginal income tax is zero up to the total income 5 000, the marginal tax is 0.2 in the income bracket $[5\,000, 10\,000]$, 0.3 in the bracket $[10\,000, 20\,000]$, 0.4 in the bracket $[20\,000, 30\,000]$ and 0.5 for incomes in the bracket $[30\,000, \infty)$.

We can describe how $f(x)$ varies in each income tax bracket through the following formula

$$
\begin{aligned}
&f(x) = 0 \quad \text{for } x \in [0, 5\,000] \\
&f(x) = 0.2(x - 5\,000) \quad \text{for } x \in [5\,000, 10\,000] \\
&f(x) = f(10\,000) + 0.3(x - 10\,000) \quad \text{for } x \in [10\,000, 20\,000] \\
&f(x) = f(20\,000) + 0.4(x - 20\,000) \quad \text{for } x \in [20\,000, 30\,000] \\
&f(x) = f(30\,000) + 0.5(x - 30\,000) \quad \text{for } x \in [30\,000, \infty)
\end{aligned}
$$

We can condense these formulas into

$$f(x) = f(\bar{x}) + m(\bar{x})(x - \bar{x}), \tag{196.1}$$

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. 196.2.

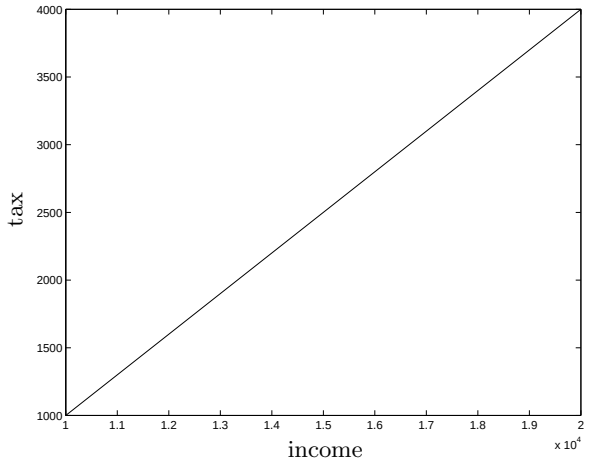


FIGURE 196.2. Income tax $f(x)$ for income x in the interval $[10\,000, 20\,000]$.

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:

$$f(x) - f(\bar{x}) = m\,(x - \bar{x}), \tag{196.2}$$

that is if $x \neq \bar{x}$, then the slope m is given by

$$m = \frac{f(x) - f(\bar{x})}{x - \bar{x}} \tag{196.3}$$

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.

196.3 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.

196.4 Definition of the Derivative

We shall now seek to define the *derivative* of a given function $f : \mathbb{R} \to \mathbb{R}$ at a given point $\bar{x}$. We shall then follow the idea that if $f(x)$ is particularly well approximated by the linear function $f(\bar{x}) + m\,(x - \bar{x})$ for x close to $\bar{x}$, then the derivative of $f(x)$ at $\bar{x}$ will be equal to m. In other words, the derivative of $f(x)$ at $\bar{x}$ will be equal to the slope m of the approximating linear function $f(\bar{x}) + m\,(x - \bar{x})$. Of course, a key point is to describe how to interpret that the linear function $f(\bar{x}) + m\,(x - \bar{x})$ approximates $f(x)$ "particularly well". We shall see that the natural requirement is to ask that the error is proportional to $|x - \bar{x}|^2$, that is that the error is quadratic in the difference $x - \bar{x}$. Geometrically, this will be the same as asking the straight line $y = f(\bar{x}) + m\,(x - \bar{x})$ to be *tangent* to the graph of $y = f(x)$ at $(\bar{x}, f(\bar{x}))$. We will see that asking the error to be quadratic in $x - \bar{x}$ is just about right. In particular, asking the error to be even smaller, for example proportional to $|x - \bar{x}|^3$, would be to ask for too much.

Before defining the derivative, we back off a little to prepare ourselves and consider different linear approximations $b + m\,(x - \bar{x})$ of the given function $f(x)$ for x close to $\bar{x}$. There are many straight lines that approximate $f(x)$ close to $\bar{x}$. We show some bad approximations and a number of good approximations in Fig. 196.3. 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.

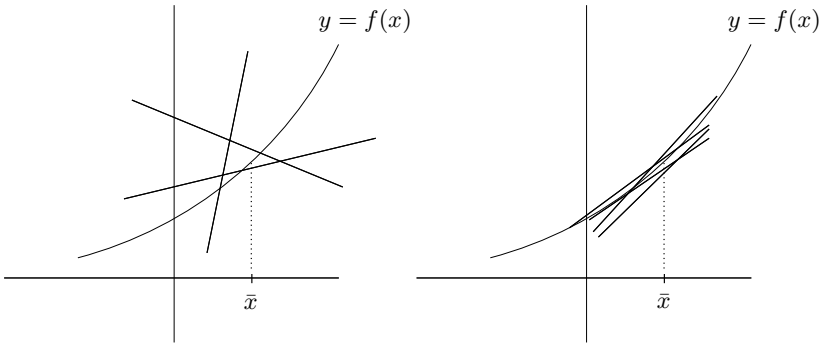


FIGURE 196.3. Linear approximations of $f(x)$ close to $\bar{x}$

The question is whether one of the many possible approximate lines is a particularly good choice or not.

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

$$y = f(\bar{x}) + m(x - \bar{x}). \tag{196.4}$$

We started this section considering approximations of $f(x)$ of this form. We plot several examples in Fig. 196.4 with different slopes m.

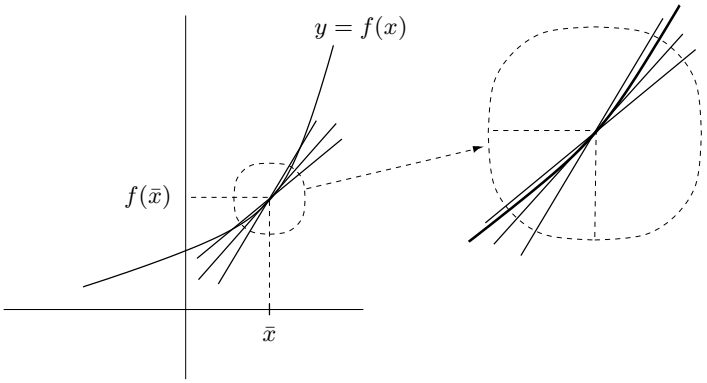


FIGURE 196.4. Linear approximations to a function that pass through the point $(\bar{x}, f(\bar{x}))$. The region near $(\bar{x}, f(\bar{x}))$ has been blown-up on the right.

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. 196.4 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

$$E_f(x, \bar{x}) = f(x) - \big(f(\bar{x}) + m(\bar{x})(x - \bar{x})\big), \tag{196.5}$$

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 (196.5) 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. 196.5. It is natural to say that the correction $E_f(x, \bar{x})$ is particularly small if it is much smaller than the term $m(x - \bar{x})$, which represents a linear correction of the constant value $f(\bar{x})$. Thus, $f(\bar{x}) + m(\bar{x})(x - \bar{x})$ is a linear approximation of $f(x)$ close to $\bar{x}$ with zero error for $x = \bar{x}$, and we seek $m(\bar{x})$ so that the correction $E_f(x, \bar{x})$ is small compared to $m(x - \bar{x})$ for x close to $\bar{x}$.

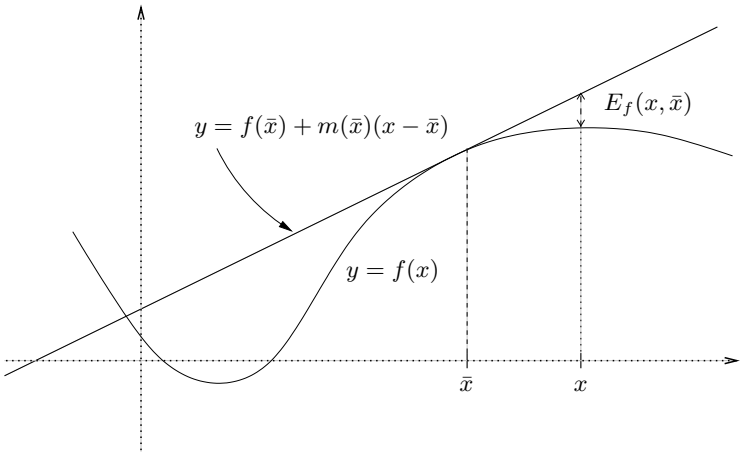


FIGURE 196.5. Graph $y = f(x)$, tangent $y = f(\bar{x}) + m(\bar{x})(x - \bar{x})$ and error $E_f(x, \bar{x})$.

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

$$|E_f(x, \bar{x})| \le K_f(\bar{x})|x - \bar{x}|^2 \quad \text{for } x \text{ close to } \bar{x}, \tag{196.6}$$

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 (196.6) for some constant $K_f(\bar{x})$ for x close to $\bar{x}$. We thus seek to choose the slope $m = m(\bar{x})$ so that the error $E_f(x,\bar{x})$ is quadratic in $x-\bar{x}$. The linear function $f(\bar{x}) + m(\bar{x})(x-\bar{x})$ will then be *tangent* to $f(x)$ at $\bar{x}$. We expect the slope of the tangent at $\bar{x}$ to depend on $\bar{x}$, which we indicate by denoting the slope by $m(\bar{x})$.

Now we are in position to define the derivative of $f(x)$ at $\bar{x}$. The function $f(x)$ is said to be *differentiable* at $\bar{x}$ if there are constants $m(\bar{x})$ and $K_f(\bar{x})$ such that for x close to $\bar{x}$,

$$f(x) = f(\bar{x}) + m(\bar{x})(x-\bar{x}) + E_f(x,\bar{x}),$$
$$\text{with } |E_f(x,\bar{x})| \le K_f(\bar{x})|x-\bar{x}|^2. \quad (196.7)$$

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 (196.7) 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.

196.5 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

$$\text{if } f(x) = x, \quad \text{then } f'(x) = 1. \quad (196.8)$$

196.6 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. 196.6,

$$\text{if } f(x) = x^2, \quad \text{then } f'(x) = 2x. \tag{196.9}$$

An alternative, shorter route to the linearization formula (196.6) in this case is

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

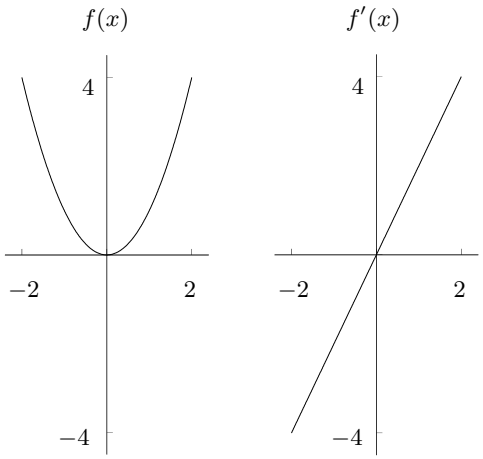


FIGURE 196.6. $f(x) = x^2$ and $f'(x) = 2x$.

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. 196.7. We compare some values of the given function

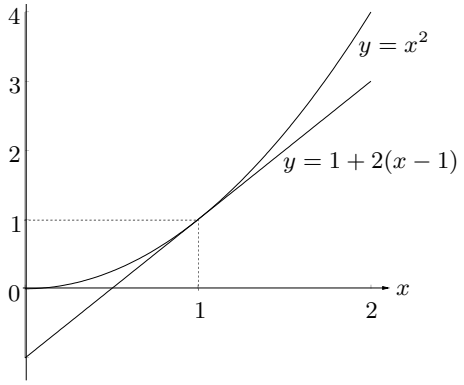


FIGURE 196.7. The linearization $1 + 2(x - 1)$ of x^2 at $\bar{x} = 1$.

x^2 to the linear approximation $1 + 2(x - 1)$ along with the error $(x - 1)^2$ in Fig. 196.8

x	$f(x)$	$f(1) + f'(2)(x - 1)$	$E_f(x, 1)$
.7	.49	.4	.09
.8	.64	.6	.04
.9	.81	.8	.01
1.0	1.0	1.0	0.0
1.1	1.21	1.2	.01
1.2	1.44	1.4	.04
1.3	1.69	1.6	.09

FIGURE 196.8. Some values of $f(x) = x^2$, $f(1) + f'(1)(x - 1) = 1 + 2(x - 1)$, and $E_f(x, 1) = (x - 1)^2$.

196.7 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 (196.6), 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:

$$\text{if } f(x) = x^n, \quad \text{then } f'(x) = nx^{n-1}. \tag{196.10}$$

For $n = 2$, we recover the formula $f'(x) = 2x$ if $f(x) = x^2$.

196.8 The Derivative of $\frac{1}{x}$ Is $-\frac{1}{x^2}$ for $x \neq 0$

We now compute the derivative of the function $f(x) = \frac{1}{x}$ for $x \neq 0$. We have for x close to $\bar{x} \neq 0$,

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

where

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

and thus $|E| \leq K|x - \bar{x}|^2$ as desired. We conclude that $f(x) = \frac{1}{x}$ is differentiable at $\bar{x}$ with derivative $f'(\bar{x}) = -\frac{1}{\bar{x}^2}$ for $\bar{x} \neq 0$, that is

$$\text{if } f(x) = \frac{1}{x}, \quad \text{then } f'(\bar{x}) = -\frac{1}{\bar{x}^2} \quad \text{for } \bar{x} \neq 0. \tag{196.11}$$

196.9 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 $3x^2$.

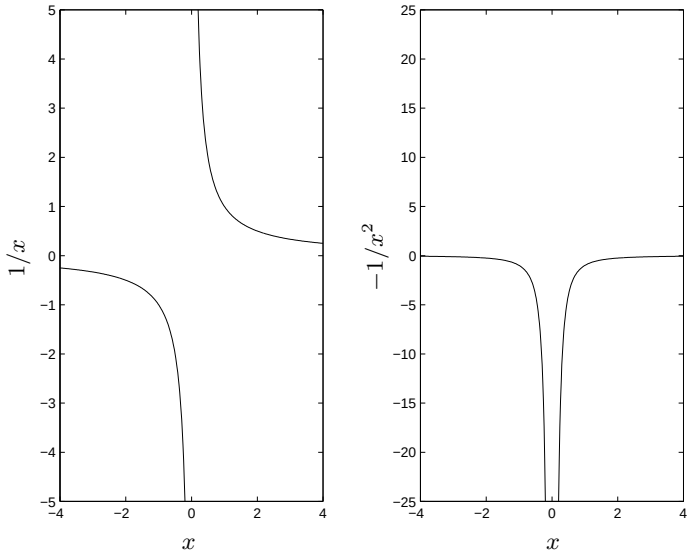


FIGURE 196.9. The function $f(x) = 1/x$ and its derivative $f'(x) = -1/x^2$ for $x > 0$.

196.10 Denoting the Derivative of $f(x)$ by $Df(x)$

We also denote the derivative $f'(x)$ of $f(x)$ by $Df(x)$, that is

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

Observe that $D(f)$ denotes the domain of f, while $Df(x)$ denotes the derivative of $f(x)$ at x.

We may write the basic formula (196.10) as

$$\text{if } f(x) = x^n, \quad \text{then } f'(x) = Df(x) = nx^{n-1}, \tag{196.12}$$

or

$$Dx^n = nx^{n-1} \quad \text{for } n = 1, 2, ... \tag{196.13}$$

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 (196.12).

EXAMPLE 196.1. 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 $0 \le t < 1$, which means that the tangent lines to $s(t)$ have positive slope for $0 \le t < 1$, 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.

196.11 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

$$\frac{df}{dx} = f'(x) \tag{196.14}$$

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 (196.22) 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

$$\frac{d}{dx}(x^n) = nx^{n-1} \tag{196.15}$$

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 $Df = \frac{df}{dx}$.

196.12 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 $\bar{x}$,

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}), \tag{196.16}$$

where

$$|E_f(x, \bar{x})| \le K_f(\bar{x})|x - \bar{x}|^2, \tag{196.17}$$

and $K_f(\bar{x})$ is a constant. Dividing by $x - \bar{x}$ assuming $x \neq \bar{x}$, we get for $x \in I$,

$$\frac{f(x) - f(\bar{x})}{x - \bar{x}} = f'(\bar{x}) + R_f(x, \bar{x}), \tag{196.18}$$

where

$$R_f(x, \bar{x}) = \frac{E_f(x, \bar{x})}{x - \bar{x}}, \tag{196.19}$$

satisfies

$$|R_f(x, \bar{x})| \leq K_f(\bar{x})|x - \bar{x}| \quad \text{for } x \in I. \tag{196.20}$$

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 (196.20) it follows that

$$\lim_{i\to\infty} R_f(x_i, \bar{x}) = 0, \tag{196.21}$$

and thus by (196.18) we have

$$f'(\bar{x}) = \lim_{i\to\infty} m_i(\bar{x}), \tag{196.22}$$

where

$$m_i(\bar{x}) = \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}} \tag{196.23}$$

is the *difference quotient* based on the two distinct points $\bar{x}$ and x_i. The difference quotient $m_i(\bar{x})$ defined by (196.23) is the slope of the *secant line* connecting the points $(\bar{x}, f(\bar{x}))$ and $(x_i, f(x_i))$, see Fig. 196.10, and can be viewed as the *average rate of change* of $f(x)$ between the points $\bar{x}$ and x_i.

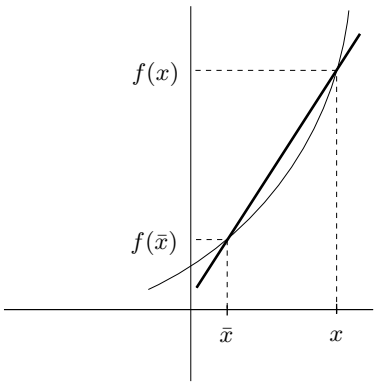


FIGURE 196.10. The secant line joining $(\bar{x}, f(\bar{x}))$ and $(x_i, f(x_i))$.

The formula

$$f'(\bar{x}) = \lim_{i\to\infty} \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}}, \tag{196.24}$$

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. 196.11.

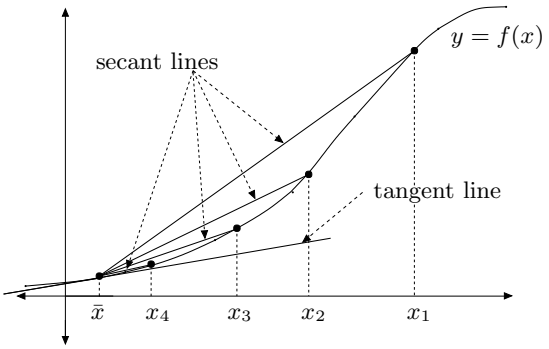


FIGURE 196.11. A sequence of secant lines approaching the tangent line at $\bar{x}$.

EXAMPLE 196.2. Let us now compute the derivative of $f(x) = x^2$ at $\bar{x}$ by using (196.22). 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 (196.22), we have

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

and we recover the well known formula $Dx^2 = 2x$.

196.13 How to Compute a Derivative?

Suppose $f(x)$ is a given function for which we are not able to analytically compute the derivative $f'(\bar{x})$ for a given $\bar{x}$. Note that we were able to carry out the analytical computation above for polynomials, but we gave no strategy to determine the derivative $f'(\bar{x})$ for more general functions

$f(x)$. The function $f(x)$ may not be given by any formula at all, and could just be given as a value $f(x)$ for each x determined in some way.

The same problem arises if we want to determine a physical velocity by doing some measurement. For example, if the speed meter of our car is out of function, how can we measure the velocity of the car at some given time $\bar{x}$? Of course the natural thing would be to measure the increment of distance $f(x) - f(\bar{x})$ over some time interval $x - \bar{x}$, where $f(x)$ is the total distance, and then use the quotient $\frac{f(x)-f(\bar{x})}{x-\bar{x}}$, the average velocity over the time interval $(\bar{x}, x)$, as an approximation of the momentary velocity at time $\bar{x}$. But how to choose the length of the time interval $x - \bar{x}$? If we choose $x - \bar{x}$ way too small, then we will not be able to measure any change in position at all, that is we will have $f(x) = f(\bar{x})$, and then conclude zero velocity, while if we take $x - \bar{x}$ too large, the computed average velocity may differ very much from the desired momentary velocity at $\bar{x}$.

We now use analysis to find the right increment $x - \bar{x}$ to use to determine the derivative $f'(\bar{x})$ of a given function $f(x)$ at $\bar{x}$, assuming that the function values $f(x)$ are given with a certain precision. From the definition of $f'(\bar{x})$, we have for x close to $\bar{x}$, $x \neq \bar{x}$,

$$f'(\bar{x}) = \frac{f(x) - f(\bar{x})}{x - \bar{x}} - \frac{E_f(x, \bar{x})}{x - \bar{x}},$$

where

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

The difference quotient

$$\frac{f(x) - f(\bar{x})}{x - \bar{x}},$$

may thus be used as an approximation of $f'(\bar{x})$ up to a linearization error of size $K_f(\bar{x})|x - \bar{x}|$.

Suppose now that we know the quantity $f(x) - f(\bar{x})$ up to an error of size δf. We thus assume that we know x and $\bar{x}$ exactly, but there is an error of size δf in the quantity $f(x) - f(\bar{x})$ resulting from errors in the function values $f(x)$ and $f(\bar{x})$ from computation or measurement. We know that frequently the value $f(x)$ for a given x, is known only approximately through computation.

The error δf in $f(x) - f(\bar{x})$ causes an error of size $|\frac{\delta f}{x-\bar{x}}|$ in the difference quotient $\frac{f(x)-f(\bar{x})}{x-\bar{x}}$. We thus have a total error in $f'(\bar{x})$ of size

$$|\frac{\delta f}{x - \bar{x}}| + K_f(\bar{x})|x - \bar{x}|, \tag{196.25}$$

resulting from the error in $f(x) - f(\bar{x})$ and the linearization error. Making the two error contributions equal, which should give the right balance, we get the equation

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

where we write $K_f = K_f(\bar{x})$, from which we compute the "optimal increment"

$$|x - \bar{x}| = \sqrt{\frac{\delta f}{K_f}}. \tag{196.26}$$

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 (196.25), we get a corresponding "best" error estimate

$$|f'(\bar{x}) - \frac{f(x) - f(\bar{x})}{x - \bar{x}}| \leq 2\sqrt{\delta f}\sqrt{K_f}. \tag{196.27}$$

Contemplating the two resulting formulas (196.26) and (196.27) 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 $Dx^2 = 2x$.

196.14 Uniform Differentiability on an Interval

We say that the function $f(x)$ is *differentiable* on the interval I if $f(x)$ is differentiable for each $\bar{x} \in I$, that is for $\bar{x} \in I$ there are constants $m(\bar{x})$ and

$K_f(\bar{x})$ such that for x close to $\bar{x}$,

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

In many cases we can choose one and the same constant $K_f(\bar{x}) = K_f$ for all $\bar{x} \in I$. We may express this by saying the $f(x)$ is uniformly differentiable on I. Allowing also x to vary in I we are led to the following definition, which we will find very useful below: We say that the function $f : I \to \mathbb{R}$ is *uniformly differentiable on the interval* I with derivative $f'(\bar{x})$ at $\bar{x}$, if there is a constant K_f such that for $x, \bar{x} \in I$,

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

Observe that the important thing is that K_f here does not depend on $\bar{x}$, but may of course depend on the function f and the interval I.

196.15 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 $x \in I$,

$$|f'(x)| \leq L. \tag{196.28}$$

We shall now show that $f(x)$ is Lipschitz continuous on I with Lipschitz constant L, that is we shall show that

$$|f(x) - f(y)| \leq L|x - y| \quad \text{for } x, y \in I. \tag{196.29}$$

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. 196.12.

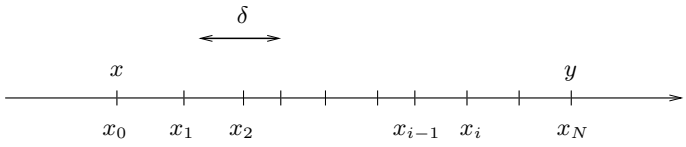


FIGURE 196.12. Subdivision of interval $[x, y]$ into subintervals of length $< \delta$.

We have by the triangle inequality

$$|f(x)-f(y)| = |\sum_{i=1}^{N}(f(x_i)-f(x_{i-1})|$$

$$\leq \sum_{i=1}^{N}|f(x_i)-f(x_{i-1})| \leq (L+K\delta)\sum_{i=1}^{N}|x_i-x_{i-1}|$$

$$= (L+K\delta)|x-y|.$$

Since this inequality holds for any $\delta > 0$, we conclude that indeed

$$|f(x)-f(y)| \leq L|x-y|, \quad \text{for } x, y \in I,$$

which proves the desired result. We summarize in the following theorem which we will use extensively below:

Theorem 196.1 *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.

196.16 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

$$\lim_{i\to\infty} \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}}, \tag{196.30}$$

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 (196.22) 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

$$|f'(\bar{x}) - \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}}| \leq K_f(\bar{x})|x_i - \bar{x}|. \tag{196.31}$$

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 (196.31), whereas the definition (196.30) 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 (196.31) 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.

196.17 Swedenborg

A Swedish counterpart of the Universal Genius Leibniz, together with Newton the Inventor of Calculus, was Emanuel Swedenborg (1688-1772). Swedenborg introduced Calculus to Sweden with independent contributions. Swedenborg produced 150 works on seventeen sciences, was a musician, mining engineer, member of the Swedish parliament, invented a glider, an undersea boat, an ear trumpet for the deaf, a mathematician who wrote the

first books in Swedish on algebra and calculus, a physiologist who discovered the function of several areas of the brain and ductless glands, creator of the (at the time) world's largest dry-dock, and suggested the nebula theory of the formation of the planets.

FIGURE 196.13. 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"

Chapter 196 Problems

196.1. Prove directly from the definition that the derivative of x^3 is $3x^2$, and that the derivative of x^4 is $4x^3$.

196.2. 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 $(\sqrt{x} - \sqrt{\bar{x}})(\sqrt{x} + \sqrt{\bar{x}}) = x - \bar{x}$.

196.3. 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 $\sqrt{x}$.

196.4. Study the symmetric difference quotient approximation

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

What is an optimal choice of the increment h, assuming $f(\bar{x} \pm h)$ is not known exactly. Hint: You may find it useful to look ahead into the next chapter (Taylor's formula of order 2).

196.5. Compute the derivative of x^n numerically for different values of x and n and study how the error depends on the increment used.

196.6. Can you compute the derivative of $\sin(x)$ and $\cos(x)$ from the definition?

196.7. Determine the smallest possible Lipschitz constant for the function $f(x) = x^3$ with $D(f) = [1, 4]$.

196.8. (*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.

196.9. 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.

197
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)

197.1 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*.

197.2 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 $\bar{x}$,

$$\begin{aligned} f(x) &= f(\bar{x}) + f'(\bar{x})(x-\bar{x}) + E_f(x,\bar{x}),\\ g(x) &= g(\bar{x}) + g'(\bar{x})(x-\bar{x}) + E_g(x,\bar{x}), \end{aligned} \tag{197.1}$$

and

$$|E_f(x,\bar{x})| \le K_f|x-\bar{x}|^2, \quad |E_g(x,\bar{x})| \le K_g|x-\bar{x}|^2, \tag{197.2}$$

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

$$(f+g)(x) = (f+g)(\bar{x}) + (f'(\bar{x})+g'(\bar{x}))(x-\bar{x}) + E_{f+g}(x,\bar{x}) \tag{197.3}$$

where

$$E_{f+g}(x,\bar{x}) = E_f(x,\bar{x}) + E_g(x,\bar{x}).$$

By (197.2), we have

$$|E_{f+g}(x,\bar{x})| \le (K_f+K_g)|x-\bar{x}|^2.$$

The formula (197.3) shows that $(f+g)(x)$ is differentiable at $\bar{x}$ and

$$(f+g)'(\bar{x}) = f'(\bar{x}) + g'(\bar{x}). \tag{197.4}$$

Next, multiplying the first line in (197.1) by a constant c, we get

$$(cf)(x) = (cf)(\bar{x}) + cf'(\bar{x})(x-\bar{x}) + cE_f(x,\bar{x}) \tag{197.5}$$

This proves that if $f(x)$ is differentiable at $\bar{x}$, then $(cf)(x)$ is differentiable at $\bar{x}$ and

$$(cf)'(\bar{x}) = cf'(\bar{x}). \tag{197.6}$$

We summarize in

Theorem 197.1 (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 $x \in I$,*

$$(f+g)'(x) = f'(x)+g'(x), \quad \textit{or} \quad D(f+g)(x) = Df(x)+Dg(x), \tag{197.7}$$

and

$$(cf)'(x) = cf'(x), \quad \textit{or} \quad D(cf)(x) = cDf(x). \tag{197.8}$$

EXAMPLE 197.1.

$$D\big(2x^3 + 4x^5 + \frac{7}{x}\big) = 6x^2 + 20x^4 - \frac{7}{x^2}.$$

EXAMPLE 197.2. Using the above theorem and the fact that $Dx^i = ix^{i-1}$, we find that the derivative of

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n = \sum_{i=0}^{n} a_ix^i$$

is

$$f'(x) = a_1 + 2a_2x^2 + \cdots + na_nx^{n-1} = \sum_{i=1}^{n} ia_ix^{i-1}.$$

197.3 The Product Rule

Multiplying the left and right-hand sides, respectively, of the two equations in (197.1), we obtain

$$\begin{aligned}(fg)(x) = f(x)g(x) &= f(\bar{x})g(\bar{x}) \\ &\quad + f'(\bar{x})g(\bar{x})(x-\bar{x}) + f(\bar{x})g'(\bar{x})(x-\bar{x}) + f'(\bar{x})g'(\bar{x})(x-\bar{x})^2 \\ &\quad + (g(\bar{x}) + g'(\bar{x})(x-\bar{x}))E_f(x,\bar{x}) + (f(\bar{x}) \\ &\quad + f'(\bar{x})(x-\bar{x}))E_g(x,\bar{x}) + E_f(x,\bar{x})E_g(x,\bar{x}).\end{aligned}$$

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 197.2 (The Product rule) *If $f(x)$ and $g(x)$ are differentiable on I, then $(fg)(x)$ is differentiable on I and*

$$(fg)'(x) = f(x)g'(x) + f'(x)g(x), \tag{197.9}$$

that is,

$$D(fg)(x) = Df(x)g(x) + f(x)Dg(x), \tag{197.10}$$

EXAMPLE 197.3.

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

197.4 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 $\bar{x}$,

$$\begin{aligned} f(y) &= f(\bar{y}) + f'(\bar{y})(y-\bar{y}) + E_f(y,\bar{y}),\\ g(x) &= g(\bar{x}) + g'(\bar{x})(x-\bar{x}) + E_g(x,\bar{x}), \end{aligned} \tag{197.11}$$

and

$$|E_f(y,\bar{y})| \le K_f|y-\bar{y}|^2, \quad |E_g(x,\bar{x})| \le K_g|x-\bar{x}|^2, \tag{197.12}$$

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

$$|g(x) - g(\bar{x})| \le L_g|x-\bar{x}|. \tag{197.13}$$

Setting $y = g(x)$ and recalling that $\bar{y} = g(\bar{x})$, we have

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

Substituting $g(x) - g(\bar{x}) = g'(\bar{x})(x-\bar{x}) + E_g(x,\bar{x})$, we thus have

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

Since (197.12) and (197.13) 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 197.3 (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

$$(f \circ g)'(x) = f'(g(x))g'(x), \quad \text{for } x \in I, \tag{197.14}$$

or

$$\frac{dh}{dx} = \frac{df}{dy}\frac{dy}{dx}, \tag{197.15}$$

where $h(x) = f(y)$ and $y = g(x)$, that is $h(x) = f(g(x)) = (f \circ g)(x)$. An alternative formulation si

$$D(f(g(x)) = Df(g(x))Dg(x), \tag{197.16}$$

where $Df = \frac{df}{dy}$.

EXAMPLE 197.4. 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 197.5.

$$\begin{aligned} D\big(7x^3 + 4x + 6\big)^{18} &= 18\big(7x^3 + 4x + 6\big)^{17} D(7x^3 + 4x + 6) \\ &= 18\big(7x^3 + 4x + 6\big)^{17}(21x^2 + 4). \end{aligned}$$

EXAMPLE 197.6. 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

$$Dh(x) = D\frac{1}{g(x)} = \frac{-1}{(g(x))^2}\, g'(x) = \frac{-g'(x)}{g(x)^2}, \tag{197.17}$$

as long as $g(x)$ is differentiable and $g(x) \neq 0$.

EXAMPLE 197.7. Using Example 197.6 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 $m = -1, -2, ...$

197.5 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 (197.17), 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 197.4 (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 197.8.

$$D\left(\frac{3x+4}{x^2-1}\right) = \frac{3 \times (x^2-1) - (3x+4) \times 2x}{(x^2-1)^2}.$$

Example 197.9.

$$\begin{aligned}
&\frac{d}{dx}\left(\frac{x^3+x}{(8-x)^6}\right)^9 \\
&\quad = 9\left(\frac{x^3+x}{(8-x)^6}\right)^8 \frac{d}{dx}\left(\frac{x^3+x}{(8-x)^6}\right) \\
&\quad = 9\left(\frac{x^3+x}{(8-x)^6}\right)^8 \frac{(8-x)^6\frac{d}{dx}(x^3+x) - (x^3+x)\frac{d}{dx}(8-x)^6}{\left((8-x)^6\right)^2} \\
&\quad = 9\left(\frac{x^3+x}{(8-x)^6}\right)^8 \frac{(8-x)^6(3x^2+1) - (x^3+x)6(8-x)^5 \times -1}{(8-x)^{12}}.
\end{aligned}$$

Example 197.10. The Chain rule can also be used recursively:

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

197.6 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 197.11. For $f(x) = x^2$, $f'(x) = 2x$ and $f''(x) = 2$.

EXAMPLE 197.12. 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 197.13. 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 197.14. The $n+1$'st derivative of a polynomial of degree n is zero.

EXAMPLE 197.15. If $f(x) = 1/x$, then

$$f(x) = x^{-1},\ Df(x) = -1 \times x^{-2},\ D^2 f(x) = 2 \times x^{-3},\ D^3 f(x) = -6 \times x^{-4}$$

$$\vdots$$

$$D^n f(x) = (-1)^n \times 1 \times 2 \times 3 \times \cdots \times n x^{-n-1} = (-1)^n n! x^{-n-1}.$$

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

$$|f(x) - (f(\bar{x}) + m(\bar{x})(x - \bar{x}))| \leq K_f(\bar{x})|x - \bar{x}|^2. \tag{197.18}$$

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 197.16. The function $f(x) = |x|$ is differentiable for $\bar{x} \neq 0$ with derivative $f'(\bar{x}) = 1$ if $\bar{x} > 0$ and $f'(\bar{x}) = -1$ if $\bar{x} < 0$. The function $f(x) = |x|$ is differentiable from the right at $\bar{x} = 0$ with derivative $f'_+(0) = 1$, and differentiable form the left at $\bar{x} = 0$ with derivative $f'_-(0) = -1$.

We say that $f : [a, b] \to \mathbb{R}$ is *differentiable on the closed interval* $[a, b]$, if $f(x)$ is differentiable on the open interval (a, b), and is differentiable from the right at a, and differentiable from the left at b. The definition extends in the obvious way to half-open/half-closed intervals $(a, b]$ and $[a, b)$. If f is either differentiable or is differentiable from the right and/or the left at every point in an interval, then we say that f is *piecewise differentiable* on the interval.

EXAMPLE 197.17. The function $|x|$ is piecewise differentiable on $\mathbb{R}$. The function $1/x$ is differentiable on $(0, \infty)$ but not differentiable on $[0, \infty)$.

197.8 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

$$f(x) = f(\bar{x}) + m_1(\bar{x})(x-\bar{x}) + m_2(\bar{x})(x-\bar{x})^2 + E_f(x,\bar{x}), \qquad (197.19)$$

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

$$|E_f(x,\bar{x})| \le K_f(\bar{x})|x-\bar{x}|^3, \qquad (197.20)$$

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 (197.19) 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 (197.19) holds, we thus have

$$f(x) = f(\bar{x}) + f'(\bar{x})(x-\bar{x}) + m_2(\bar{x})(x-\bar{x})^2 + E_f(x,\bar{x}). \qquad (197.21)$$

Let us next try to determine the constant $m_2(\bar{x})$. To this end we differentiate the relation (197.19) with respect to x to get

$$f'(x) = f'(\bar{x}) + 2m_2(\bar{x})(x-\bar{x}) + \frac{d}{dx}E_f(x,\bar{x}). \qquad (197.22)$$

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

$$|\frac{d}{dx}E_f(x,\bar{x})| \le M_f(\bar{x})|x-\bar{x}|^2, \qquad (197.23)$$

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 (197.23) 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

$$f(x) = f(\bar{x}) + f'(\bar{x})(x-\bar{x}) + \frac{1}{2}f''(\bar{x})(x-\bar{x})^2 + E_f(x,\bar{x}), \qquad (197.24)$$

for x close to $\bar{x}$, where $|E_f(x,\bar{x})| \le K_f(\bar{x})|x-\bar{x}|^3$ with $K_f(\bar{x})$ a constant.

EXAMPLE 197.18. Consider the function $f(x) = \frac{1}{x}$ for x close to $\bar{x} = 1$. We shall use the fact that if $y \neq -1$, then

$$\frac{1}{1+y} = 1 - y\frac{1}{1+y}$$

which is readily verified by multiplying by $1+y$, and thus

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

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

Choosing $y = x - 1$, we get

$$\frac{1}{x} = \frac{1}{1+(x-1)} = 1 - (x-1) + (x-1)^2 - \frac{(x-1)^3}{1+(x-1)}, \quad (197.25)$$

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. 197.1 and list some values of the approximation in Fig. 197.2.

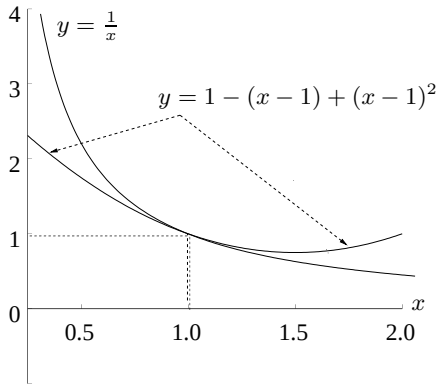


FIGURE 197.1. The quadratic approximation $1 - (x-1) + (x-1)^2$ of $1/x$ near $\bar{x} = 1$

x	$1/x$	$1 - (x-1) + (x-1)^2$	$E_f(x,1)$
.7	1.428571	1.39	.038571
.8	1.25	1.22	.03
.9	1.111111	1.11	.00111
1.0	1.0	1.0	0.0
1.1	.909090	.91	.000909
1.2	.833333	.84	.00666
1.3	.769230	.79	.02077

FIGURE 197.2. 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 $\bar{x}$,

$$f(x) = f(\bar{x}) + f'(\bar{x})(x-\bar{x}) + \frac{1}{2}f''(\bar{x})(x-\bar{x})^2 + E_f(x,\bar{x}), \qquad (197.26)$$

where the error function $E_f(x,\bar{x})$ is cubic in $x-\bar{x}$, more precisely,

$$|E_f(x,\bar{x})| \leq K_f(\bar{x})|x-\bar{x}|^3, \quad \text{for } x \text{ close to } \bar{x}. \qquad (197.27)$$

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 (197.19).

197.9 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 $\bar{x}$

$$f(x) = f(\bar{x}) + f'(\bar{x})(x-\bar{x}) + E_f(x,\bar{x}), \qquad (197.28)$$

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 (197.3).

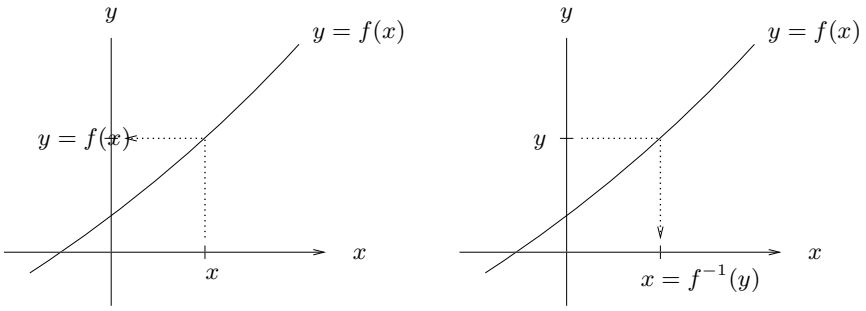


FIGURE 197.3. The function $y = f(x)$ and its inverse $x = f^{-1}(y)$

Can we compute the derivative of the function $x = f^{-1}(y)$ with respect to y close to $\bar{y} = f(\bar{x})$? Rewriting (197.28), 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

$$f^{-1}(y) = f^{-1}(\bar{y}) + \frac{1}{f'(\bar{x})}(y-\bar{y}) - \frac{1}{f'(\bar{x})}E_f(f^{-1}(y), f^{-1}(\bar{y})), \quad (197.29)$$

Suppose now that f^{-1} is Lipschitz continuous in an open interval J around $\bar{y}$, so that

$$|f^{-1}(y) - f^{-1}(\bar{y})| \le L_{f^{-1}}|y-\bar{y}| \quad \text{for } y \in J.$$

Then for y close to $\bar{y}$,

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

which proves by (197.29) 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

$$Df^{-1}(\bar{y}) = \frac{1}{f'(\bar{x})}, \quad (197.30)$$

where $\bar{y} = f(\bar{x})$. We summarize:

Theorem 197.5 *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 197.19. 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 $x > 0$,

$$\frac{d}{dx}\sqrt{x} = D\sqrt{x} = \frac{1}{2\sqrt{x}}, \quad \text{or} \quad Dx^{\frac{1}{2}} = \frac{1}{2}x^{-\frac{1}{2}}. \quad (197.31)$$

197.10 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

$$(y(x))^q = x^p \quad \text{for } x > 0. \quad (197.32)$$

Assuming $y(x)$ to be differentiable with respect to x with derivative $y'(x)$, we would get differentiating both sides of (197.32) 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

$$Dx^r = rx^{r-1} \quad \text{for } r \text{ rational, and } x > 0, \tag{197.33}$$

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 (197.30).

197.11 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,

$$f(x_1, x_2) = 15x_1 + 3x_2, \tag{197.34}$$

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 197.20. Suppose $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is given by $f(x_1, x_2) = x_1^2 + x_2^3 + x_1 x_2$. We compute

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 2x_1 + x_2, \quad \frac{\partial f}{\partial x_2}(x_1, x_2) = 3x_2^2 + x_1,$$

where we follow the principle just explained: to compute $\frac{\partial f}{\partial x_1}$, keep x_2 constant and differentiate with respect to x_1, and to compute $\frac{\partial f}{\partial x_2}$, keep x_1 constant and differentiate with respect to x_2.

More generally, we may in a natural way extend the concept of differentiability of a real-valued function $f(x)$ of one real variable x to differentiability of a real valued function $f(x_1, x_2)$ of two real variables x_1 and x_2 as follows: We say that function $f(x_1, x_2)$ is *differentiable* at $\bar{x} = (\bar{x}_1, \bar{x}_2)$ if there are constants $m_1(\bar{x}_1, \bar{x}_2)$, $m_2(\bar{x}_1, \bar{x}_2)$ and $K_f(\bar{x}_1, \bar{x}_2)$, such that for (x_1, x_2) close to $(\bar{x}_1, \bar{x}_2)$,

$$f(x_1, x_2) = f(\bar{x}_1, \bar{x}_2) + m_1(\bar{x}_1, \bar{x}_2)(x_1 - \bar{x}_1) + m_2(\bar{x}_1, \bar{x}_2)(x_2 - \bar{x}_2) + E_f(x, \bar{x}),$$

where

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

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 $f(x_1, ...x_d)$ of d real variables x_1,...,x_d, and we can speak about (and compute) partial derivatives of $f(x_1, ...x_d)$ with respect to x_1,...,x_d following the same basic idea. To compute the partial derivative $\frac{\partial f}{\partial x_j}$ with respect to x_j for some $j = 1, ..., d$, we keep all variables but x_j constant and compute the usual derivative with respect to x_j. We shall return below to the concept of partial derivative below, and through massive experience learn that it plays a basic role in mathematical modeling.

197.12 A Sum Up So Far

We have proved above that

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

We have also proved rules for how to differentiate linear combinations, products, quotients, compositions, and inverses of differentiable functions. This is just about all so far. We lack in particular answers to the following questions:

- What function $u(x)$ satisfies $u'(x) = \frac{1}{x}$?
- What is the derivative of the function a^x, where $a > 0$ is a constant?

Chapter 197 Problems

197.1. 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

$$\sqrt{x^{11} + \sqrt{\frac{x^{111}}{x^{-1.1} + x^{1.1}}}}.$$

197.2. Compute the partial derivatives of the function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined by $f(x_1, x_2) = x_1^2 + x_2^4$.

197.3. 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.

197.4. Suppose you know how to compute the derivative of 2^x at $x = 0$. What is the derivative then at $x \neq 0$? Hint: $2^{x+\frac{1}{n}} = 2^x 2^{\frac{1}{n}}$.

197.5. 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 $(0, 2)$?

197.6. Compute the partial derivatives of the function $f : \mathbb{R}^3 \to \mathbb{R}$ given by $f(x_1, x_2, x_3) = 2x_1^2 x_3 + 5x_2^3 x_3^4$.

198
Newton's Method

Brains first and then Hard Work. (The House at Pooh Corner, Milne)

198.1 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.

198.2 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 196.1 we know that $g(x)$ is Lipschitz continuous on I with Lipschitz constant L, and since $L < 1$, the function $g(x)$ has a unique fixed point $\bar{x} \in I$ satisfying $\bar{x} = g(\bar{x})$.

We know that $\bar{x} = \lim_{i\to\infty} x_i$, where $\{x_i\}_{i=1}^{\infty}$ is a sequence generated using Fixed Point Iteration: $x_{i+1} = g(x_i)$ for $i = 1, 2, ...$ To analyze the convergence of Fixed Point Iteration, we assume that $g(x)$ admits the fol-

lowing quadratic approximation close to $\bar{x}$ following the pattern of (197.26),

$$g(x) = g(\bar{x}) + g'(\bar{x})(x-\bar{x}) + \frac{1}{2}g''(\bar{x})(x-\bar{x})^2 + E_g(x,\bar{x}), \tag{198.1}$$

where $|E_g(x,\bar{x})| \le K_g(\bar{x})|x-\bar{x}|^3$. Choosing $x = x_i$, setting $e_i = x_i - \bar{x}$ and using $\bar{x} = g(\bar{x})$, we have for i large enough,

$$e_{i+1} = x_{i+1} - \bar{x} = g(x_i) - g(\bar{x}) = g'(\bar{x})e_i + \frac{1}{2}g''(\bar{x})e_i^2 + E_g(x_i,\bar{x}), \tag{198.2}$$

where $|E_g(x_i,\bar{x})| \le K_g(\bar{x})|e_i|^3$. 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

$$|e_{i+1}| \approx |g'(\bar{x})||e_i|, \tag{198.3}$$

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, (198.2) 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

$$|e_{i+1}| \approx \frac{1}{2}|g''(\bar{x})|e_i^2. \tag{198.4}$$

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.

198.3 Newton's Method

In Chapter Fixed Point Iteration, we saw that the problem of finding a root of an equation $f(x) = 0$, where $f(x)$ is a given function, can be reformulated as a fixed point equation $x = g(x)$, with $g(x) = x - \alpha f(x)$ and α a non-zero constant to choose. In fact, one may choose $\alpha(x)$ to depend in x and reformulate $f(x) = 0$ as

$$g(x) = x - \alpha(x)f(x),$$

if only $\alpha(\bar{x}) \neq 0$, where $\bar{x}$ is the root being computed. From above, we understand that a natural strategy is to choose α so as to make $g'(\bar{x})$ as

small as possible. The ideal would be $g'(\bar{x}) = 0$. Differentiating the equation $g(x) = x - \alpha(x)f(x)$ with respect to x, we get

$$g'(x) = 1 - \alpha'(x)f(x) - \alpha(x)f'(x).$$

Assuming that $f'(\bar{x}) \neq 0$, and using $f(\bar{x}) = 0$,

$$\alpha(\bar{x}) = \frac{1}{f'(\bar{x})}.$$

Setting $\alpha(x) = \frac{1}{f'(x)}$ leads to *Newton's method* for computing a root of $f(x) = 0$: for $i = 0, 1, 2, ...$

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}, \tag{198.5}$$

where x_0 is a given initial root approximation. Newton's method corresponds to Fixed Point Iteration with

$$g(x) = x - \frac{f(x)}{f'(x)}. \tag{198.6}$$

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 198.1. 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. 198.1. 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 nearby initial values converging to different roots.

198.4 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 (198.6):

$$g'(\bar{x}) = 1 - \frac{f'(\bar{x})^2 - f(\bar{x})f''(\bar{x})}{f'(\bar{x})^2} = \frac{f(\bar{x})f''(\bar{x})}{f'(\bar{x})^2} = 0,$$

where we used that $f(\bar{x}) = 0$ and the assumption that $f'(\bar{x}) \neq 0$. We conclude that Newton's method converges quadratically if $f'(\bar{x}) \neq 0$. This

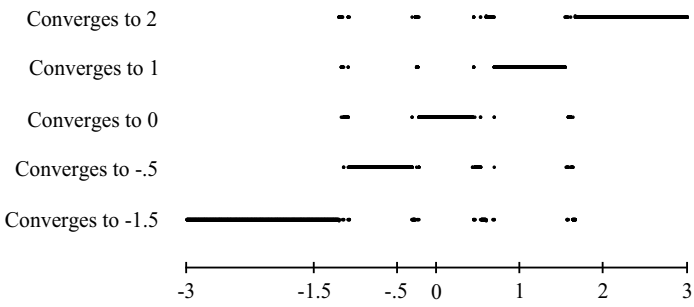


FIGURE 198.1. This plot shows the roots of $f(x) = (x-2)(x-1)x(x+.5)(x+1.5)$ found by Newton's method for 5000 equally spaced initial guesses in $[-3, 3]$. The horizontal position of the points shows the location of the initial guess and the vertical position indicates the twenty first Newton iterate.

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 (198.5) 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 $\bar{x}$.

198.5 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. 198.2, that is let x_{i+1} satisfy $f(x_i) + f'(x_i)(x_{i+1} - x_i) = 0$, so that

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}, \tag{198.7}$$

which is Newton's method. We conclude that the iterate x_{i+1} in Newton's method is the intersection of the tangent line to $f(x)$ at x_i with the x-axis. In words: trying to find $\bar{x}$, so that $f(\bar{x}) = 0$, we replace $f(x)$ by the linear approximation

$$\hat{f}(x) = f(x_i) + f'(x_i)(x - x_i),$$

that is by the tangent line at $x = x_i$, and then compute x_{i+1} as the solution of the equation $\hat{f}(x) = 0$. We shall find that this approach to Newton's method is easy to generalize to systems of equations corresponding to finding roots of $f(x)$ where $f : \mathbb{R}^n \to \mathbb{R}^n$.

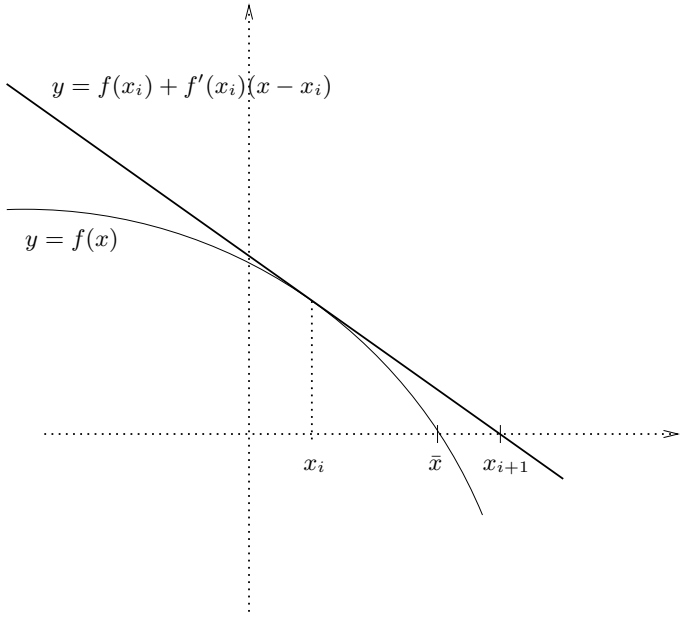


FIGURE 198.2. An illustration of one step of Newton's method from x_i to x_{i+1}.

198.6 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

$$x_i - \bar{x} \approx \frac{f(x_i)}{f'(\bar{x})}, \tag{198.8}$$

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. 198.3.

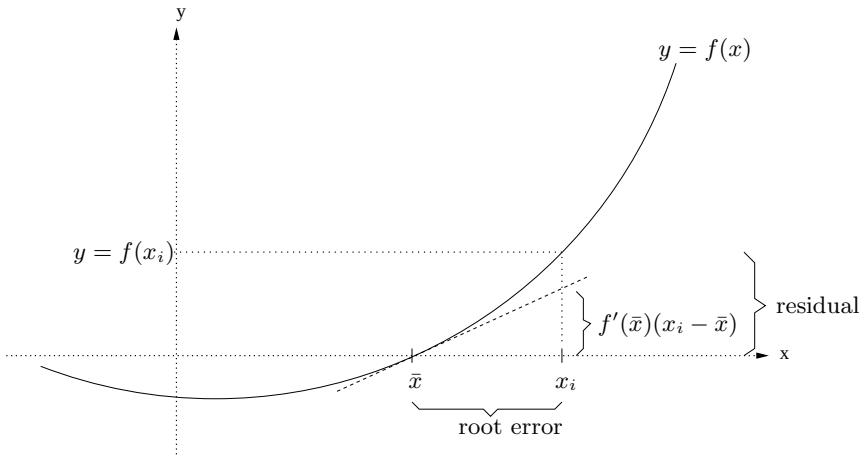


FIGURE 198.3. The root error and the residual

The relation (198.8) 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 198.1 *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 198.2. We apply Newton's method to $f(x) = (x-1)^2 - 10^{-15}x$ with root $\bar{x} \approx 1.00000003162278$. Here $f'(1) = -10^{-15}$ and

$f'(\bar{x}) \approx 0.0000000316$, so that $f'(x_n)$ is very small for all x_n close to $\bar{x}$, and the problem seems to be very ill-conditioned. We plot the errors and residuals versus iteration in Fig. 198.4. We see that the residuals become small quite a bit faster than the errors.

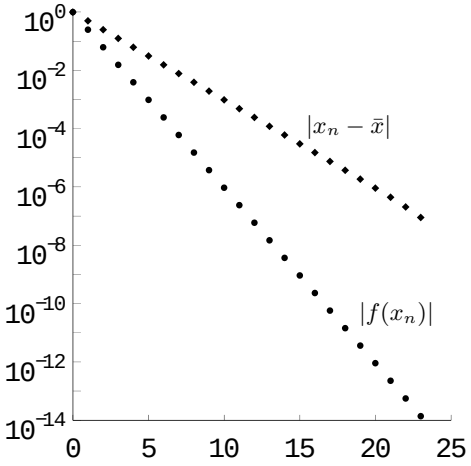


FIGURE 198.4. Plots of the residuals • and errors ♦ versus iteration number for Newton's method applied to $f(x) = (x-1)^2 - 10^{-15}x$ with initial value $x_0 = 1$

Introducing the approximation (198.8) into the definition of Newton's method,

$$x_{i+1} = x_i - f(x_i)/f'(x_i),$$

we get the relation

$$|x_i - \bar{x}| \approx |x_{i+1} - x_i|. \tag{198.9}$$

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.

EXAMPLE 198.3. We apply Newton's method to $f(x) = x^2 - 2$ and show the error and error estimate (198.9) in Fig. 198.5. The error estimate does a pretty good job.

198.7 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

i	$\lvert x_i - \bar{x}\rvert$	$\lvert x_{i+1} - x_i\rvert$
0	.586	.5
1	.086	.083
2	2.453×10^{-3}	2.451×10^{-3}
3	2.124×10^{-6}	2.124×10^{-6}
4	1.595×10^{-12}	1.595×10^{-12}
5	0	0

FIGURE 198.5. The error and error estimate for Newton's method for $f(x) = x^2 - 2$ with $x_0 = 2$.

other words, suppose we want to guarantee that

$$|x_i - \bar{x}| \leq TOL, \tag{198.10}$$

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 (198.10)? The following criteria based on (198.8) presents itself: stop the iterative process at step i if

$$|(f'(\bar{x}_i))^{-1} f(\bar{x}_i)| \leq TOL. \tag{198.11}$$

Up to the change of argument from $\bar{x}$ to $\bar{x}_i$, this criterion guarantees the desired error control (198.10).

As an alternative stopping criterion for Newton's method, we may use (198.9), that is accept the approximation x_i with tolerance TOL if

$$|x_{i+1} - x_i| \leq TOL. \tag{198.12}$$

198.8 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 198 Problems

198.1. (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.

198.2. (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?

198.3. (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.

198.4. Use Newton's method to compute all the roots of $f(x) = x^5 + 3x^4 - 3x^3 - 5x^2 + 5x - 1$.

198.5. Use Newton's method to compute the smallest positive root of $f(x) = \cos(x) + \sin(x)^2(50x)$.

198.6. 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.

198.7. 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 $f(x)$.

198.8. (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).

198.9. 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 $n = 1, 2, \cdots$.

198.10. 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.

199
The Integral

> The two questions, the first that of finding the description of the curve from its elements, the second that of finding the figure from the given differences, both reduce to the same thing. From this it can be taken that the whole of the theory of the inverse method of the tangents is reducible to quadratures. (Leibniz 1673)
>
> Utile erit scribit $\int$ pro omnia. (Leibniz, October 29 1675)

199.1 Primitive Functions and Integrals

In this chapter, we begin the study of the subject of *differential equations*, which is one of the common ties binding together all areas of science and engineering, and it would be hard to overstate its importance. We have been preparing for this chapter for a long time, starting from the beginning with Chapter *A very short course in Calculus*, through all of the chapters on functions, sequences, limits, real numbers, derivatives and basic differential equation models. So we hope the gentle reader is both excited and ready to embark on this new exploration.

We begin our study with the simplest kind of differential equation, which is of fundamental importance:

> Given the **function** $f : I \to \mathbb{R}$ defined on the interval $I = [a, b]$, find a **function** $u(x)$ on I, such that the derivative $u'(x)$ of $u(x)$ is equal to $f(x)$ for $x \in I$.

We can formulate this problem more concisely as: given $f : I \to \mathbb{R}$ find $u : I \to \mathbb{R}$ such that

$$u'(x) = f(x) \tag{199.1}$$

for all $x \in I$. We call the solution $u(x)$ of the differential equation $u'(x) = f(x)$ for $x \in I$, a *primitive function* of $f(x)$, or an *integral* of $f(x)$. Sometimes the term *antiderivative* is also used.

To understand what we mean by "solving" (199.1), we consider two simple examples. If $f(x)=1$ for $x \in \mathbb{R}$, then $u(x) = x$ is a solution of $u'(x) = f(x)$ for $x \in \mathbb{R}$, since $Dx = 1$ for all $x \in \mathbb{R}$. Likewise if $f(x) = x$, then $u(x) = x^2/2$ is a solution of $u'(x) = f(x)$ for $x \in \mathbb{R}$, since $Dx^2/2 = x$ for $x \in \mathbb{R}$. Thus the function x is a primitive function of the constant function 1, and $x^2/2$ is a primitive function of the function x.

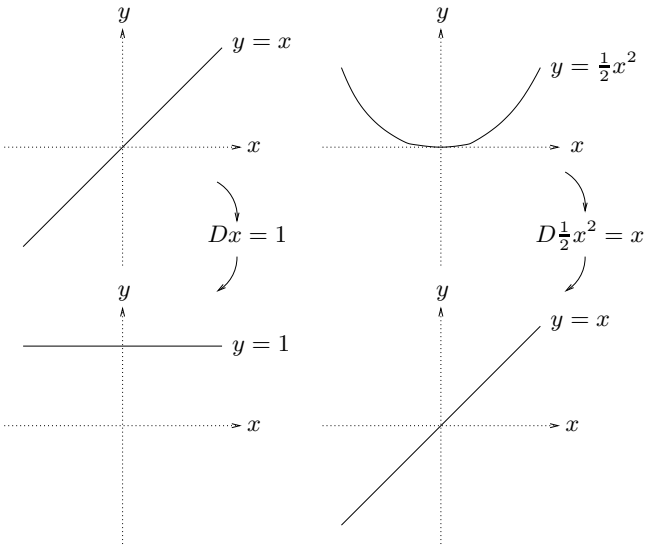


FIGURE 199.1. $Dx = 1$ and $D(x^2/2) = x$.

We emphasize that the solution of (199.1) is a **function** defined on an interval. We can interpret the problem in physical terms if we suppose that $u(x)$ represents some accumulated quantity like a sum of money in a bank, or an amount of rain, or the height of a tree, while x represents some changing quantity like time. Then solving (199.1) amounts to computing the total accumulated quantity $u(x)$ from knowledge of the rate of growth $u'(x) = f(x)$ at each instant x. This interpretation suggests that finding the total accumulated quantity $u(x)$ amounts to adding little pieces of momentary increments or changes of the quantity $u(x)$. Thus we expect that finding the integral $u(x)$ of a function $f(x)$ satisfying $u'(x) = f(x)$ will amount to some kind of *summation*.

A familiar example of this problem occurs when $f(x)$ is a velocity and x represents time so that the solution $u(x)$ of $u'(x) = f(x)$, represents the distance traveled by a body moving with instantaneous velocity $u'(x) = f(x)$. As the examples above show, we can solve this problem in simple cases, for example when the velocity $f(x)$ is equal to a constant v for all x and therefore the distance traveled during a time x is $u(x) = vx$. If we travel with constant velocity 4 miles/hour for two hours, then the distance traveled is 8 miles. We reach these 8 miles by accumulating distance foot-by-foot, which would be very apparent if we are walking!

An important observation is that the differential equation (199.1) alone is not sufficient information to determine the solution $u(x)$. Consider the interpretation when f represents velocity and u distance traveled by a body. If we want to know the position of the body, we need to know only the distance traveled but also the starting position. In general, a solution $u(x)$ to (199.1) is determined only up to a constant, because the derivative of a constant is zero. If $u'(x) = f(x)$, then also $(u(x) + c)' = f(x)$ for any constant c. For example, both $u(x) = x^2$ and $u(x) = x^2 + 1$ satisfy $u'(x) = 2x$. Graphically, we can see that there are many "parallel" functions that have the same slope at every point. The constant may be specified by

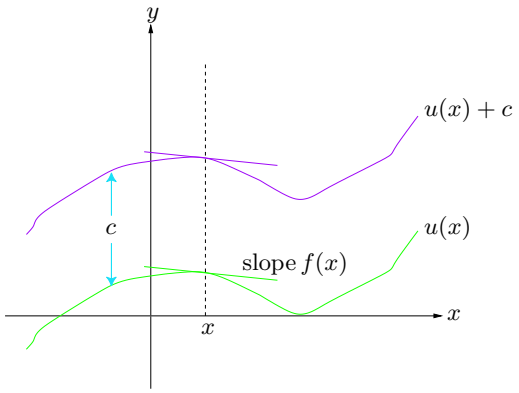


FIGURE 199.2. Two functions that have the same slope at every point.

specifying the value of the function $u(x)$ at some point. For example, the solution of $u'(x) = x$ is $u(x) = x^2 + c$ with c a constant, and specifying $u(0) = 1$ gives that $c = 1$.

More generally, we now formulate our basic problem as follows: Given $f : [a, b] \to \mathbb{R}$ and u_a, find $u : [a, b] \to \mathbb{R}$ such that

$$\begin{cases} u'(x) = f(x) & \text{for } a < x \leq b, \\ u(a) = u_a, \end{cases} \tag{199.2}$$

where u_a is a given *initial value*. The problem (199.2) is the simplest example of an *initial value problem* involving a differential equation and an initial value. The terminology naturally couples to situations in which x represents time and $u(a) = u_a$ amounts to specifying $u(x)$ at the initial time $x = a$. Note that we often keep the initial value terminology even if x represents a quantity different from time, and in case x represents a space coordinate we may alternatively refer to (199.2) as a *boundary value problem* with now $u(a) = u_a$ representing a given *boundary value*

We shall now prove that the initial value problem (199.2) has a unique solution $u(x)$ if the given function $f(x)$ is Lipschitz continuous on $[a, b]$. This is the *Fundamental Theorem of Calculus*, which stated in words says that a Lipschitz continuous function has a (unique) primitive function. Leibniz referred to the Fundamental Theorem as the "inverse method of tangents" because he thought of the problem as trying to find a curve $y = u(x)$ given the slope $u'(x)$ of its tangent at every point x.

We shall give a constructive proof of the Fundamental Theorem, which not only proves that $u : I \to \mathbb{R}$ exists, but also gives a way to compute $u(x)$ for any given $x \in [a, b]$ to any desired accuracy by computing a sum involving values of $f(x)$. Thus the version of the Fundamental Theorem we prove contains two results: (i) the existence of a primitive function and (ii) a way to compute a primitive function. Of course, (i) is really a consequence of (ii) since if we know how to compute a primitive function, we also know that it exists. These results are analogous to defining $\sqrt{2}$ by constructing a Cauchy sequence of approximate solutions of the equation $x^2 = 2$ by the Bisection algorithm. In the proof of the Fundamental Theorem we shall also construct a Cauchy sequence of approximate solutions of the differential equation (199.2) and show that the limit of the sequence is an exact solution of (199.2).

We shall express the solution $u(x)$ of (199.2) given by the Fundamental Theorem in terms of the data $f(x)$ and u_a as follows:

$$u(x) = \int_a^x f(y)\,dy + u_a \quad \text{for } a \le x \le b, \tag{199.3}$$

where we refer to

$$\int_a^x f(y)\,dy$$

as the *integral* of f over the interval $[a, x]$, a and x as the *lower and upper limits of integration* respectively, $f(y)$ as the *integrand* and y the *integration variable*. This notation was introduced on October 29 1675 by Leibniz, who thought of the integral sign $\int$ as representing "summation" and dy as the "increment" in the variable y. The notation of Leibniz is part of the big success of Calculus in science and education, and (like a good cover of a record) it gives a direct visual expression of the mathematical content of the integral in very suggestive form that indicates both the construction of

the integral and how to operate with integrals. Leibniz choice of notation plays an important role in making Calculus into a "machine" which "works by itself".

We recapitulate: There are two basic problems in Calculus. The first problem is to determine the derivative $u'(x)$ of a given function $u(x)$. We have met this problem above and we know a set of rules that we can use to attack this problem. The other problem is to find a function $u(x)$ given its derivative $u'(x)$. In the first problem we assume knowledge of $u(x)$ and we want to find $u'(x)$. In the second problem we assume knowledge of $u'(x)$ and we want to find $u(x)$.

As an interesting aside, the proof of the Fundamental Theorem also shows that the integral of a function over an interval may be interpreted as the area underneath the graph of the function over the interval. This couples the problem of finding a primitive function, or computing an integral, to that of computing an area, that is to *quadrature*. We expand on this geometric interpretation below.

Note that in (199.2), we require the differential equation $u'(x) = f(x)$ to be satisfied for x in the half-open interval $(a, b]$ excluding the left end-point $x = a$, where the differential equation is replaced by the specification $u(a) = u_a$. The proper motivation for this will become clear as we develop the proof of the Fundamental Theorem. Of course, the derivative $u'(b)$ at the right end-point $x = b$, is taken to be the left-hand derivative of u. By continuity, we will in fact have also $u'(a) = f(a)$, with $u'(a)$ the right-hand derivative.

199.2 Primitive Function of $f(x) = x^m$ for $m = 0, 1, 2, ...$

For some special functions $f(x)$, we can immediately find primitive functions $u(x)$ satisfying $u'(x) = f(x)$ for x in some interval. For example, if $f(x) = 1$, then $u(x) = x + c$, with c a constant, for $x \in \mathbb{R}$. Further, if $f(x) = x$, then $u(x) = x^2/2 + c$ for $x \in \mathbb{R}$. More generally, if $f(x) = x^m$, where $m = 0, 1, 2, 3, ..$, then $u(x) = x^{m+1}/(m + 1) + c$. Using the notation (199.3) for $x \in \mathbb{R}$ we write

$$\int_0^x 1\, dy = x, \quad \int_0^x y\, dy = \frac{x^2}{2}, \tag{199.4}$$

and more generally for $m = 0, 1, 2, ...,$

$$\int_0^x y^m\, dy = \frac{x^{m+1}}{m+1}, \tag{199.5}$$

because both right and left hand sides vanish for $x = 0$.

199.3 Primitive Function of $f(x) = x^m$ for $m = -2, -3, ...$

We recall that if $v(x) = x^{-n}$, where $n = 1, 2, 3, ..$, then $v'(x) = -nx^{-(n+1)}$, where now $x \neq 0$. Thus a primitive function of $f(x) = x^m$ for $m = -2, -3, ..$ is given by $u(x) = x^{m+1}/(m+1)$ for $x > 0$. We can state this fact as follows: For $m = -2, -3, ...,$

$$\int_1^x y^m \, dy = \frac{x^{m+1}}{m+1} - \frac{1}{m+1} \quad \text{for } x > 1, \tag{199.6}$$

where we start the integration arbitrarily at $x = 1$. The starting point really does not matter as long as we avoid 0. We have to avoid 0 because the function x^m with $m = -2, -3, ..,$ tends to infinity as x tends to zero. To compensate for starting at $x = 1$, we subtract the corresponding value of $x^{m+1}/(m+1)$ at $x = 1$ from the right hand side. We can write analogous formulas for $0 < x < 1$ and $x < 0$.

Summing up, we see that the polynomials x^m with $m = 0, 1, 2, ...,$ have the primitive functions $x^{m+1}/(m+1)$, which again are polynomials. Further, the rational functions x^m for $m = -2, -3, ...,$ have the primitive functions $x^{m+1}/(m+1)$, which again are rational functions.

199.4 Primitive Function of $f(x) = x^r$ for $r \neq -1$

Given our success so far, it would be easy to get overconfident. But we encounter a serious difficulty even with these early examples. Extending the previous arguments to rational powers of x, since $Dx^s = sx^{s-1}$ for $s \neq 0$ and $x > 0$, we have for $r = s - 1 \neq -1$,

$$\int_1^x y^r \, dy = \frac{x^{r+1}}{r+1} - \frac{1}{r+1} \quad \text{for } x > 1. \tag{199.7}$$

This formula breaks down for $r = -1$ and therefore we do not know a primitive function of $f(x) = x^r$ with $r = -1$ and moreover we don't even know that one exists. In fact, it turns out that most of the time we cannot solve the differential equation (199.2) in the sense of writing out $u(x)$ in terms of known functions. Being able to integrate simple rational functions is special. The Fundamental Theorem of Calculus will give us a way past this difficulty by providing the means to approximate the unknown solution to any desired accuracy.

199.5 A Quick Overview of the Progress So Far

Any function obtained by linear combinations, products, quotients and compositions of functions of the form x^r with rational power $r \neq 0$ and $x > 0$, can be differentiated analytically. If $u(x)$ is such a function, we thus obtain an analytical formula for $u'(x)$. If we now choose $f(x) = u'(x)$, then of course $u(x)$ satisfies the differential equation $u'(x) = f(x)$, so that we can write recalling Leibniz notation:

$$u(x) = \int_0^x f(y)\,dy + u(0) \quad \text{for } x \geq 0,$$

which apparently states that the function $u(x)$ is a primitive function of its derivative $f(x) = u'(x)$ (assuming that $u(x)$ is defined for all $x \geq 0$ so that no denominator vanishes for $x \geq 0$).

We give an example: Since $D(1+x^3)^{\frac{1}{3}} = (1+x^3)^{-\frac{2}{3}}x^2$ for $x \in \mathbb{R}$, we can write

$$(1+x^3)^{\frac{1}{3}} = \int_0^x \frac{y^2}{(1+y^3)^{\frac{2}{3}}}\,dy + 1 \quad \text{for } x \in \mathbb{R}.$$

In other words, we know primitive functions $u(x)$ satisfying the differential equation $u'(x) = f(x)$ for $x \in I$, for any function $f(x)$, which itself is a derivative of some function $v(x)$ so that $f(x) = v'(x)$ for $x \in I$. The relation between $u(x)$ and $v(x)$ is then

$$u(x) = v(x) + c \quad \text{for } x \in I,$$

for some constant c.

On the other hand, if $f(x)$ is an arbitrary function of another from, then we may not be able to produce an analytical formula for the corresponding primitive function $u(x)$ very easily or not at all. The Fundamental Theorem now tells us how to compute a primitive function of an arbitrary Lipschitz continuous function $f(x)$. We shall see that in particular, the function $f(x) = x^{-1}$ has a primitive function for $x > 0$ which is the famous *logarithm function* $\log(x)$. The Fundamental Theorem therefore gives in particular a constructive procedure for computing $\log(x)$ for $x > 0$.

199.6 A "Very Quick Proof" of the Fundamental Theorem

We shall now enter into the proof of the Fundamental Theorem. It is a good idea at this point to review the Chapter *A very short course in Calculus*. We shall give a sequence of successively more complete versions of the proof of the Fundamental Theorem with increasing precision and generality in each step.

The problem we are setting out to solve has the following form: given a function $f(x)$, find a function $u(x)$ such that $u'(x) = f(x)$ for all x in an interval. In this problem, we start with $f(x)$ and seek a function $u(x)$ such that $u'(x) = f(x)$. However in the early "quick" versions of the proofs, it will appear that we have turned the problem around by starting with a given function $u(x)$, differentiating u to get $f(x) = u'(x)$, and then recovering $u(x)$ as a primitive function of $f(x) = u'(x)$. This naturally appears to be quite meaningless circular reasoning, and some Calculus books completely fall into this trap. But we are doing this to make some points clear. In the final proof, we will in fact start with $f(x)$ and construct a function $u(x)$ that satisfies $u'(x) = f(x)$ as desired!

Let now $u(x)$ be differentiable on $[a, b]$, let $x \in [a, b]$, and let $a = y_0 < y_1 < ... < y_m = x$ be a *subdivision* of $[a, x]$ into subintervals $[a, y_1)$, $[y_1, y_2)$,...,$[y_{m-1}, x)$. By repeatedly subtracting and adding $u(y_j)$, we obtain the following identity which we refer to as a *telescoping sum* with the terms cancelling two by two:

$$\begin{aligned} u(x) - u(a) &= u(y_m) - u(y_0) \\ &= u(y_m) - u(y_{m-1}) + u(y_{m-1}) - u(y_{m-2}) + u(y_{m-2}) \\ &\qquad - \cdots + u(y_2) - u(y_1) + u(y_1) - u(y_0). \end{aligned} \tag{199.8}$$

We can write this identity in the form

$$u(x) - u(a) = \sum_{i=1}^{m} \frac{u(y_i) - u(y_{i-1})}{y_i - y_{i-1}}(y_i - y_{i-1}), \tag{199.9}$$

or as

$$u(x) - u(a) = \sum_{i=1}^{m} f(y_{i-1})(y_i - y_{i-1}), \tag{199.10}$$

if we set

$$f(y_{i-1}) = \frac{u(y_i) - u(y_{i-1})}{y_i - y_{i-1}} \quad \text{for } i = 1, ..., m. \tag{199.11}$$

Recalling the interpretation of the derivative as the ratio of the change in a function to a change in its input, we obtain our first version of the Fundamental Theorem as the following analog of (199.10) and (199.11):

$$u(x) - u(a) = \int_a^x f(y)\, dy \quad \text{where} \quad f(y) = u'(y) \quad \text{for } a < y < x.$$

In the integral notation, the sum $\sum$ corresponds to the integral sign $\int$, the increments $y_i - y_{i-1}$ correspond to dy, the y_{i-1} to the integration variable y, and the difference quotient $\frac{u(y_i)-u(y_{i-1})}{y_i-y_{i-1}}$ corresponds to the derivative $u'(y_{i-1})$.

This is the way that Leibniz was first led to the Fundamental Theorem at the age of 20 (without having studied any Calculus at all) as presented in his *Art of Combinations* from 1666.

Note that (199.8) expresses the idea that "the whole is equal to the sum of the parts" with "the whole" being equal to $u(x)-u(a)$ and the "parts" being the differences $(u(y_m)-u(y_{m-1}))$, $(u(y_{m-1})-u(y_{m-2}))$,...,$(u(y_2)-u(y_1))$ and $(u(y_1)-u(y_0))$. Compare to the discussion in Chapter *A very short Calculus course* including Leibniz' teen-age dream.

199.7 A "Quick Proof" of the Fundamental Theorem

We now present a more precise version of the above "proof". To exercise flexibility in the notation, which is a useful ability, we change notation slightly. Let $u(x)$ be uniformly differentiable on $[a,b]$, let $\bar{x}\in[a,b]$, and let $a=x_0<x_1<...<x_m=\bar{x}$ be a partition of $[a,\bar{x}]$. We thus change from y to x and from x to $\bar{x}$. With this notation x serves the role of a variable and $\bar{x}$ is a particular value of x. We recall the identity (199.9) in its new dress:

$$u(\bar{x})-u(a)=\sum_{i=1}^{m}\frac{u(x_i)-u(x_{i-1})}{x_i-x_{i-1}}(x_i-x_{i-1}). \tag{199.12}$$

By the uniform differentiability of u:

$$u(x_i)-u(x_{i-1})=u'(x_{i-1})(x_i-x_{i-1})+E_u(x_i,x_{i-1}),$$

where

$$|E_u(x_i,x_{i-1})|\le K_u(x_i-x_{i-1})^2, \tag{199.13}$$

with K_u a constant, we can write the identity as follows:

$$u(\bar{x})-u(a)=\sum_{i=1}^{m}u'(x_{i-1})(x_i-x_{i-1})+\sum_{i=1}^{m}E_u(x_i,x_{i-1}). \tag{199.14}$$

Setting h equal to the largest increment x_i-x_{i-1}, so that $x_i-x_{i-1}\le h$ for all i, we find

$$\sum_{i=1}^{m}|E_u(x_i,x_{i-1})|\le\sum_{i=1}^{m}K_u(x_i-x_{i-1})h=K_u(\bar{x}-a)h.$$

The formula (199.14) can thus be written

$$u(\bar{x})-u(a)=\sum_{i=1}^{m}u'(x_{i-1})(x_i-x_{i-1})+E_h, \tag{199.15}$$

where

$$|E_h| \le K_u(\bar{x} - a)h. \tag{199.16}$$

The Fundamental Theorem is the following analog of this formula:

$$u(\bar{x}) - u(a) = \int_a^{\bar{x}} u'(x)\,dx, \tag{199.17}$$

with the sum $\sum$ corresponding to the integral sign $\int$, the increments $x_i - x_{i-1}$ corresponding to dx, and x_i corresponding to the integration variable x. We see by (199.16) that the additional term E_h in (199.15) tends to zero as the maximal increment h tends to zero. We thus expect (199.17) to be a limit form of (199.15) as h tends to zero.

199.8 A Proof of the Fundamental Theorem of Calculus

We now give a full proof of the Fundamental theorem. We assume for simplicity that $[a, b] = [0, 1]$ and the initial value $u(0) = 0$. We comment on the general problem at the end of the proof. So the problem we consider is: Given a Lipschitz continuous function $f : [0, 1] \to \mathbb{R}$, find a solution $u(x)$ of the initial value problem,

$$\begin{cases} u'(x) = f(x) & \text{for } 0 < x \le 1, \\ u(0) = 0. \end{cases} \tag{199.18}$$

We shall now construct an approximation to the solution $u(x)$ and give a meaning to the solution formula

$$u(\bar{x}) = \int_0^{\bar{x}} f(x)\,dx \quad \text{for } 0 \le \bar{x} \le 1.$$

To this end, let n be a natural number and let $0 = x_0 < x_1 < ... < x_N = 1$ be the subdivision of the interval $[0, 1]$ with *nodes* $x_i^n = ih_n$, $i = 0, ..., N$, where $h_n = 2^{-n}$ and $N = 2^n$. We thus divide the given interval $[0, 1]$ into subintervals $I_i^n = (x_{i-1}^n, x_i^n]$ of equal lengths $h_n = 2^{-n}$, see Fig. 199.3.

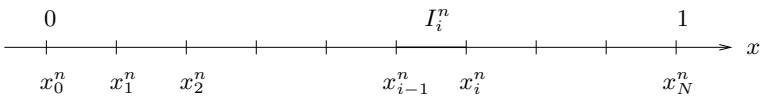


FIGURE 199.3. Subintervals I_i^n of lengths $h_n = 2^{-n}$

The approximation to $u(x)$ is a continuous piecewise linear function $U^n(x)$ defined by the formula

$$U^n(x_j^n) = \sum_{i=1}^{j} f(x_{i-1}^n)h_n \qquad \text{for } j = 1, ..., N, \tag{199.19}$$

where $U^n(0) = 0$. This formula gives the values of $U^n(x)$ at the nodes $x = x_j^n$ and we extend $U^n(x)$ linearly between the nodes to get the rest of the values, see Fig. 199.4.

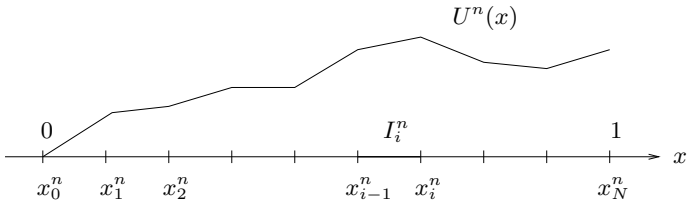


FIGURE 199.4. Piecewise linear function $U^n(x)$

We see that $U^n(x_j^n)$ is a sum of contributions $f(x_{i-1}^n)h_n$ for all intervals I_i^n with $i \leq j$. By construction,

$$U^n(x_i^n) = U^n(x_{i-1}^n) + f(x_{i-1}^n)h_n \quad \text{for } i = 1, ..., N, \tag{199.20}$$

so given the function $f(x)$, we can compute the function $U^n(x)$ by using the formula (199.20) successively with $i = 1, 2, ..., N$, where we first compute $U^n(x_1^n)$ using the value $U^n(x_0^n) = U^n(0) = 0$, then $U^n(x_2^n)$ using the value $U^n(x_1^n)$ and so on. We may alternatively use the resulting formula (199.19) involving summation, which of course just amounts to computing the sum by successively adding the terms of the sum.

The function $U^n(x)$ defined by (199.19) is thus a continuous piecewise linear function, which is computable from the nodal values $f(x_i^n)$, and we shall now motivate why $U^n(x)$ should have a good chance of being an approximation of a function $u(x)$ satisfying (199.18). If $u(x)$ is uniformly differentiable on $[0, 1]$, then

$$u(x_i^n) = u(x_{i-1}^n) + u'(x_{i-1}^n)h_n + E_u(x_i^n, x_{i-1}^n) \quad \text{for } i = 1, ..., N, \tag{199.21}$$

where $|E_u(x_i^n, x_{i-1}^n)| \leq K_u(x_i^n - x_{i-1}^n)^2 = K_u h_n^2$, and consequently

$$u(x_j^n) = \sum_{i=1}^{j} u'(x_{i-1}^n)h_n + E_h \quad \text{for } j = 1, ..., N, \tag{199.22}$$

where $|E_h| \leq K_u h_n$, since $\sum_{i=1}^{j} h_n = jh_n \leq 1$. Assuming that $u'(x) = f(x)$ for $0 < x \leq 1$, the connection between (199.20) and (199.21) and (199.19) and (199.22) becomes clear considering that the terms $E_u(x_i^n, x_{i-1}^n)$ and E_h are small. We thus expect $U^n(x_j^n)$ to be an approximation of $u(x_j^n)$ at the nodes x_j^n, and therefore $U^n(x)$ should be an increasingly accurate approximation of $u(x)$ as n increases and $h_n = 2^{-n}$ decreases.

To make this approximation idea precise, we first study the convergence of the functions $U^n(x)$ as n tends to infinity. To do this, we fix $\bar{x} \in [0, 1]$

and consider the sequence of numbers $\{U^n(\bar{x})\}_{n=1}^{\infty}$. We want to prove that this is a Cauchy sequence and thus we want to estimate $|U^n(\bar{x}) - U^m(\bar{x})|$ for $m > n$.

We begin by estimating the difference $|U^n(\bar{x}) - U^{n+1}(\bar{x})|$ for two consecutive indices n and $m = n + 1$. Recall that we used this approach in the proof of the Contraction Mapping theorem. We have

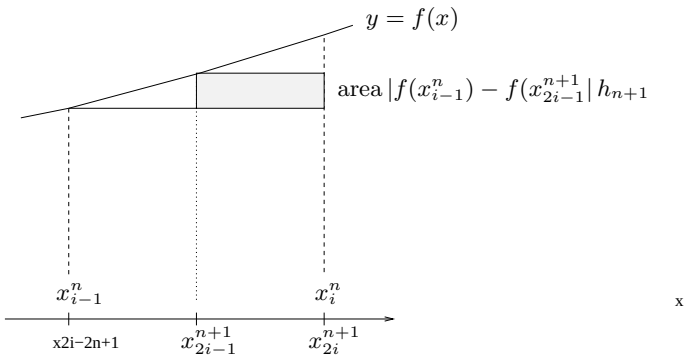


FIGURE 199.5. The difference between $U^{n+1}(x)$ and $U^n(x)$

$$U^n(x_i^n) = U^n(x_{i-1}^n) + f(x_{i-1}^n)h_n,$$

and since $x_{2i}^{n+1} = x_i^n$ and $x_{2i-2}^{n+1} = x_{i-1}^n$,

$$\begin{aligned} U^{n+1}(x_i^n) &= U^{n+1}(x_{2i}^{n+1}) = U^{n+1}(x_{2i-1}^{n+1}) + f(x_{2i-1}^{n+1})h_{n+1} \\ &= U^{n+1}(x_{i-1}^n) + f(x_{2i-2}^{n+1})h_{n+1} + f(x_{2i-1}^{n+1})h_{n+1}. \end{aligned}$$

Subtracting and setting $e_i^n = U^n(x_i^n) - U^{n+1}(x_i^n)$, we have

$$e_i^n = e_{i-1}^n + (f(x_{i-1}^n)h_n - f(x_{2i-2}^{n+1})h_{n+1} - f(x_{2i-1}^{n+1})h_{n+1}),$$

that is, since $h_{n+1} = \frac{1}{2}h_n$,

$$e_i^n - e_{i-1}^n = (f(x_{i-1}^n) - f(x_{2i-1}^{n+1}))h_{n+1}. \tag{199.23}$$

Assuming that $\bar{x} = x_j^n$ and using (199.23) and the facts that $e_0^n = 0$ and $|f(x_{i-1}^n) - f(x_{2i-1}^{n+1})| \le L_f h_{n+1}$, we get

$$\begin{aligned} |U^n(\bar{x}) - U^{n+1}(\bar{x})| = |e_j^n| &= |\sum_{i=1}^{j}(e_i^n - e_{i-1}^n)| \\ &\le \sum_{i=1}^{j}|e_i^n - e_{i-1}^n| = \sum_{i=1}^{j}|f(x_{i-1}^n) - f(x_{2i-1}^{n+1})|h_{n+1} \\ &\le \sum_{i=1}^{j} L_f h_{n+1}^2 = \frac{1}{4}L_f h_n \sum_{i=1}^{j} h_n = \frac{1}{4}L_f \bar{x} h_n, \end{aligned} \tag{199.24}$$

where we also used the fact that $\sum_{i=1}^{j} h_n = \bar{x}$. Iterating this estimate and using the formula for a geometric sum, we get

$$|U^n(\bar{x}) - U^m(\bar{x})| \le \frac{1}{4} L_f \bar{x} \sum_{k=n}^{m-1} h_k = \frac{1}{4} L_f \bar{x} (2^{-n} + ... + 2^{-m+1})$$
$$= \frac{1}{4} L_f \bar{x} 2^{-n} \frac{1 - 2^{-m+n}}{1 - 2^{-1}} \le \frac{1}{4} L_f \bar{x} 2^{-n} 2 = \frac{1}{2} L_f \bar{x} h_n,$$

that is

$$|U^n(\bar{x}) - U^m(\bar{x})| \le \frac{1}{2} L_f \bar{x} h_n. \tag{199.25}$$

This estimate shows that $\{U^n(\bar{x})\}_{n=1}^{\infty}$ is a Cauchy sequence and thus converges to a real number. We decide, following Leibniz, to denote this real number by

$$\int_0^{\bar{x}} f(x)\, dx,$$

which thus is the limit of

$$U^n(\bar{x}) = \sum_{i=1}^{j} f(x_{i-1}^n) h_n$$

as n tends to infinity, where $\bar{x} = x_j^n$. In other words,

$$\int_0^{\bar{x}} f(x)\, dx = \lim_{n \to \infty} \sum_{i=1}^{j} f(x_{i-1}^n) h_n.$$

Letting m tend to infinity in (199.25), we can express this relation in quantitative form as follows:

$$\Big| \int_0^{\bar{x}} f(x)\, dx - \sum_{i=1}^{j} f(x_{i-1}^n) h_n \Big| \le \frac{1}{2} L_f \bar{x} h_n.$$

At this point, we have defined the integral $\int_0^{\bar{x}} f(x)\, dx$ for a given Lipschitz continuous function $f(x)$ on $[0,1]$ and a given $\bar{x} \in [0,1]$, as the limit of the sequence $\{U^n(\bar{x})\}_{n=1}^{\infty}$ as n tends to infinity. We can thus define a function $u : [0,1] \to \mathbb{R}$ by the formula

$$u(\bar{x}) = \int_0^{\bar{x}} f(x)\, dx \quad \text{for } \bar{x} \in [0,1]. \tag{199.26}$$

We now proceed to check that the function $u(x)$ defined in this way indeed satisfies the differential equation $u'(x) = f(x)$. We proceed in two steps. First we show that the function $u(x)$ is Lipschitz continuous on $[0,1]$ and then we show that $u'(x) = f(x)$.

Before plunging into these proofs, we need to address a subtle point. Looking back at the construction of $u(x)$, we see that we have defined $u(\bar{x})$ for $\bar{x}$ of the form $\bar{x} = x_j^n$, where $j = 0, 1, ..., 2^n$, $n = 1, 2, ...,$. These are the rational numbers with finite decimal expansion in the base of 2, and they are *dense* in the sense that for any real number $x \in [0, 1]$ and any $\epsilon > 0$, there is a point of the form x_j^n so that $|x - x_j^n| \leq \epsilon$. Recalling the Chapter *Real numbers*, we understand that if we can show that $u(x)$ is Lipschitz continuous on the dense set of numbers of the form x_j^n, then we can extend $u(x)$ as a Lipschitz function to the set of real numbers $[0, 1]$.

We thus assume that $\bar{x} = x_j^n$ and $\bar{y} = x_k^n$ with $j > k$, and we note that

$$U^n(\bar{x}) - U^n(\bar{y}) = \sum_{i=1}^{j} f(x_{i-1}^n)h_n - \sum_{i=1}^{k} f(x_{i-1}^n)h_n = \sum_{i=k+1}^{j} f(x_{i-1}^n)h_n$$

and using the triangle inequality

$$|U^n(\bar{x}) - U^n(\bar{y})| \leq \sum_{i=k+1}^{j} |f(x_{i-1}^n)|h_n \leq M_f \sum_{i=k+1}^{j} h_n = M_f|\bar{x} - \bar{y}|,$$

where M_f is a positive constant such that $|f(x)| \leq M_f$ for all $x \in [0, 1]$. Letting n tend to infinity, we see that

$$u(\bar{x}) - u(\bar{y}) = \int_0^{\bar{x}} f(x)\,dx - \int_0^{\bar{y}} f(x)\,dx = \int_{\bar{y}}^{\bar{x}} f(x)\,dx, \tag{199.27}$$

where of course,

$$\int_{\bar{y}}^{\bar{x}} f(x)\,dx = \lim_{n\to\infty} \sum_{i=k+1}^{j} f(x_{i-1}^n)h_n,$$

and also

$$|u(\bar{x}) - u(\bar{y})| \leq |\int_{\bar{y}}^{\bar{x}} f(x)\,dx| \leq \int_{\bar{y}}^{\bar{x}} |f(x)|\,dx \leq M_f|\bar{x} - \bar{y}|, \tag{199.28}$$

where the second inequality is the so-called *triangle inequality for integrals* to be proved in the next section. We thus have

$$|u(\bar{x}) - u(\bar{y})| \leq M_f|\bar{x} - \bar{y}|, \tag{199.29}$$

which proves the Lipschitz continuity of $u(x)$.

We now prove that the function $u(x)$ defined for $x \in [0, 1]$ by the formula

$$u(x) = \int_a^x f(y)\,dy,$$

where $f : [0,1] \to \mathbb{R}$ is Lipschitz continuous, satisfies the differential equation

$$u'(x) = f(x) \quad \text{for } x \in [0,1],$$

that is

$$\frac{d}{dx}\int_0^x f(y)\,dy = f(x). \tag{199.30}$$

To this end, we choose $x, \bar{x} \in [0,1]$ with $x \geq \bar{x}$ and use (199.27) and (199.28) to see that

$$u(x) - u(\bar{x}) = \int_0^x f(z)dz - \int_0^{\bar{x}} f(y)dy = \int_{\bar{x}}^x f(y)dy,$$

and

$$\begin{aligned} |u(x) - u(\bar{x}) - f(\bar{x})(x-\bar{x})| &= \Big| \int_{\bar{x}}^x f(y)\,dy - f(\bar{x})(x-\bar{x}) \Big| \\ &= \Big| \int_{\bar{x}}^x (f(y) - f(\bar{x}))\,dy \Big| \leq \int_{\bar{x}}^x |f(y) - f(\bar{x})|\,dy \\ &\leq \int_{\bar{x}}^x L_f|y-\bar{x}|\,dy = \frac{1}{2}L_f(x-\bar{x})^2, \end{aligned}$$

where we again used the triangle inequality for integrals. This proves that u is uniformly differentiable on $[0,1]$, and that $K_u \leq \frac{1}{2}L_f$.

Finally to prove uniqueness, we recall from (199.15) and (199.16) that a function $u : [0,1] \to \mathbb{R}$ with Lipschitz continuous derivative $u'(x)$ and $u(0) = 0$, can be represented as

$$u(\bar{x}) = \sum_{i=1}^m u'(x_{i-1})(x_i - x_{i-1}) + E_h,$$

where

$$|E_h| \leq K_u(\bar{x} - a)h.$$

Letting n tend to infinity, we find that

$$u(\bar{x}) = \int_0^{\bar{x}} u'(x)\,dx \quad \text{for } \bar{x} \in [0,1], \tag{199.31}$$

which expresses the fact that a uniformly differentiable function with Lipschitz continuous derivative is the integral of its derivative. Suppose now that $u(x)$ and $v(x)$ are two uniformly differentiable functions on $[0,1]$ satisfying $u'(x) = f(x)$, and $v'(x) = f(x)$ for $0 < x \leq 1$, and $u(0) = u_0$, $v(0) = u_0$, where $f : [0,1] \to \mathbb{R}$ is Lipschitz continuous. Then the difference $w(x) = u(x) - v(x)$ is a uniformly differentiable function on $[0,1]$ satisfying $w'(x) = 0$ for $a < x \leq b$ and $w(0) = 0$. But we just showed that

$$w(x) = \int_a^x w'(y)\,dy,$$

and thus $w(x) = 0$ for $x \in [0,1]$. This proves that $u(x) = v(x)$ for $x \in [0,1]$ and the uniqueness follows.

Recall that we proved the Fundamental Theorem for special circumstances, namely on the interval $[0,1]$ with initial value 0. We can directly generalize the construction above by replacing $[0,1]$ by an arbitrary bounded interval $[a,b]$, replacing h_n by $h_n = 2^{-n}(b-a)$, and assuming instead of $u(0) = 0$ that $u(a) = u_a$, where u_a is a given real number. We have now proved the formidable Fundamental Theorem of Calculus.

Theorem 199.1 (Fundamental Theorem of Calculus) *If $f : [a,b] \to \mathbb{R}$ is Lipschitz continuous, then there is a unique uniformly differentiable function $u : [a,b] \to \mathbb{R}$, which solves the initial value problem*

$$\begin{cases} u'(x) = f(x) & \text{for } x \in (a,b], \\ u(a) = u_a, \end{cases} \tag{199.32}$$

where $u_a \in \mathbb{R}$ is given. The function $u : [a,b] \to \mathbb{R}$ can be expressed as

$$u(\bar{x}) = u_a + \int_a^{\bar{x}} f(x)\,dx \quad \text{for } \bar{x} \in [a,b],$$

where

$$\int_0^{\bar{x}} f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{j} f(x_{i-1}^n)h_n,$$

with $\bar{x} = x_j^n$, $x_i^n = a + ih_n$, $h_n = 2^{-n}(b-a)$. More precisely, if the Lipschitz constant of $f : [a,b] \to \mathbb{R}$ is L_f, then for $n = 1, 2, ...,$

$$\Big| \int_a^{\bar{x}} f(x)\,dx - \sum_{i=1}^{j} f(x_{i-1}^n)h_n \Big| \leq \frac{1}{2}(\bar{x} - a)L_f h_n. \tag{199.33}$$

Furthermore if $|f(x)| \leq M_f$ for $x \in [a,b]$, then $u(x)$ is Lipschitz continuous with Lipschitz constant M_f and $K_u \leq \frac{1}{2}L_f$, where K_u is the constant of uniform differentiability of $u : [a,b] \to \mathbb{R}$.

199.9 Comments on the Notation

We can change the names of the variables and rewrite (199.26) as

$$u(x) = \int_0^x f(y)\,dy. \tag{199.34}$$

We will often use the Fundamental Theorem in the form

$$\int_a^b u'(x)\,dx = u(b) - u(a), \tag{199.35}$$

which states that the integral $\int_a^b f(x)\,dx$ is equal to the difference $u(b) - u(a)$, where $u(x)$ is a primitive function of $f(x)$. We will sometimes use the notation $[u(x)]_{x=a}^{x=b} = u(b) - u(a)$ or shorter $[u(x)]_a^b = u(b) - u(a)$, and write

$$\int_a^b u'(x)\,dx = \big[u(x)\big]_{x=a}^{x=b} = \big[u(x)\big]_a^b.$$

Sometimes the notation

$$\int f(x)\,dx,$$

without limits of integration, is used to denote a primitive function of $f(x)$. With this notation we would have for example

$$\int dx = x + C, \quad \int x\,dx = \frac{x^2}{2} + C, \quad \int x^2\,dx = \frac{x^3}{3} + C,$$

where C is a constant. We will not use this notation in this book. Note that the formula $x = \int dx$ may be viewed to express that "the whole is equal to the sum of the parts".

199.10 Alternative Computational Methods

Note that we might as well compute $U^n(x_i^n)$ from knowledge of $U^n(x_{i-1}^n)$, using the formula

$$U^n(x_i^n) = U^n(x_{i-1}^n) + f(x_i^n)h_n, \tag{199.36}$$

obtained by replacing $f(x_{i-1}^n)$ by $f(x_i^n)$, or

$$U^n(x_i^n) = U^n(x_{i-1}^n) + \frac{1}{2}(f(x_{i-1}^n) + f(x_i^n))h_n \tag{199.37}$$

using the mean value $\frac{1}{2}(f(x_{i-1}^n) + f(x_i^n))$. These alternatives may bring certain advantages, and we will return to them in Chapter *Numerical quadrature*. The proof of the Fundamental Theorem is basically the same with these variants and by uniqueness all the alternative constructions give the same result.

199.11 The Cyclist's Speedometer

An example of a physical situation modeled by the initial value problem (199.2) is a cyclist biking along a straight line with $u(x)$ representing the position at time x, $u'(x)$ being the speed at time x and specifying the position $u(a) = u_a$ at the initial time $x = a$. Solving the differential equation

(199.2) amounts to determining the position $u(x)$ of the cyclist at time $a < x \leq b$, after specifying the position at the initial time $x = a$ and knowing the speed $f(x)$ at each time x. A standard bicycle speedometer may be viewed to solve this problem, viewing the speedometer as a device which measures the instantaneous speed $f(x)$, and then outputs the total traveled distance $u(x)$. Or is this a good example? Isn't it rather so that the speedometer measures the traveled distance and then reports the momentary (average) speed? To answer this question would seem to require a more precise study of how a speedometer actually works, and we urge the reader to investigate this problem.

199.12 Geometrical Interpretation of the Integral

In this section, we interpret the proof of the Fundamental Theorem as saying that the integral of a function is the area underneath the graph of the function. More precisely, the solution $u(\bar{x})$ given by (199.3) is equal to the area under the graph of the function $f(x)$ on the interval $[a, \bar{x}]$, see Fig. 199.6. For the purpose of this discussion, it is natural to assume that $f(x) \geq 0$.

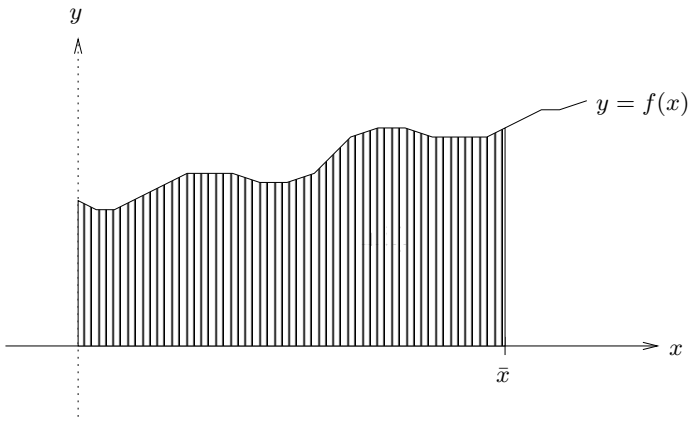


FIGURE 199.6. Area under $y = f(x)$.

Of course, we also have to explain what we mean by the area under the graph of the function $f(x)$ on the interval $[a, \bar{x}]$. To do this, we first interpret the approximation $U^n(\bar{x})$ of $u(\bar{x})$ as an area. We recall from the previous section that

$$U^n(x_j^n) = \sum_{i=1}^{j} f(x_{i-1}^n)h_n,$$

where $x_j^n = \bar{x}$. Now, we can view $f(x_{i-1}^n)h_n$ as the area of a rectangle with base h_n and height $f(x_{i-1}^n)$, see Fig. 199.7.

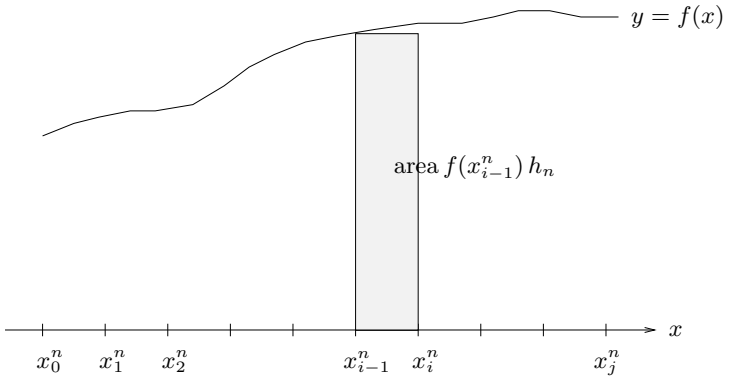


FIGURE 199.7. Area $f(x_{i-1}^n)\, h_n$ of rectangle.

We can thus interpret the sum

$$\sum_{i=1}^{j} f(x_{i-1}^n)h_n$$

as the area of a collection of rectangles which form a staircase approximation of $f(x)$, as displayed in Fig. 199.8. The sum is also referred to as a *Riemann sum*.

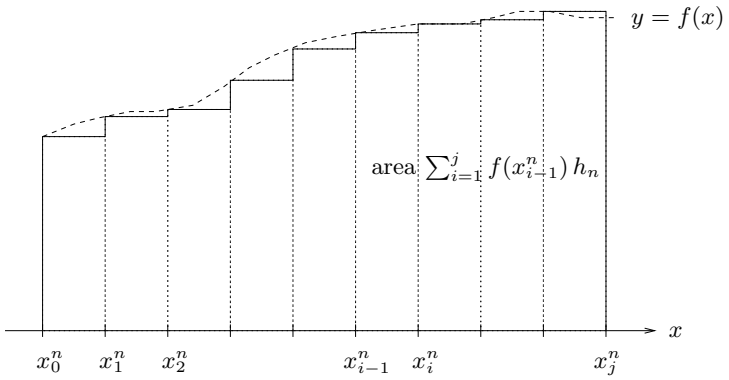


FIGURE 199.8. Area $\sum_{i=1}^{j} f(x_{i-1}^n)\, h_n$ under a staircase approximation of $f(x)$.

Intuitively, the area under the staircase approximation of $f(x)$ on $[a, \bar{x}]$, which is $U^n(\bar{x})$, will approach the area under the graph of $f(x)$ on $[a, \bar{x}]$ as n tends to infinity and $h_n = 2^{-n}(b-a)$ tends to zero. Since $\lim_{n\to\infty} U^n(\bar{x}) =$

$u(\bar{x})$, this leads us to *define* the area under $f(x)$ on the interval $[0, \bar{x}]$ as the limit $u(\bar{x})$.

Note the logic used here: The value $U^n(\bar{x})$ represents the area under a staircase approximation of $f(x)$ on $[a, \bar{x}]$. We know that $U^n(\bar{x})$ tends to $u(\bar{x})$ as n tends to infinity, and on intuitive grounds we feel that the limit of the area under the staircase should be equal to the area under the graph of $f(x)$ on $[a, \bar{x}]$. We then simply define the area under $f(x)$ on $[a, \bar{x}]$ to be $u(\bar{x})$. By definition we thus interpret the integral of $f(x)$ on $[0, \bar{x}]$ as the area under the graph of the function $f(x)$ on $[a, \bar{x}]$. Note that *this is an interpretation*. It is not a good idea to say the integral *is* an area. This is because the integral can represent many things, such as a distance, a quantity of money, a weight, or some thing else. If we interpret the integral as an area, then we also interpret a distance, a quantity of money, a weight, or some thing else, as an area. We understand that we cannot take this interpretation to be literally true, because a distance cannot *be equal* to an area, but it can be *interpreted* as an area. We hope the reader gets the (subtle) difference.

As an example, we compute the area A under the graph of the function $f(x) = x^2$ between $x = 0$ and $x = 1$ as follows

$$A = \int_0^1 x^2\,dx = [\frac{x^3}{3}]_{x=0}^{x=1} = \frac{1}{3}.$$

This is an example of the magic of Calculus, behind its enormous success. We were able to compute an area, which in principle is the sum of very many very small pieces, without actually having to do the tedious and laborious computation of the sum. We just found a primitive function $u(x)$ of x^2 and computed $A = u(3) - u(0)$ without any effort at all. Of course we understand the telescoping sum behind this illusion, but if you don't see this, you must be impressed, right? To get a perspective and close a circle, we recall the material in Leibniz' teen-age dream in Chapter *A very short course in Calculus*.

199.13 The Integral as a Limit of Riemann Sums

The Fundamental Theorem of Calculus states that the integral of $f(x)$ over the interval $[a, b]$ is equal to a limit of Riemann sums:

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{2^n} f(x_{i-1}^n)h_n,$$

where $x_i^n = a + ih_n$, $h_n = 2^{-n}(b - a)$, or more precisely, for $n = 1, 2, ...,$

$$\Big| \int_a^b f(x)\,dx - \sum_{i=1}^{2^n} f(x_{i-1}^n)h_n \Big| \le \frac{1}{2}(b - a)L_f h_n,$$

where L_f is the Lipschitz constant of f. We can thus define the integral $\int_a^b f(x)\,dx$ as a limit of Riemann sums without invoking the underlying differential equation $u'(x) = f(x)$. This approach is useful in defining integrals of functions of several variables (so-called multiple integrals like double integrals and triple integrals), because in these generalizations there is no underlying differential equation.

In our formulation of the Fundamental Theorem of Calculus, we emphasized the coupling of the integral $\int_a^x f(y)\,dy$ to the related differential equation $u'(x) = f(x)$, but as we just said, we could put this coupling in the back-ground, and define the integral as a limit of Riemann sums without invoking the underlying differential equation. This connects with the idea that the integral of a function can be interpreted as the area under the graph of the function, and will find a natural extension to multiple integrals in Chapters *Double integrals* and *Multiple integrals*.

Defining the integral as a limit of Riemann sums poses a question of uniqueness: since there are different ways of constructing Riemann sums one may ask if all limits will be the same. We will return to this question in Chapter *Numerical quadrature* and (of course) give an affirmative answer.

199.14 An Analog Integrator

James Thompson, brother of Lord Kelvin, constructed in 1876 an analog mechanical integrator based on a rotating disc coupled to a cylinder through another orthogonal disc adjustable along the radius of the first disc, see Fig. 199.9. The idea was to get around the difficulties of realizing the Analytical Engine, the mechanical digital computer envisioned by Babbage in the 1830s. Lord Kelvin tried to use a system of such analog integrators to compute different problems of practical interest including that of tide prediction, but met serious problems to reach sufficient accuracy. Similar ideas ideas were taken up by Vannevar Bush at MIT Massachussetts Institute of Technology in the 1930s, who constructed a *Differential Analyzer* consisting of a collection of analog integrators, which was programmable to solve differential equations, and was used during the Second World War for computing trajectories of projectiles. A decade later the digital computer took over the scene, and the battle between arithmetic and geometry initiated between the Pythagorean and Euclidean schools more than 2000 years ago, had finally come an end.

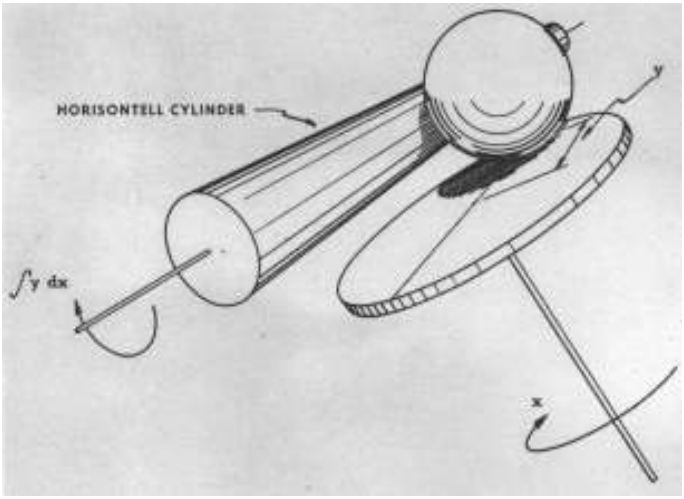


FIGURE 199.9. The principle of an Analog Integrator

Chapter 199 Problems

199.1. Determine primitive functions on $\mathbb{R}$ to (a) $(1+x^2)^{-2}2x$, (b) $(1+x)^{-99}$, (c) $(1+(1+x^3)^2)^{-2}2(1+x^3)3x^2$.

199.2. Compute the area under the graph of the function $(1+x)^{-2}$ between $x=1$ and $x=2$.

199.3. A car travels along the x-axis with speed $v(t)=t^{\frac{3}{2}}$ starting at $x=0$ for $t=0$. Compute the position of the car for $t=10$.

199.4. Carry out the proof of the Fundamental Theorem for the variations (199.36) and (199.37).

199.5. Construct a *mechanical integrator* solving the differential equation $u'(x)=f(x)$ for $x>0$, $u(0)=0$ through an analog mechanical devise. Hint: Get hold of a rotating cone and a string.

199.6. Explain the principle behind Thompson's analog integrator.

199.7. Construct a *mechanical* speedometer reporting the speed and traveled distance. Hint: Check the construction of the speedometer of your bike.

199.8. Find the solutions of the initial value problem $u'(x)=f(x)$ for $x>0$, $u(0)=1$, in the following cases: (a) $f(x)=0$, (b) $f(x)=1$, (c) $f(x)=x^r$, $r>0$.

199.9. Find the solution to the second order initial value problem $u''(x)=f(x)$ for $x>0$, $u(0)=u'(0)=1$, in the following cases: (a) $f(x)=0$, (b) $f(x)=1$, (c) $f(x)=x^r$, $r>0$. Explain why two initial conditions are specified.

199.10. Solve initial value problem $u'(x) = f(x)$ for $x \in (0, 2]$, $u(0) = 1$, where $f(x) = 1$ for $x \in [0, 1)$ and $f(x) = 2$ for $x \in [1, 2]$. Draw a graph of the solution and calculate $u(3/2)$. Show that $f(x)$ is not Lipschitz continuous on $[0, 2]$ and determine if $u(x)$ is Lipschitz continuous on $[0, 2]$.

199.11. The time it takes for a light beam to travel through an object is $t = \frac{d}{c/n}$, where c is the speed of light in vacuum, n is the refractive index of the object and d is its thickness. How long does it take for a light beam to travel the shortest way through the center of a glass of water, if the refractive index of the water varies as a certain function $n_w(r)$ with the distance r from the center of glass, the radius of the glass is R and the thickness and that the walls have constant thickness h and constant refractive index equal to n_g.

199.12. Assume that f and g are Lipschitz continuous on $[0, 1]$. Show that $\int_0^1 |f(x) - g(x)|dx = 0$ if and only if $f = g$ on $[0, 1]$. Does this also hold if $\int_0^1 |f(x) - g(x)|dx$ is replaced by $\int_0^1 (f(x) - g(x))dx$?

FIGURE 199.10. David Hilbert (1862-1943) at the age of 24: "A mathematical theory is not to be considered complete until you have made it so clear that you can explain it to the first man whom you meet on the street".

200
Properties of the Integral

For more than two thousand years some familiarity with mathematics has been regarded as an indispensable part of the intellectual equipment of every cultured person. Today the traditional place of mathematics in education is in great danger. Unfortunately, professional representatives of mathematics share the responsibility. The teaching of mathematics has sometimes degenerated into empty drill in problem solving, which may develop formal ability but does not lead to real understanding or to greater intellectual independence..... Teachers, students and the general public demand constructive reform, not resignation along the lines of least resistance. (Richard Courant, in Preface to What is Mathematics?, 1941)

200.1 Introduction

In this chapter, we gather together various useful properties of the integral. We may prove these properties in two ways: (i) by using the connection between the integral and the derivative and using properties of the derivative, or (ii) using that the integral is the limit of Riemann sum approximations, that is, using the area interpretation of the integral. We indicate both types of proofs to help the reader getting familiar with different aspects of the integral, and leave some of the work to the problem section.

Throughout the chapter we assume that $f(x)$ and $g(x)$ are Lipschitz continuous on the interval $[a, b]$, and we assume that

$$\sum_{i=1}^{N} f(x_{i-1}^n)h_n \qquad \text{and} \qquad \sum_{i=1}^{N} g(x_{i-1}^n)h_n$$

are Riemann sum approximations of $\int_a^b f(x)\,dx$ and $\int_a^b g(x)\,dx$ with step length $h_n = 2^{-n}(b-a)$ and $x_i^n = a + ih_n$, $i = 0, 1, ..., N = 2^n$, as in the previous chapter.

200.2 Reversing the Order of Upper and Lower Limits

So far we have defined the integral $\int_a^b f(x)\,dx$ assuming that $a \le b$, that is that the upper limit of integration b is larger than (or equal to) the lower limit a. It is useful to *extend* the definition to cases with $a > b$ by defining

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx. \tag{200.1}$$

In other words, we decide that switching the limits of integration should change the sign of an integral. As a motivation we may consider the case $f(x) = 1$ and $a > b$, and recall that $\int_b^a 1\,dx = a - b > 0$. Using the same formula with a and b interchanged, we would have $\int_a^b 1\,dx = b - a = -(a-b) = -\int_b^a 1\,dx$, which motivates the sign change under the switch of upper and lower limits. The motivation carries over to the general case using the Riemann sum approximation. Notice that here we do not *prove* anything, we simply introduce a *definition*. Of course we seek a definition which is natural, easy to remember and which allows efficient symbolic computation. The definition we chose fulfills these conditions.

Example 200.1. We have

$$\int_2^1 2x\,dx = -\int_1^2 2x\,dx = -\left[x^2\right]_1^2 = -(4-1) = -3.$$

200.3 The Whole Is Equal to the Sum of the Parts

We shall now prove that if $a \le c \le b$, then

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx. \tag{200.2}$$

One way to do this is to use the area interpretation of the integral and simply notice that the area under $f(x)$ from a to b should be equal to the sum of the area under $f(x)$ from a to c and the area under $f(x)$ from c to b.

We can also give an alternative proof using that that $\int_a^b f(x)\,dx = u(b)$, where $u(x)$ satisfies $u'(x) = f(x)$ for $a \le x \le b$, and $u(a) = 0$. Letting now $w(x)$ satisfy $w'(x) = f(x)$ for $c \le x \le b$, and $w(c) = u(c)$, we have by uniqueness that $w(x) = u(x)$ for $c \le x \le b$, and thus

$$u(b) = w(b) = u(c) + \int_c^b f(y)\,dy = \int_a^c f(y)\,dy + \int_c^b f(y)\,dy,$$

which is the desired result.

EXAMPLE 200.2. We have

$$\int_0^2 x\,dx = \int_0^1 x\,dx + \int_1^2 x\,dx,$$

which expresses the identity

$$2 = \left(\frac{1}{2}\right) + \left(2 - \frac{1}{2}\right).$$

Note that by the definition (200.1), (200.2) actually holds for any a, b and c.

200.4 Integrating Piecewise Lipschitz Continuous Functions

A functions is said to be *piecewise Lipschitz continuous* on a finite interval $[a, b]$ if $[a, b]$ can be divided up into a finite number of sub-intervals on which the function is Lipschitz continuous, allowing the function to have jumps at the ends of the subintervals, see Fig. 200.1.

We now extend (in the obvious way) the definition of the integral $\int_a^b f(x)\,dx$ to a piecewise Lipschitz continuous function $f(x)$ on an interval $[a, b]$ starting with the case of two subintervals with thus $f(x)$ Lipschitz continuous separately on two adjoining intervals $[a, c]$ and $[c, b]$, where $a \le c \le b$. We define

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx,$$

which obviously fits with (200.2). The extension is analogous for several subintervals with the integral over the whole interval being the sum of the integrals over the subintervals.

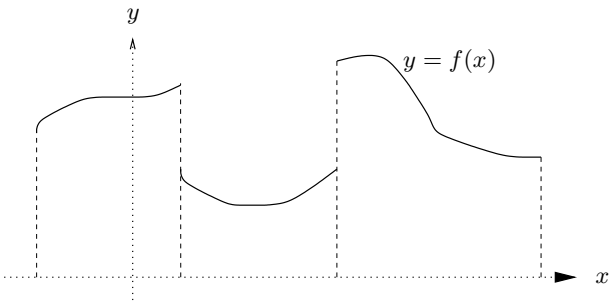


FIGURE 200.1. A piecewise Lipschitz continuous function.

200.5 Linearity

We shall now prove the following property of *linearity* of the integral: If α and β are real numbers then,

$$\int_a^b (\alpha f(x) + \beta g(x))\,dx = \alpha \int_a^b f(x)\,dx + \beta \int_a^b g(x)\,dx. \tag{200.3}$$

With $\alpha = \beta = 1$ this property expresses that the area (from a to b) underneath the sum of two functions is equal to the sum of the areas underneath each function. Further, with $g(x) = 0$ and $\alpha = 2$ say, the property expresses that the area under the function $2f(x)$ is equal to 2 times the area under the function $f(x)$.

More generally, the linearity of the integral is inherited from the linearity of the Riemann sum approximation, which we may express as

$$\sum_{i=1}^{N} (\alpha f(x_{i-1}^n) + \beta g(x_{i-1}^n) h_n = \alpha \sum_{i=1}^{N} f(x_{i-1}^n) h_n + \beta \sum_{i=1}^{N} g(x_{i-1}^n) h_n, \tag{200.4}$$

and which follows from basic rules for computing with real numbers.

A differential equation proof goes as follows: Define

$$u(x) = \int_a^x f(y)\,dy \quad \text{and} \quad v(x) = \int_a^x g(y)\,dy, \tag{200.5}$$

that is, $u(x)$ is a primitive function of $f(x)$ satisfying $u'(x) = f(x)$ for $a < x \leq b$ and $u(a) = 0$, and $v(x)$ is a primitive function of $g(x)$ satisfying $v'(x) = g(x)$ for $a < x \leq b$ and $v(a) = 0$. Now, the function $w(x) = \alpha u(x) + \beta v(x)$ is a primitive function of the function $\alpha f(x) + \beta g(x)$, since by the linearity of the derivative, $w'(x) = \alpha u'(x) + \beta v'(x) = \alpha f(x) + \beta g(x)$, and $w(a) = \alpha u(a) + \beta v(a) = 0$. Thus, the left hand side of (200.3) is equal to $w(b)$, and since $w(b) = \alpha u(b) + \beta v(b)$, the desired equality follows from setting $x = b$ in (200.5).

EXAMPLE 200.3. We have

$$\int_0^b (2x + 3x^2)\,dx = 2\int_0^b x\,dx + 3\int_0^b x^2\,dx = 2\frac{b^2}{2} + 3\frac{b^3}{3} = b^2 + b^3.$$

200.6 Monotonicity

The *monotonicity* property of the integral states that if $f(x) \geq g(x)$ for $a \leq x \leq b$, then

$$\int_a^b f(x)\,dx \geq \int_a^b g(x)\,dx. \tag{200.6}$$

This is the same as stating that if $f(x) \geq 0$ for $x \in [a, b]$, then

$$\int_a^b f(x)\,dx \geq 0, \tag{200.7}$$

which follows from the fact that all Riemann sum approximations $\sum_{i=1}^j f(x_{i-1}^n)h_n$ of $\int_a^b f(x)\,dx$ are all non-negative if $f(x) \geq 0$ for $x \in [a, b]$.

200.7 The Triangle Inequality for Integrals

We shall now prove the following *triangle inequality for integrals*:

$$\Big|\int_a^b f(x)\,dx\Big| \leq \int_a^b |f(x)|\,dx, \tag{200.8}$$

stating that moving the absolute value inside the integral increases the value (or leaves the value unchanged). This property follows from applying the usual triangle inequality to Riemann sum approximations to get

$$\Big|\sum_{i=1}^N f(x_{i-1}^n)h_n\Big| \leq \sum_{i=1}^N \Big|f(x_{i-1}^n)\Big|h_n$$

and then passing to the limit. Evidently there may be cancellations on the left hand side if $f(x)$ has changes sign, while on the right hand side we always add nonnegative contributions, making the right hand side at least as big as the left hand side.

Another proof uses the monotonicity as follows: Apply (200.7) to the function $|f| - f \geq 0$ to obtain

$$\int_a^{\bar{x}} f(x)\,dx \leq \int_a^{\bar{x}} |f(x)|\,dx.$$

Replacing f by the function $-f$ we obtain

$$-\int_a^{\bar{x}} f(x)\,dx = \int_a^{\bar{x}} (-f(x))\,dx \leq \int_a^{\bar{x}} |-f(x)|\,dx = \int_a^{\bar{x}} |f(x)|\,dx,$$

which proves the desired result.

200.8 Differentiation and Integration Are Inverse Operations

The Fundamental Theorem says that integration and differentiation are *inverse operations* in the sense that first integrating and then differentiating, or first differentiating and then integrating, gives the net result of doing nothing! We make this clear by repeating a part of the proof of the Fundamental Theorem to prove that if $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous then for $x \in [a, b]$,

$$\frac{d}{dx}\int_a^x f(y)\,dy = f(x). \tag{200.9}$$

In other words, integrating a function $f(x)$ and then differentiating the primitive function, gives back the function $f(x)$. Surprise? We illustrate in Fig. 200.2. To properly understand the equality (200.9), it is important to

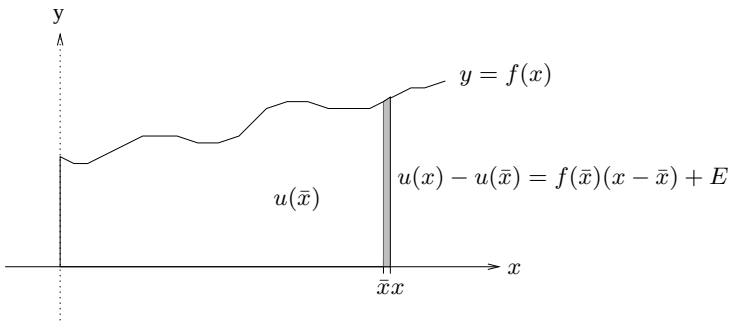


FIGURE 200.2. The derivative of $\int_0^x f(y)\,dy$ at $x = \bar{x}$ is $f(\bar{x})$: $|E| \leq \frac{1}{2}L_f|x - \bar{x}|^2$.

realize that $\int_a^x f(y)\,dy$ is a function of x and thus depends on x. The area under the function f from a to x, of course depends on the upper limit x.

We may express (200.9) in words as follows: Differentiating an integral with respect to the upper limit of integration gives the value of the integrand at the upper limit of integration.

To prove (200.9) we now choose x and $\bar{x}$ in $[a, b]$ with $x \geq \bar{x}$, and use (200.2) to see that

$$u(x) - u(\bar{x}) = \int_a^x f(z)\,dz - \int_a^{\bar{x}} f(y)dy = \int_{\bar{x}}^x f(y)\,dy$$

so that

$$\begin{aligned}|u(x) - u(\bar{x}) - f(\bar{x})(x-\bar{x})| &= \Big| \int_{\bar{x}}^{x} f(y)\,dy - f(\bar{x})(x-\bar{x}) \Big| \\ &= \Big| \int_{\bar{x}}^{x} (f(y) - f(\bar{x}))\,dy \Big| \\ &\le \int_{\bar{x}}^{x} |f(y) - f(\bar{x})|\,dy \\ &\le \int_{\bar{x}}^{x} L_f |y-\bar{x}|\,dy = \frac{1}{2} L_f (x-\bar{x})^2.\end{aligned}$$

This proves that $u(x)$ is uniformly differentiable on $[a,b]$ with derivative $u'(x) = f(x)$ and constant $K_u \le \frac{1}{2}L_f$.

We also note that (200.1) implies

$$\frac{d}{dx}\int_{x}^{a} f(y)\,dy = -f(x). \tag{200.10}$$

which we may express in words as: Differentiating an integral with respect to the lower limit of integration gives minus the value of the integrand at the lower limit of integration.

Example 200.4. We have

$$\frac{d}{dx}\int_{0}^{x} \frac{1}{1+y^2}\,dy = \frac{1}{1+x^2}.$$

Example 200.5. Note that we can combine (199.30) with the Chain Rule:

$$\frac{d}{dx}\int_{0}^{x^3} \frac{1}{1+y^2}\,dy = \frac{1}{1+(x^3)^2}\frac{d}{dx}\big(x^3\big) = \frac{3x^2}{1+x^6}.$$

200.9 Change of Variables or Substitution

We recall that the Chain rule tells us how to differentiate the composition of two functions. The analogous property of the integral is known as the *change of variables*, or *substitution* formula and plays an important role. For example, it can be used to compute many integrals analytically. The idea is that an integral may be easier to compute analytically if we change scales in the independent variable.

Let now $g : [a,b] \to I$, be uniformly differentiable on an interval $[a,b]$, where I is an interval, and let $f : I \to \mathbb{R}$ be Lipschitz continuous. Typically, g is strictly increasing (or decreasing) and maps $[a,b]$ onto I, so that $g :$

$[a,b] \to I$ corresponds to a change of scale, but more general situations are allowed. The *change of variables* formula reads

$$\int_a^x f(g(y))g'(y)\,dy = \int_{g(a)}^{g(x)} f(z)\,dz \quad \text{for } x \in [a,b]. \tag{200.11}$$

This is also called *substitution* since the left hand side $L(x)$ is formally obtained by in the right hand hand side $H(x)$ setting $z = g(y)$ and replacing dz by $g'(y)\,dy$ motivated by the relation

$$\frac{dz}{dy} = g'(y),$$

and noting that as y runs from a to x then z runs from $g(a)$ to $g(x)$.

To verify (200.11), we now prove that $H'(x) = L'(x)$ and use the fact that $H(a) = L(a) = 0$ and the uniqueness of the integral. The Chain rule and (199.30) imply that

$$H'(x) = f(g(x))\,g'(x).$$

Further,

$$L'(x) = f(g(x))\,g'(x),$$

which thus proves the desired equality.

We now give a two examples. We will meet many more examples below.

Example 200.6. To integrate

$$\int_0^2 (1+y^2)^{-2}2y\,dy$$

we make the observation that

$$\frac{d}{dy}(1+y^2) = 2y.$$

Thus, if we set $z = g(y) = 1+y^2$, then applying (200.11) noting that $g(0) = 1$ and $g(2) = 5$ and formally $dz = 2ydy$, we have that

$$\int_0^2 (1+y^2)^{-2}2y\,dy = \int_0^2 (g(y))^{-2}g'(y)\,dy = \int_1^5 z^{-2}\,dz.$$

Now, the right hand integral can easily be evaluated:

$$\int_1^5 z^{-2}\,dz = [-z^{-1}]_{z=1}^{z=5} = -\big(\frac{1}{5} - 1\big),$$

and thus

$$\int_0^2 (1+y^2)^{-2}2y\,dy = \frac{4}{5}.$$

EXAMPLE 200.7. We have setting $y = g(x) = 1 + x^4$ noting that then formally $dy = g'(x)dx = 4x^3dx$ and $g(0) = 1$ and $g(1) = 2$, to get

$$\int_0^1 (1+x^4)^{-1/2}x^3\,dx = \frac{1}{4}\int_0^1 (g(x))^{-1/2}g'(x)\,dx = \frac{1}{4}\int_1^2 y^{-1/2}\,dy$$
$$= \frac{1}{2}[y^{1/2}]_1^2 = \frac{\sqrt{2}-1}{2}.$$

200.10 Integration by Parts

We recall that the Product rule is a basic property of the derivative, showing how to compute the derivative of a product of two functions. The corresponding formula for integration is called *integration by parts*. The formula is

$$\int_a^b u'(x)v(x)\,dx = u(b)v(b) - u(a)v(a) - \int_a^b u(x)v'(x)\,dx. \qquad (200.12)$$

The formula follows by applying the Fundamental Theorem to the function $w(x) = u(x)v(x)$, in the form

$$\int_a^b w'(x)\,dx = u(b)v(b) - u(a)v(a),$$

together with the product formula $w'(x) = u'(x)v(x) + u(x)v'(x)$ and (200.3). Below we often write

$$u(b)v(b) - u(a)v(a) = \Big[u(x)v(x)\Big]_{x=a}^{x=b},$$

and we can thus state the formula for integration by parts as

$$\int_a^b u'(x)v(x)\,dx = \Big[u(x)v(x)\Big]_{x=a}^{x=b} - \int_a^b u(x)v'(x)\,dx. \qquad (200.13)$$

This formula is very useful and we will use many times below.

EXAMPLE 200.8. Computing

$$\int_0^1 4x^3(1+x^2)^{-3}\,dx$$

by guessing at a primitive function for the integrand would be a pretty daunting task. However we can use integration by parts to compute the integral. The trick here is to realize that

$$\frac{d}{dx}(1+x^2)^{-2} = -4x(1+x^2)^{-3}.$$

If we rewrite the integral as

$$\int_0^1 x^2 \times 4x(1+x^2)^{-3}\,dx$$

then we can apply integration by parts with $u(x) = x^2$ and $v'(x) = 4x(1+x^2)^{-3}$, so $u'(x) = 2x$ and $v(x) = -(1+x^2)^{-2}$, to get

$$\begin{aligned}\int_0^1 4x^3(1+x^2)^{-3}\,dx &= \int_0^1 u(x)v'(x)\,dx \\ &= [x^2(-(1+x^2)^{-2})]_{x=0}^{x=1} - \int_0^1 2x(-(1+x^2)^{-2})\,dx \\ &= -\frac{1}{4} - \int_0^1 (-(1+x^2)^{-2})2x\;dx.\end{aligned}$$

To do the remaining integral, we use the substitution $z = 1+x^2$ with $dz = 2x\,dx$ to get

$$\begin{aligned}\int_0^1 4x^3(1+x^2)^{-3}\,dx &= -\frac{1}{4} + \int_1^2 z^{-2}\,dz \\ &= -\frac{1}{4} + \big[-z^{-1}\big]_{z=1}^{z=2} = -\frac{1}{4} - \frac{1}{2} + 1 = \frac{1}{4}.\end{aligned}$$

200.11 The Mean Value Theorem

The *Mean Value theorem* states that if $u(x)$ is a differentiable function on $[a,b]$ then there is a point $\bar{x}$ in (a,b) such that the slope $u'(\bar{x})$ of the tangent of the graph of $u(x)$ at $\bar{x}$ is equal to the slope of the secant line, or chord, connecting the points $(a,u(a))$ and $(b,u(b))$. In other words,

$$\frac{u(b)-u(a)}{b-a} = u'(\bar{x}). \tag{200.14}$$

This is geometrically intuitive, see Fig. 200.3, and expresses that the average velocity over $[a,b]$ is equal to momentary velocity $u'(\bar{x})$ at some intermediate point $\bar{x} \in [a,b]$.

To get from the point $(a,u(a))$ to the point $(b,u(b))$, f has to "bend" around in such a way that the tangent becomes parallel to the secant line at least at one point.

Assuming that $u'(x)$ is Lipschitz continuous on $[a,b]$, we shall now prove that there is a real number $\bar{x} \in [a,b]$ such that

$$u(b) - u(a) = (b-a)u'(\bar{x})$$

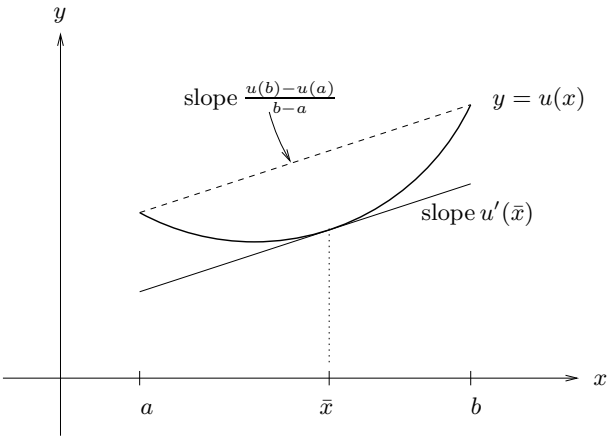


FIGURE 200.3. Illustration of the Mean Value theorem.

which is equivalent to (200.14). The proof is based on the formula

$$u(b) = u(a) + \int_a^b u'(x)\,dx \tag{200.15}$$

which holds if $u(x)$ is uniformly differentiable on $[a, b]$. If for all $x \in [a, b]$, we would have

$$\frac{u(b) - u(a)}{b - a} > u'(x),$$

then we would have (explain why)

$$u(b) - u(a) = \int_b^a \frac{u(b) - u(a)}{b - a}\,dx > \int_b^a u'(x)\,dx = u(b) - u(a),$$

which is a contradiction. Arguing in the same way, we conclude it is also impossible that

$$\frac{u(b) - u(a)}{b - a} < u'(x)$$

for all $x \in [a, b]$. So there must be numbers c and d in $[a, b]$ such that

$$u'(c) \leq \frac{u(b) - u(a)}{b - a} \leq u'(d).$$

Since $u'(x)$ is Lipschitz continuous for $x \in [a, b]$, it follows by the Intermediate Value theorem that there is a number $\bar{x} \in [a, b]$ such that

$$u'(\bar{x}) = \frac{u(b) - u(a)}{b - a}.$$

We have now proved:

Theorem 200.1 (Mean Value theorem) *If $u(x)$ is uniformly differentiable on $[a,b]$ with Lipschitz continuous derivative $u'(x)$, then there is a (at least one) $\bar{x} \in [a,b]$, such that*

$$u(b) - u(a) = (b-a)u'(\bar{x}). \tag{200.16}$$

The Mean Value Theorem is often written in terms of an integral by setting $f(x) = u'(x)$ in (200.16), which gives

Theorem 200.2 (Mean Value theorem for integrals) *If $f(x)$ is Lipschitz continuous on $[a,b]$, then there is some $\bar{x} \in [a,b]$ such that*

$$\int_a^b f(x)\,dx = (b-a)f(\bar{x}). \tag{200.17}$$

The Mean Value theorem turns out to be very useful in several ways. To illustrate, we discuss two results that can be proved easily using the Mean Value Theorem.

200.12 Monotone Functions and the Sign of the Derivative

The first result says that the sign of the derivative of a function indicates whether the function is increasing or decreasing in value as the input increases. More precisely, the Mean Value theorem implies that if $f'(x) \geq 0$ for all $x \in [a,b]$ then $f(b) \geq f(a)$. Moreover if $x_1 \leq x_2$ are in $[a,b]$, then $f(x_1) \leq f(x_2)$. A function with this property is said to be *non-decreasing* on $[a,b]$. If in fact $f'(x) > 0$ for all $x \in (a,b)$, then $f(x_1) < f(x_2)$ for $x_1 < x_2$ in $[a,b]$ (with strict inequalities), and we say that $f(x)$ is *(strictly) increasing* on the interval $[a,b]$. Corresponding statements hold if $f'(x) \leq 0$ and $f'(x) < 0$, with non-decreasing and (strictly) increasing replaced with *non-increasing* and *(strictly) decreasing*, respectively. Functions that are either (strictly) increasing or (strictly) decreasing on an interval $[a,b]$ are said to be *(strictly) monotone* on $[a,b]$.

200.13 A Function with Zero Derivative Is Constant

As a particular consequence of the preceding section, we conclude that if $f'(x) = 0$ for all $x \in [a,b]$, so that $f(x)$ is both non-increasing and non-decreasing on $[a,b]$, then $f(x)$ is constant on $[a,b]$. Thus, a function with derivative vanishing everywhere is a constant function.

200.14 A Bounded Derivative Implies Lipschitz Continuity

As a second consequence of the Mean Value Theorem, we give an alternate, and shorter, proof that a function with a Lipschitz continuous derivative is Lipschitz continuous. Assume that $u : [a, b] \to \mathbb{R}$ has a Lipschitz continuous derivative $u'(x)$ on $[a, b]$ satisfying $|u'(x)| \leq M$ for $x \in [a, b]$. By the Mean Value theorem, we have

$$|u(x) - u(\bar{x})| \leq M|x - \bar{x}| \quad \text{for } x, \bar{x} \in [a, b].$$

We conclude that $u(x)$ is Lipschitz continuous on $[a, b]$ with Lipschitz constant $M = \max_{x \in [a,b]} |u'(x)|$.

200.15 Taylor's Theorem

In earlier chapters, we analyzed a linear approximation to a function u,

$$u(x) \approx u(\bar{x}) + u'(\bar{x})(x - \bar{x}), \tag{200.18}$$

as well as a quadratic approximation

$$u(x) \approx u(\bar{x}) + u'(\bar{x})(x - \bar{x}) + \frac{u''(\bar{x})}{2}(x - \bar{x})^2. \tag{200.19}$$

These approximations are very useful tools for dealing with nonlinear functions. *Taylor's theorem*, invented by Brook Taylor (1685-1731), see Fig. 200.4, generalizes these approximations to any degree.Taylor sided up with Newton in a long scientific fight with associates of Leibniz about "who's best in Calculus?".

Theorem 200.3 (Taylor's theorem) *If $u(x)$ is $n+1$ times differentiable on the interval I with $u^{(n+1)}$ Lipschitz continuous, then for $x, \bar{x} \in I$, we have*

$$u(x) = u(\bar{x}) + u'(\bar{x})(x - \bar{x}) + \cdots + \frac{u^{(n)}(\bar{x})}{n!}(x - \bar{x})^n + \int_{\bar{x}}^{x} \frac{(x - y)^n}{n!}\, u^{(n+1)}(y)\, dy. \tag{200.20}$$

The polynomial

$$P_n(x) = u(\bar{x}) + u'(\bar{x})(x - \bar{x}) + \cdots + \frac{u^{(n)}(\bar{x})}{n!}(x - \bar{x})^n$$

is called the *Taylor polynomial*, or *Taylor expansion*, of $u(x)$ at $\bar{x}$ of degree n,

FIGURE 200.4. Brook Taylor, inventor of the Taylor expansion: "I am the best".

The term

$$R_n(x) = \int_{\bar{x}}^{x} \frac{(x-y)^n}{n!}\, u^{(n+1)}(y)\, dy$$

is called the *remainder term* of order n. We have for $x \in I$,

$$u(x) = P_n(x) + R_n(x).$$

It follows directly that

$$\left(\frac{d^k}{dx^k}\right) P_n(\bar{x}) = \left(\frac{d^k}{dx^k}\right) u(\bar{x}) \quad \text{for } k = 0, 1 \cdots, n.$$

Thus Taylor's theorem gives a polynomial approximation $P_n(x)$ of degree n of a given function $u(x)$, such that the derivatives of order $\leq n$ of $P_n(x)$ and $u(x)$ agree at $x = \bar{x}$.

EXAMPLE 200.9. The Taylor polynomial of order 2 at $x = 0$ for $u(x) = \sqrt{1+x}$ is given by

$$P_2(x) = 1 + \frac{1}{2} - \frac{1}{8}x^2,$$

since $u(0) = 1$, $u'(0) = \frac{1}{2}$, and $u''(0) = -\frac{1}{4}$.

The proof of Taylor's theorem is a wonderful application of integration by parts, discovered by Taylor. We start by noting that Taylor's theorem with $n = 0$ is just the Fundamental Theorem

$$u(x) = u(\bar{x}) + \int_{\bar{x}}^{x} u'(y)\, dy,$$

Using that $\frac{d}{dy}(y-x)=1$, we get integrating by parts

$$\begin{aligned}
u(x) &= u(\bar{x}) + \int_{\bar{x}}^{x} u'(y)\,dy \\
&= u(\bar{x}) + \int_{\bar{x}}^{x} \frac{d}{dy}(y-x)u'(y)\,dy \\
&= u(\bar{x}) + [(y-x)u'(y)]_{y=\bar{x}}^{y=x} - \int_{\bar{x}}^{x} (y-x)u''(y)\,dy \\
&= u(\bar{x}) + (x-\bar{x})u'(\bar{x}) + \int_{\bar{x}}^{x} (x-y)u''(y)\,dy,
\end{aligned}$$

which is Taylor's theorem with $n=1$. Continuing in this manner, integrating by parts, using the notation $k_n(y)=(y-x)^n/n!$, and noting that for $n \geq 1$

$$\frac{d}{dy}k_n(y) = k_{n-1}(y),$$

we get

$$\begin{aligned}
\int_{\bar{x}}^{x} \frac{(x-y)^{n-1}}{(n-1)!}u^{(n)}(y)\,dy &= (-1)^{n-1}\int_{\bar{x}}^{x} k_{n-1}(y)u^{(n)}(y)\,dy \\
&= (-1)^{n-1}\int_{\bar{x}}^{x} \frac{d}{dy}k_n(y)u^{(n)}(y)\,dy \\
&= [(-1)^{n-1}k_n(y)u^{(n)}(y)]_{y=\bar{x}}^{y=x} - (-1)^{n-1}\int_{\bar{x}}^{x} k_n(y)u^{(n+1)}(y)\,dy \\
&= \frac{u^{(n)}(\bar{x})}{n!}(x-\bar{x})^n + \int_{\bar{x}}^{x} \frac{(x-y)^n}{n!}u^{(n+1)}(y)\,dy.
\end{aligned}$$

This proves Taylor's theorem.

EXAMPLE 200.10. We compute a fourth order polynomial approximation to $f(x)=\frac{1}{1-x}$ near $x=0$. We have

$$\begin{aligned}
f(x) = \frac{1}{1-x} &\implies f(0)=1, \\
f'(x) = \frac{1}{(1-x)^2} &\implies f'(0)=1, \\
f''(x) = \frac{2}{(1-x)^3} &\implies f''(0)=2, \\
f'''(x) = \frac{6}{(1-x)^4} &\implies f'''(0)=6, \\
f''''(x) = \frac{24}{(1-x)^5} &\implies f''''(0)=24,
\end{aligned}$$

and therefore

$$\begin{aligned} P_4(x) &= 1 + 1(x-0)^1 + \frac{2}{2}(x-0)^2 + \frac{6}{6}(x-0)^3 + \frac{24}{24}(x-0)^4 \\ &= 1 + x + x^2 + x^3 + x^4. \end{aligned}$$

We plot the function and the polynomial in Fig. 200.5. Characteristically,

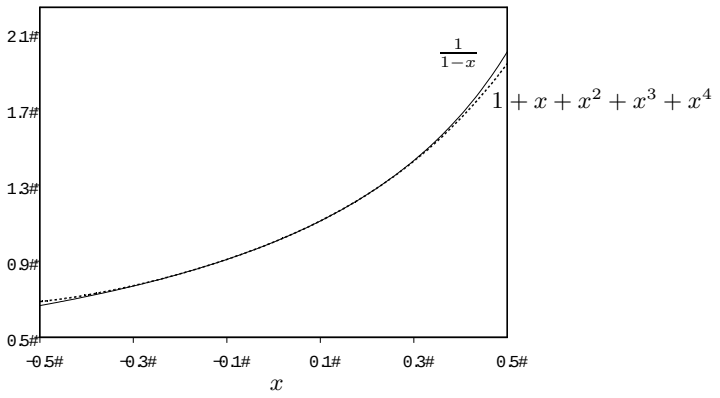


FIGURE 200.5. Plots of $f(x) = 1/(1-x)$ and its Taylor polynomial $1 + x + x^2 + x^3 + x^4$.

the Taylor polynomial is a very accurate approximation near the $\bar{x}$ but the error becomes larger as x moves further away from $\bar{x}$.

200.16 October 29, 1675

On October 29, 1675, Leibniz got a bright idea while sitting at his desk in Paris. He wrote "Utile erit scribit $\int$ pro omnia", which translates to "It is useful to write $\int$ instead of omnia". This is the moment when the modern notation of calculus was created. Earlier than this date, Leibniz had been working with a notation based on a, l and "omnia" which represented in modern notation dx, dy and $\int$ respectively. This notation resulted in formulas like

$$\text{omn.}l = y, \quad \text{omn.}yl = \frac{y^2}{2}, \quad \text{omn.}xl = x\text{omn.}l - \text{omn.omn.}la,$$

where "omn.", short for omnia, indicated a discrete sum and l and a denoted increments of finite size (often $a = 1$). In the new notation, these formulas became

$$\int dy = y, \quad \int y\,dy = \frac{y^2}{2}, \quad \int x\,dy = xy - \int y\,dx. \tag{200.21}$$

This opened up the possibility of dx and dy being arbitrarily small and the sum being replaced by the "integral".

200.17 The Hodometer

The Romans constructed many roads to keep the Empire together and the need of measuring distances between cities and traveled distance on the road, became very evident. For this purpose the *Hodometer* was constructed by Vitruvius, see Fig. 200.6. For each turn of the wagon wheel, the vertical gear is shifted one step, and for each turn of the vertical gear the horizontal gear is shifted one step. The horizontal gear has a set of holes with one stone in each, and for each shift one stone drops down to a box under the wagon; at the end of the day one computes the number of stones in the box, and the device is so calibrated that this number is equal to the number of traveled miles. Evidently, one may view the hodometer as a kind of simple analog integrator.

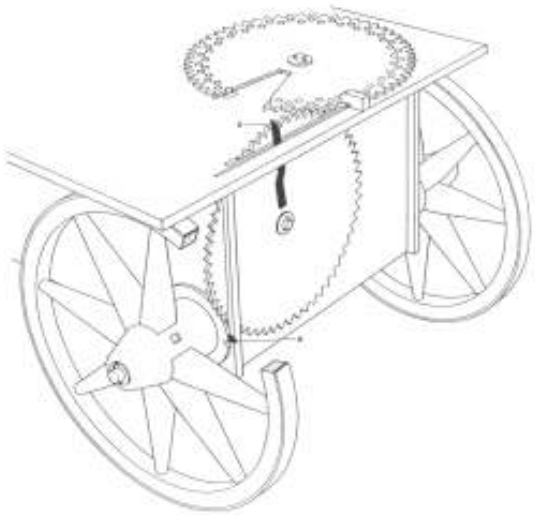


FIGURE 200.6. The principle of the Hodometer

Chapter 200 Problems

200.1. Compute the following integrals: a) $\int_0^1 (ax + bx^2)dx$, b) $\int_{-1}^1 |x|dx$, c) $\int_{-1}^1 |x-1|dx$, d) $\int_{-1}^1 |x+a|dx$, e) $\int_{-1}^1 (x-a)^{10}dx$.

200.2. Compute the following integrals by integration by parts. Verify that you get the same result by directly finding the primitive function. a) $\int_0^1 x^2 dx = \int_0^1 x\cdot x dx$, b) $\int_0^1 x^3 dx = \int_0^1 x\cdot x^2 dx$, c) $\int_0^1 x^3 dx = \int_0^1 x^{3/2}\cdot x^{3/2} dx$, d) $\int_0^1 (x^2-1)dx = \int_0^1 (x+1)\cdot(x-1)dx$.

200.3. For computing the integral $\int_0^1 x(x-1)^{1000}dx$, what would you rather do; find the primitive function directly or integrate by parts?

200.4. Compute the following integrals: a) $\int_{-1}^2 (2x-1)^7 dx$, b) $\int_0^1 f'(7x)dx$, c) $\int_{-10}^{-7} f'(17x+5)dx$.

200.5. Compute the integral $\int_0^1 x(x^2-1)^{10}dx$ in two ways, first by integration by parts, then by a clever substitution using the chain rule.

200.6. Find Taylor polynomials at $\overline{x}$ of the following functions: a) $f(x) = x$, $\overline{x} = 0$, b) $f(x) = x + x^2 + x^3$, $\overline{x} = 1$, c) $f(x) = \sqrt{\sqrt{x+1}+1}$, $\overline{x} = 0$.

200.7. Find a Taylor expansion of the function $f(x) = x^r - 1$ around a suitable choice of $\overline{x}$, and use the result to compute the limit $\lim_{x\to 1} \frac{x^r-1}{x-1}$. Compare this to using l'Hopital's rule to compute the limit. Can you see the connection between the two methods?

200.8. Motivate the basic properties of linearity and subinterval additivity of the integral using the area interpretation of the integral.

200.9. Prove the basic properties of linearity and subinterval additivity of the integral using that the integral is a limit of discrete sums together with basic properties of discrete sums.

200.10. Make sense out of Leibniz formulas (200.21). Prove, as did Leibniz, the second from a geometrical argument based on computing the area of a right-angled triangle by summing thin slices of variable height y and thickness dy, and the third from computing similarly the area of a rectangle as the sum of the two parts below and above a curve joining two opposite corners of the rectangle.

200.11. Prove the following variant of Taylor's theorem: If $u(x)$ is $n+1$ times differentiable on the interval I, with $u^{(n+1)}(x)$ Lipschitz continuous, then for $\bar{x} \in I$, we have

$$u(x) = u(\bar{x}) + u'(\bar{x})(x-\bar{x}) + \cdots + \frac{u^{(n)}(\bar{x})}{n!}(x-\bar{x})^n + \frac{u^{(n+1)}(\hat{x})}{(n+1)!}(x-\bar{x})^{n+1}$$

where $\hat{x} \in [\bar{x}, x]$. Hint: Use the Mean Value theorem for integrals.

200.12. Prove that if $x = f(y)$ with inverse function $y = f^{-1}(x)$, and $f(0) = 0$, then

$$\int_0^{\bar{y}} f(y)\,dy = \bar{y}\bar{x} - \int_0^{\bar{x}} f^{-1}(x)\,dx.$$

Compare with (200.21) Hint: use integration by parts.

200.13. Show that $x \mapsto F(x) = \int_0^x f(x)dx$ is Lipschitz continuous on $[0, a]$ with Lipschitz constant L_F if $|f(x)| \leq L_F$ for $x \in [0, a]$.

200.14. Why can we think of the primitive function as being "nicer" than the function itself?

200.15. Under what conditions is the following generalized integration by parts formula valid

$$\int_I \frac{d^n f}{dx^n}\varphi dx = (-1)^n \int_I f\frac{d^n \varphi}{dx^n}dx,\ \ n = 0, 1, 2, \ldots?$$

200.16. Show the following inequality:

$$|\int_I u(x)v(x)\ dx| \leq \sqrt{\int_I u^2(x)\ dx}\ \sqrt{\int_I v^2(x)\ dx},$$

which is referred to as *Cauchy's inequality*. Hint: Let $\overline{u} = u/\sqrt{\int_I u^2(x)\ dx}$, $\overline{v} = v/\sqrt{\int_I v^2(x)\ dx}$, and show that $|\int_I \overline{u}(x)\overline{v}(x)\ dx| \leq 1$ by considering the expression $\int_I(\overline{u}(x) - \int_I \overline{u}(y)\overline{v}(y)\ dy\ \overline{v}(x))\ dx$. Would it be helpful to use the notation $(u, v) = \int_I u(x)v(x)\ dx$, and $\|u\| = \sqrt{\int_I u^2(x)\ dx}$?

200.17. Show that if v is Lipschitz continuous on the bounded interval I and $v = 0$ at one of the endpoints of the interval, then

$$\|v\|_{L^2(I)} \leq C_I\|v'\|_{L^2(I)},$$

for some constant C_I, where the so-called $L^2(I)$ norm of a function v is defined as $\|v\|_{L^2(I)} = \sqrt{\int_I v^2(x)dx}$. What is the value of the constant? Hint: Express v in terms of v' and use the result from the previous problem.

200.18. Check that the inequality from the previous problem holds for the following functions on $I = [0, 1]$: a) $v(x) = x(1 - x)$, b) $v(x) = x^2(1 - x)$, c) $v(x) = x(1 - x)^2$.

200.19. Prove quadratic convergence of Newton's method (198.5) for computing a root $\bar{x}$ of the equation $f(x) = 0$ using Taylor's theorem. Hint: Use the fact that $x_{i+1} - \bar{x} = x_i - \bar{x} + \frac{f(x_i)-f(\bar{x})}{f'(x_i)}$ and Taylor's theorem to see that $f(x_i) - f(\bar{x}) = f'(x_i)(x_i - \bar{x}) + \frac{1}{2}f''(\tilde{x}_i)(x_i - \bar{x})^2$ for some $\tilde{x}_i \approx x_i$.

200.20. Prove (200.3) from (200.4).

201
The Logarithm $\log(x)$

> Nevertheless technicalities and detours should be avoided, and the presentation of mathematics should be just as free from emphasis on routine as from forbidding dogmatism, which refuses to disclose motive or goal and which is an unfair obstacle to honest effort. (R. Courant)

201.1 The Definition of $\log(x)$

We return to the question of the existence of a primitive function of $f(x) = 1/x$ for $x > 0$ posed above. Since the function $f(x) = 1/x$ is Lipschitz continuous on any given interval $[a, b]$ with $0 < a < b$, we know by the Fundamental Theorem that there is a unique function $u(x)$ which satisfies $u'(x) = 1/x$ for $a \leq x \leq b$ and takes on a specific value at some point in $[a, b]$, for example $u(1) = 0$. Since $a > 0$ may be chosen as small as we please and b as large as we please, we may consider the function $u(x)$ to be defined for $x > 0$. We now define the *natural logarithm* $\log(x)$ (or $\ln(x)$) for $x > 0$ as the primitive function $u(x)$ of $1/x$ vanishing for $x = 1$, i.e., $\log(x)$ satisfies

$$\frac{d}{dx}(\log(x)) = \frac{1}{x} \quad \text{for } x > 0, \quad \log(1) = 0. \tag{201.1}$$

Using the definition of the integral, we may express $\log(x)$ as an integral:

$$\log(x) = \int_1^x \frac{1}{y}\, dy \quad \text{for } x > 0. \tag{201.2}$$

In the next chapter we shall use this formula to compute approximations of $\log(x)$ for a given $x > 0$ by computing approximations of the corresponding integral. We plot $\log(x)$ in Fig. 201.1.

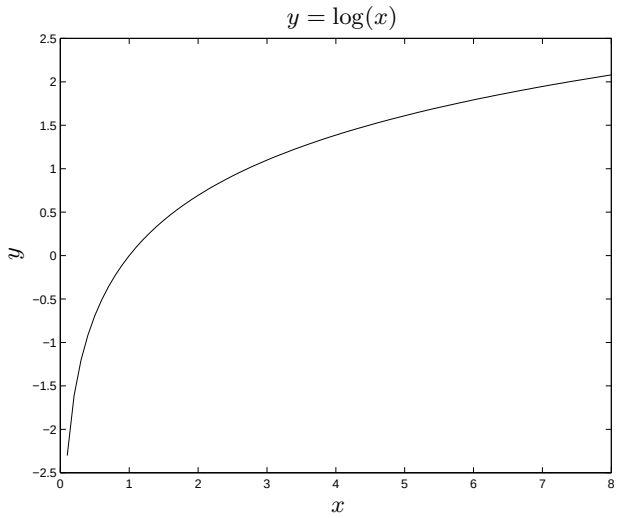


FIGURE 201.1. Plot of $\log(x)$.

201.2 The Importance of the Logarithm

The logarithm function $\log(x)$ is a basic function in science, simply because it solves a basic differential equation, and thus occurs in many applications. More concretely, the logarithm has some special properties that made previous generations of scientists and engineers use the logarithm intensely, including memorizing long tables of its values. The reason is that one can compute products of real numbers by adding logarithms of real numbers, and thus the operation of multiplication can be replaced by the simpler operation of addition. The *slide rule* is an analog computing device built on this principle, that used to be sign of the engineer visible in the waist-pocket, recall Fig. **??**. Today the modern computer has replaced the slide rule and does not use logarithms to multiply real numbers. However, the first computer, the mechanical Difference Machine by Babbage from the 1830s, see Fig. **??**, was used for computing accurate tables of values of the logarithm. The logarithm was discovered by John Napier and presented in *Mirifici logarithmorum canonis descriptio* in 1614. A illuminating citation from the foreword is given in Fig. 201.2.

FIGURE 201.2. Napier, Inventor of the Logarithm: "Seeing there is nothing (right well-beloved Students of the Mathematics) that is so troublesome to mathematical practice, nor that doth more molest and hinder calculators, than the multiplications, divisions, square and cubical extractions of great numbers, which besides the tedious expense of time are for the most part subject to many slippery errors, I began therefore to consider in my mind by what certain and ready art I might remove those hindrances. And having thought upon many things to this purpose, I found at length some excellent brief rules to be treated of (perhaps) hereafter. But amongst all, none more profitable than this which together with the hard and tedious multiplications, divisions, and extractions of roots, doth also cast away from the work itself even the very numbers themselves that are to be multiplied, divided and resolved into roots, and putteth other numbers in their place which perform as much as they can do, only by addition and subtraction, division by two or division by three".

201.3 Important Properties of $\log(x)$

We now derive the basic properties of the logarithm function $\log(x)$ using (201.1) or (201.2). We first note that $u(x) = \log(x)$ is *strictly increasing* for $x > 0$, because $u'(x) = 1/x$ is positive for $x > 0$. This can be seen in Fig. 201.1. Recalling (201.1), we conclude that for $a, b > 0$,

$$\int_a^b \frac{dy}{y} = \log(b) - \log(a).$$

Next we note that the Chain rule implies that for any constant $a > 0$

$$\frac{d}{dx}(\log(ax) - \log(x)) = \frac{1}{ax} \cdot a - \frac{1}{x} = 0,$$

and consequently $\log(ax) - \log(x)$ is constant for $x > 0$. Since $\log(1) = 0$, we see by setting $x = 1$ that the constant value is equal to $\log(a)$, and so

$$\log(ax) - \log(x) = \log(a) \quad \text{for } x > 0.$$

Choosing $x = b > 0$, we thus obtain the following fundamental relation satisfied by the logarithm $\log(x)$:

$$\log(ab) = \log(a) + \log(b) \quad \text{for } a, b > 0 \tag{201.3}$$

We can thus compute the logarithm of the product of two numbers by adding the logarithms of the two numbers. We already indicated that this is the principle of the slide rule or using a table of logarithms for multiplying two numbers. More precisely (as first proposed by Napier), to multiply two numbers a and b we first find their logarithms $\log(a)$ and $\log(b)$ from the table, then add them to get $\log(ab)$ using the formula (201.3), and finally we find from the table which real number has the logarithm equal to $\log(ab)$, which is equal to the desired product ab. Clever, right?

The formula (201.3) implies many things. For example, choosing $b = 1/a$, we get

$$\log(a^{-1}) = -\log(a) \quad \text{for } a > 0. \tag{201.4}$$

Choosing $b = a^{n-1}$ with $n = 1, 2, 3, ..$, we get

$$\log(a^n) = \log(a) + \log(a^{n-1}),$$

so that by repeating this argument

$$\log(a^n) = n \log(a) \quad \text{for } n = 1, 2, 3, ... \tag{201.5}$$

By (201.4) the last equality holds also for $n = -1, -2, ...$

More generally, we have for any $r \in \mathbb{R}$ and $a > 0$,

$$\log(a^r) = r \log(a). \tag{201.6}$$

We prove this using the change of variables $x = y^r$ with $dx = ry^{r-1}dy$:

$$\log(a^r) = \int_1^{a^r} \frac{1}{x}\, dx = \int_1^a \frac{ry^{r-1}}{y^r}\, dy = r \int_1^a \frac{1}{y}\, dy = r \log(a).$$

Finally, we note that $1/x$ also has a primitive function for $x < 0$ and for $a, b > 0$, setting $y = -x$,

$$\begin{aligned} \int_{-a}^{-b} \frac{dy}{y} = \int_a^b \frac{-dx}{-x} = \int_a^b \frac{dx}{x} &= \log(b) - \log(a) \\ &= \log(-(-b)) - \log(-(-a)). \end{aligned}$$

Accordingly, we may write for any $a \neq 0$ and $b \neq 0$ that have the same sign,

$$\int_a^b \frac{dx}{x} = \log(|b|) - \log(|a|). \tag{201.7}$$

It is important to understand that (201.7) does *not* hold if a and b have opposite signs.

Chapter 201 Problems

201.1. Prove that $\log(4) > 1$ and $\log(2) \geq 1/2$.

201.2. Prove that

$$\log(x) \to \infty \quad \text{as } x \to \infty,$$
$$\log(x) \to -\infty \quad \text{as } x \to 0^+.$$

Hint: Using that $\log(2) \geq 1/2$ it follows from (201.5) that $\log(2^n)$ tends to infinity as n tends to infinity.

201.3. Give an alternative proof of (201.3) using that

$$\log(ab) = \int_1^{ab} \frac{1}{y}dy = \int_1^a \frac{1}{y}dy + \int_a^{ab} \frac{1}{y}dy = \log(a) + \int_a^{ab} \frac{1}{y}dy,$$

and changing the variable y in the last integral to $z = ay$.

201.4. Prove that $\log(1+x) \leq x$ for $x > 0$, and that $\log(1+x) < x$ for $x \neq 0$ and $x > -1$. Hint: Differentiate.

201.5. Show using the Mean Value theorem, that $\log(1+x) \leq x$ for $x > -1$. Can prove this directly from the definition of the logarithm by sketching the area under the graph?

201.6. Prove that $\log(a) - \log(b) = \log(\frac{a}{b})$ for $a, b > 0$.

201.7. Write down the Taylor polynomial of order n for $\log(x)$ at $x = 1$.

201.8. Find a primitive function of $\frac{1}{x^2-1}$. Hint: use that $\frac{1}{x^2-1} = \frac{1}{(x-1)(x+1)} = \frac{1}{2}(\frac{1}{x-1} - \frac{1}{x+1})$.

201.9. Prove that $\log(x^r) = r\log(x)$ for $r = \frac{p}{q}$ rational by using (201.5) cleverly.

201.10. Solve the initial value problem $u'(x) = 1/x^a$ for $x > 0$, $u(1) = 0$, for values of the exponent a close to 1. Plot the solutions. Study for which values of a the solution $u(x)$ tends to infinity when x tends to infinity.

201.11. Solve the following equations: (a) $\log(x^2) + \log(3) = \log(\sqrt{x}) + \log(5)$, (b) $\log(7x) - 2\log(x) = \log(3)$, (c) $\log(x^3) - \log(x) = \log(7) - \log(x^2)$.

201.12. Compute the derivatives of the following functions: a) $f(x) = \log(x^3 + 6x)$, b) $f(x) = \log(\log(x))$, c) $f(x) = \log(x + x^2)$, d) $f(x) = \log(1/x)$, e) $f(x) = x\log(x) - x$.

202
Numerical Quadrature

"And I know it *seems* easy", said Piglet to himself, "but it isn't *everyone* who could do it". (House at the Pooh Corner, Milne)

Errare humanum est.

202.1 Computing Integrals

In some cases, we can compute a primitive function (or antiderivative or integral) of a given function analytically, that is we can give give a formula for the primitive function in terms of known functions. For example we can give a formula for a primitive function of a polynomial as another polynomial. We will return in Chapter *Techniques of integration* to the question of finding analytical formulas for primitive functions of certain classes of functions. The Fundamental Theorem states that any given Lipschitz continuous function has a primitive function, but does not give any analytical formula for the primitive function. The logarithm,

$$\log(x) = \int_1^x \frac{dy}{y}, \quad \text{where } x > 0,$$

is the first example of this case we have encountered. We know that the logarithm function $\log(x)$ exists for $x > 0$, and we have derived some of its properties indirectly through its defining differential equation, but the question remains how to determine the value of $\log(x)$ for a given $x > 0$. Once we have solved this problem, we may add $\log(x)$ to a list of "elementary" functions that we can play with. Below we will add to this list the

exponential function, the trigonometric functions, and other more exotic functions.

This situation is completely analogous to solving algebraic equations for numbers. Some equations have rational roots and in that case, we feel that we can solve the equation "exactly" by analytic (symbolic) computation. We have a good understanding of rational numbers, even when they have infinite decimal expansions, and we can determine their values, or the pattern in the decimal expansion, with a finite number of arithmetic operations. But most equations have irrational roots with infinite, non-repeating decimal expansions that we can only approximate to a desired level of accuracy in practice. Likewise in a situation in which a function is known only as a primitive function of a given function, the best we can do is to seek to compute its values approximately to a desired level of accuracy. One way to compute values of such a function is through the definition of the integral as a Riemann sum. This is known as *numerical quadrature* or *numerical integration*, and we now explore this possibility.

Suppose thus that we want to compute the integral

$$\int_a^b f(x)\,dx, \tag{202.1}$$

where $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L_f. If we can give a formula for the primitive function $F(x)$ of $f(x)$, then the integral is simply $F(b) - F(a)$. If we cannot give a formula for $F(x)$, then we turn to the Fundamental Theorem and compute an approximate the value of the integral using a Riemann sum approximation

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{N} f(x_{i-1}^n)h_n, \tag{202.2}$$

where $x_i^n = a + ih_n$, $h_n = 2^{-n}(b - a)$, and $N = 2^n$ describes a uniform partition of $[a, b]$, with the *quadrature error*

$$Q_n = \Big| \int_a^b f(x)\,dx - \sum_{i=1}^{N} f(x_{i-1}^n)h_n \Big| \leq \frac{b-a}{2} L_f h_n, \tag{202.3}$$

which tends to zero as we increase the number of steps and $h_n \to 0$. Put another way, if we desire the value of the integral to within a *tolerance* $TOL > 0$ and we know the Lipschitz constant L_f, then we will have $Q_n \leq TOL$ if the mesh size h_n satisfies the *stopping criterion*

$$h_n \leq \frac{2\,TOL}{(b-a)L_f}. \tag{202.4}$$

We refer to the Riemann sum approximation (202.2), compare also with Fig. 202.1, as the *rectangle rule*, which is the simplest method for approximating an integral among many possible methods. The search for more

sophisticated methods for approximating an integral is driven by consideration of the *computational cost* associated to computing the approximation. The cost is typically measured in terms of time because there is a limit on the time we are willing to wait for a solution. In the rectangle rule, the computer spends most of the time evaluating the function f and since each step requires one evaluation of f, the cost is determined ultimately by the number of steps. Considering the cost leads to the optimization problem of trying to compute an approximation of a given accuracy at a relatively low cost.

To reduce the cost, we may construct more sophisticated methods for approximating integrals. But even if we restrict ourselves to the rectangle rule, we can introduce variations that can lower the computational cost of computing an approximation. There are two quantities that we can vary: the point at which we evaluate the function on each interval and the size of the intervals. To understand how these changes could help, consider the illustration of the rectangle rule in Fig. 202.1. Here f varies quite a bit

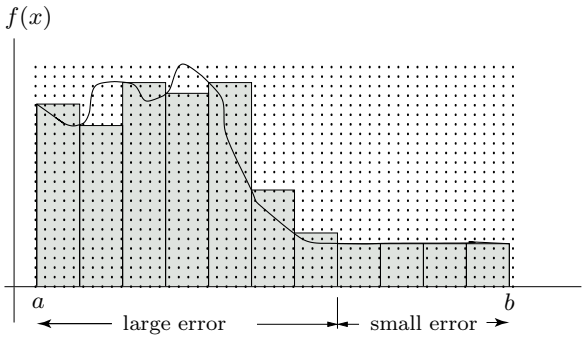


FIGURE 202.1. An illustration of the rectangle rule.

on part of $[a, b]$ and is fairly constant on another part. Consider the approximation to the area under f on the first subinterval on the left. By evaluating f at the left-hand point on the subinterval, we clearly overestimate the area to the maximum degree possible. Choosing to evaluate f at some point inside the subinterval would likely give a better approximation. The same is true of the second subinterval, where choosing the left-hand point clearly leads to an underestimate of the area. On the other hand, consider the approximations to the area in the last four subintervals on the right. Here f is nearly constant and the approximation is very accurate. In fact, we could approximate the area underneath f on this part of $[a, b]$ using one rectangle rather than four. In other words, we would get just as accurate an approximation by using one large subinterval instead of four subintervals. This would cost four times less.

So we generalize the rectangle rule to allow non-uniform partitions and different points at which to evaluate f. We choose a partition $a = x_0 < x_1 < x_2 \cdots < x_N = b$ of $[a, b]$ into N subintervals $I_j = [x_{j-1}, x_j]$ of lengths $h_j = x_j - x_{j-1}$ for $j = 1, ..., N$. Note that N can be any integer and the subintervals may vary in size. By the Mean Value theorem for integrals there is $\bar{x}_j \in I_j$ such that

$$\int_{x_{j-1}}^{x_j} f(x)\,dx = f(\bar{x}_j)h_j, \tag{202.5}$$

and thus we have

$$\int_a^b f(x)\,dx = \sum_{i=1}^{N} \int_{x_{j-1}}^{x_j} f(x)\,dx = \sum_{j=1}^{N} f(\bar{x}_j)h_j.$$

Since the $\bar{x}_j$ are not known in general, we replace $\bar{x}_j$ by a given point $\hat{x}_j \in I_j$. For example, in the original method we use the left end-point $\hat{x}_j = x_{j-1}$, but we could choose the right end-point $\hat{x}_j = x_j$ or the mid-point $\hat{x}_j = \frac{1}{2}(x_{j-1} + x_j)$. We then get the approximation

$$\int_a^b f(x)\,dx \approx \sum_{j=1}^{N} f(\hat{x}_j)h_j. \tag{202.6}$$

We call

$$\sum_{j=1}^{N} f(\hat{x}_j)h_j \tag{202.7}$$

a *quadrature* formula for computing the integral $\int_a^b f(x)\,dx$. We recall that we refer to (202.7) as a *Riemann sum*. The quadrature formula is characterized by the *quadrature points* $\hat{x}_j$ and the *weights* h_j. Note that if $f(x) = 1$ for all x then the quadrature formula is exact and we conclude that $\sum_{j=1}^{N} h_j = b - a$.

We now estimate the *quadrature error*

$$Q_h = \Big| \int_a^b f(x)\,dx - \sum_{j=1}^{N} f(\hat{x}_j)h_j \Big|,$$

where the subscript h refers to the sequence of step sizes h_j. Recalling (202.5), we can do this by estimating the error over each subinterval and then summing. We have

$$\Big| \int_{x_{j-1}}^{x_j} f(x)\,dx - f(\hat{x}_j)h_j \Big| = |f(\bar{x}_j)h_j - f(\hat{x}_j)h_j| = h_j|f(\bar{x}_j) - f(\hat{x}_j)|.$$

We assume that $f'(x)$ is Lipschitz continuous on $[a, b]$. The Mean Value theorem implies that for $x \in [x_{j-1}, x_j]$,

$$f(x) = f(\hat{x}_j) + f'(y)(x - \hat{x}_j),$$

for some $y \in [x_{j-1}, x_j]$. Integrating over $[x_{j-1}, x_j]$ we obtain

$$\Big| \int_{x_{j-1}}^{x_j} f(x)\,dx - f(\hat{x}_j)h_j \Big| \leq \max_{y \in I_j} |f'(y)| \int_{x_{j-1}}^{x_j} |x - \hat{x}_j|\,dx.$$

To simplify the sum on the right, we use the fact that

$$\int_{x_{j-1}}^{x_j} |x - \hat{x}_j|\,dx$$

is maximized if $\hat{x}_j$ is the left (or right) endpoint, in which case

$$\int_{x_{j-1}}^{x_j} (x - x_{j-1})\,dx = \frac{1}{2}h_j^2.$$

We find that

$$\Big| \int_{x_{j-1}}^{x_j} f(x)\,dx - f(\hat{x}_j)h_j \Big| \leq \frac{1}{2} \max_{y \in I_j} |f'(y)| h_j^2.$$

Summing, we conclude

$$Q_h = \left| \int_a^b f(x)\,dx - \sum_{j=1}^{N} f(\hat{x}_j)h_j \right| \leq \frac{1}{2} \sum_{j=1}^{N} \left(\max_{y \in I_j} |f'(y)|\, h_j \right) h_j. \quad (202.8)$$

This generalizes the estimate of the Fundamental Theorem to non-uniform partitions. We can see that (202.8) implies that Q_h tends to zero as the maximal step size tends to zero by estimating further:

$$Q_h \leq \frac{1}{2} \max_{[a,b]} |f'| \sum_{j=1}^{N} h_j \max_{1 \leq j \leq N} h_j = \frac{1}{2}(b - a) \max_{[a,b]} |f'| \max_{1 \leq j \leq N} h_j. \quad (202.9)$$

So Q_h tends to zero at the same rate that $\max h_j$ tends to zero.

202.2 The integral as a Limit of Riemann Sums

We now return to the (subtle) question posed at the end of the Chapter The Integral: Will all limits of Riemann sum approximations (as the maximal subinterval tends to zero) of a certain integral be the same? We recall that we defined the integral using a particular uniform partition and we now

ask if any limit of non-uniform partitions will be the same. The affirmative answer follows from the last statement of the previous section: The quadrature error Q_h tends to zero as $\max h_j$ tends to zero, under the assumption that $\max_{[a,b]} |f'|$ is finite, that is $|f'(x)|$ is bounded on $[a,b]$. This proves the uniqueness of limits of Riemann sum approximations of a certain integral as the maximal subinterval tends to zero, under the assumption that the derivative of the integrand is bounded. This assumption can naturally be relaxed to assuming that the integrand is Lipschitz continuous. We sum up:

Theorem 202.1 *The limit (as the maximal subinterval tends to zero) of Riemann sum approximations of an integral of a Lipschitz continuous function, is unique.*

202.3 The Midpoint Rule

We now analyze the quadrature formula in which the quadrature point is chosen to be the midpoint of each subinterval, $\hat{x}_j = \frac{1}{2}(x_{j-1} + x_j)$. It turns out that this choice gives a formula that is more accurate than any other rectangle rule on a given mesh provided f has a Lipschitz continuous second derivative. Taylor's theorem implies that for $x \in [x_{j-1}, x_j]$,

$$f(x) = f(\hat{x}_j) + f'(\hat{x}_j)(x - \hat{x}_j) + \frac{1}{2}f''(y)(x - \hat{x}_j)^2,$$

for some $y \in [x_{j-1}, x_j]$ if we assume that f'' is Lipschitz continuous. We argue as above by integrating over $[x_{j-1}, x_j]$. Now however we use the fact that

$$\int_{x_{j-1}}^{x_j} (x - \hat{x}_j)\, dx = \int_{x_{j-1}}^{x_j} (x - (x_j + x_{j-1})/2)\, dx = 0$$

which holds only when $\hat{x}_j$ is the midpoint of $[x_{j-1}, x_j]$. This gives

$$\begin{aligned}\Big| \int_{x_{j-1}}^{x_j} f(x)\, dx - f(\hat{x}_j)h_j \Big| &\leq \frac{1}{2} \max_{y \in I_j} |f''(y)| \int_{x_{j-1}}^{x_j} (x - \hat{x}_j)^2\, dx \\ &\leq \frac{1}{24} \max_{y \in I_j} |f''(y)| h_j^3.\end{aligned}$$

Now summing the errors on each subinterval, we obtain the following estimate on the total error

$$Q_h \leq \frac{1}{24} \sum_{j=1}^{N} \left(\max_{y \in I_j} |f''(y)|\, h_j^2 \right)\, h_j. \tag{202.10}$$

To understand the claim that this formula is more accurate than any other rectangle rule, we estimate further

$$Q_h \leq \frac{1}{24}(b-a) \max_{[a,b]} |f''|\, \max h_j^2$$

which says that the error decreases as $\max h_j$ decreases like $\max h_j^2$. Compare this to the general result (202.9), which says that the error decreases like $\max h_j$ for general rectangle rules. If we halve the step size $\max h_j$ then in a general rectangle rule the error decreases by a factor of two but in the midpoint rule the error decreases by a factor of *four*. We say that the midpoint rule converges at a *quadratic* rate while the general rectangle rule converges at a *linear* rate.

We illustrate the accuracy of these methods and the error bounds by approximating

$$\log(4) = \int_1^4 \frac{dx}{x} \approx \sum_{j=1}^{N} \frac{h_j}{\hat{x}_j}$$

both with the original rectangle rule with $\hat{x}_j$ equal to the left-hand endpoint x_{j-1} of each subinterval and the midpoint rule. In both cases, we use a constant stepsize $h_i = (4-1)/N$ for $i = 1, 2, \cdots, N$. It is straightforward to evaluate (202.8) and (202.10) because $|f'(x)| = 1/x^2$ and $|f''(x)| = 2/x^3$ are both decreasing functions. We show the results for four different values of N.

	Rectangle rule		Midpoint rule	
N	True error	Error bound	True error	Error bound
25	.046	.049	.00056	.00056
50	.023	.023	.00014	.00014
100	.011	.011	.000035	.000035
200	.0056	.0057	.0000088	.0000088

These results show that the error bounds (202.8) and (202.10) can give quite accurate estimates of the true error. Also note that the midpoint rule is much more accurate than the general rectangle rule on a given mesh and moreover the error in the midpoint rule goes to zero quadratically with the error decreasing by a factor of 4 each time the number of steps is doubled.

202.4 Adaptive Quadrature

In this section, we consider the optimization problem of trying to compute an approximation of an integral to within a given accuracy at a relatively low cost. To simplify the discussion, we use the original rectangle rule with $\hat{x}_j$ equal to the left-hand endpoint x_{j-1} of each subinterval to compute the approximation. The optimization problem becomes to compute an approximation with error smaller than a given tolerance TOL using the least number of steps. Since we do not know the error of the approximation, we use the quadrature estimate (202.8) to estimate the error. The optimization problem is therefore to find a partition $\{x_j\}_{j=0}^N$ using the smallest number

of points N that satisfies the stopping criterion

$$\sum_{j=1}^{N} \left(\max_{y \in I_j} |f'(y)| \, h_j \right) h_j \leq \text{TOL}. \tag{202.11}$$

This equation suggests that we should adjust or *adapt* the stepsizes h_j depending on the size of $\max_{I_j} |f'|$. If $\max_{I_j} |f'|$ is large, then h_j should be small, and vice versa. Trying to find such an optimized partition is referred to as *adaptive* quadrature, because we seek a partition suitably adapted to the nature of the integrand $f(x)$.

There are several possible strategies for finding such a partition and we consider two here.

In the first strategy, or adaptive algorithm, we estimate the sum in (202.11) as follows

$$\sum_{j=1}^{N} \left(\max_{I_j} |f'| \, h_j \right) h_j \leq (b-a) \max_{1 \leq j \leq N} \left(\max_{I_j} |f'| \, h_j \right),$$

where we use the fact that $\sum_{j=1}^{N} h_j = b - a$. It follows that (202.11) is satisfied if the steps are chosen by

$$h_j = \frac{\text{TOL}}{(b-a) \max_{I_j} |f'|} \quad \text{for } j = 1, ..., N. \tag{202.12}$$

In general, this corresponds to a nonlinear equation for h_j since $\max_{I_j} |f'|$ depends on h_j.

We apply this adaptive algorithm to the computation of $\log(b)$ and obtain the following results

TOL	b	Steps	Approximate Area	Error
.05	4.077	24	1.36	.046
.005	3.98	226	1.376	.0049
.0005	3.998	2251	1.38528	.0005
.00005	3.9998	22501	1.3861928	.00005

The reason b varies slightly in these results is due to the strategy we use to implement (202.12). Namely, we specify the tolerance and then search for the value of N that gives the closest b to 4.

We plot the sequence of mesh sizes for TOL = .01 in Fig. 202.2, where the adaptivity is plainly visible. In contrast, if we compute with a uniform mesh, we find using (202.11) that we need $N = 9/\text{TOL}$ points to guarantee an accuracy of TOL. For example, this means using 900 points to guarantee an accuracy of .01, which is significantly more than needed for the adapted mesh.

The second adaptive algorithm is based on an *equidistribution of error* in which the steps h_j are chosen so that the contribution to the error from each sub-interval is roughly equal. Intuitively, this should lead to the least number of intervals since the largest error reduction is gained if we subdivide the interval with largest contribution to the error. In this case, we estimate the sum on the left-hand side of (202.11) by

$$\sum_{j=1}^{N} \left(\max_{I_j} |f'| \, h_j \right) h_j \leq N \max_{1 \leq j \leq N} \left(\max_{I_j} |f'| \, h_j^2 \right)$$

and determine the steps h_j by

$$h_j^2 = \frac{\text{TOL}}{N \max\limits_{I_j} |f'|} \quad \text{for } j = 1, ..., N. \tag{202.13}$$

As above, we have to solve a nonlinear equation for h_j, now with the additional complication of the explicit presence of the total number of steps N.

We implement (202.13) to compute of $\log(b)$ with $b \approx 4$ and obtain the following results:

TOL	b	Steps	Appr. Area	Error
.05	4.061	21	1.36	.046
.005	4.0063	194	1.383	.005
.0005	3.9997	1923	1.3857	.0005
.00005	4.00007	19220	1.38626	.00005

We plot the sequence of step sizes for TOL = .01 in (202.2). We see that at

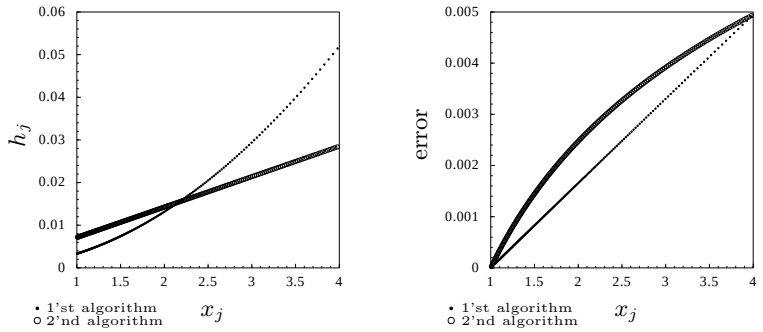


FIGURE 202.2. On the left, we plot the step sizes generated by two adaptive algorithms for the integration of $\log(4)$ using TOL = .01. On the right, we plot the errors of the same computations versus x.

every tolerance level, the second adaptive strategy (202.13) gives the same

accuracy at $x_N \approx 4$ as (202.12) while using fewer steps. It thus seems that the second algorithm is more efficient.

We compare the efficiency of the two adaptive algorithms by estimating the number of steps N required to compute $\log(x)$ to a given tolerance TOL in each case. We begin by noting that the equality

$$N = \frac{h_1}{h_1} + \frac{h_2}{h_2} + \cdots + \frac{h_N}{h_N},$$

implies that, assuming $x_N > 1$,

$$N = \int_1^{x_N} \frac{dy}{h(y)},$$

where $h(y)$ is the piecewise constant *mesh function* with the value $h(s) = h_j$ for $x_{j-1} < s \leq x_j$. In the case of the second algorithm, we substitute the value of h given by (202.13) into the integral to get, recalling that $f(y) = 1/y$ so that $f'(y) = -1/y^2$,

$$N \approx \frac{\sqrt{N}}{\sqrt{\text{TOL}}} \int_1^{x_N} \frac{dy}{y},$$

or

$$N \approx \frac{1}{\text{TOL}} \left(\log(x_N)\right)^2. \tag{202.14}$$

Making a similar analysis of the first adaptive algorithm, we get

$$N \approx \frac{x_{N-1}}{\text{TOL}}\left(1 - \frac{1}{x_N}\right). \tag{202.15}$$

We see that in both cases, N is inversely proportional to TOL. However, the number of steps needed to reach the desired accuracy using the first adaptive algorithm increases much more quickly as x_N increases than the number needed by the second algorithm, i.e. at a linear rate as opposed to a logarithmic rate. Note that the case $0 < x_N < 1$ may be reduced to the case $x_N > 1$ by replacing x_N by $1/x_N$ since $\log(x) = -\log(1/x)$.

If we use (202.12) or (202.13) to choose the steps h_j over the interval $[a, x_N]$, then of course the quadrature error over any smaller interval $[a, x_i]$ with $i \leq N$, is also smaller than TOL. For the first algorithm (202.12), we can actually show the stronger estimate

$$\left| \int_a^{x_i} f(y)\, dy - \sum_{j=1}^{i} f(x_j)\, h_j \right| \leq \frac{x_i - a}{x_N - a}\, TOL, \quad 1 \leq i \leq N, \tag{202.16}$$

i.e., the error grows at most linearly with x_i as i increases. However, this does not hold in general for the second adaptive algorithm. In Fig. 202.2, we plot the errors versus x_i for $x_i \leq x_N$ resulting from the two adaptive algorithms with TOL $= .01$. We see the linear growth predicted for the first algorithm (202.12) while the error from the second algorithm (202.13) is larger for $1 < x_i < x_N$.

Chapter 202 Problems

202.1. Estimate the error using endpoint and midpoint quadrature for the following integrals: (a) $\int_0^2 2s\,ds$, (b) $\int_0^2 s^3\,ds$, and (c) $\int_0^2 \exp(-s)\,ds$ using $h = .1$, .01, .001 and .0001. Discuss the results.

202.2. Compute approximations of the following integrals using adaptive quadrature (a) $\int_0^2 2s\,ds$, (b) $\int_0^2 s^3\,ds$, and (c) $\int_0^2 \exp(-s)\,ds$. Discuss the results.

202.3. Compare theoretically and experimentally the number of steps of (202.12) and (202.13) for the computation of integrals of the form $\int_x^1 f(y)\,dy$ for $x > 0$, where $f(y) \sim y^{-\alpha}$ with $\alpha > 1$.

202.4. The *trapezoidal rule* takes the form

$$\int_{x_{j-1}}^{x_j} f(x)dx \approx (x_j - x_{j-1})(f(x_{j-1}) + f(x_j))/2. \tag{202.17}$$

Show that the quadrature is exact if $f(x)$ is a first order polynomial, and give an estimate of the quadrature error analogous to that of the midpoint rule. Compare the the midpoint and the trapezoidal method.

202.5. Design different adaptive quadrature algorithms based on the midpoint rule and make comparisons.

202.6. Consider a quadrature formula of the form

$$\int_a^b f(x)dx \approx (b-a)(f(\hat{x}_1) + f(\hat{x}_2))/2. \tag{202.18}$$

Determine the quadrature points $\hat{x}_1$ and $\hat{x}_2$, so that the quadrature formula is exact for $f(x)$ a second order polynomial. This quadrature rule is called the two-point Gauss rule. Check for which order of polynomials the resulting quadrature formula is exact.

202.7. Compute the value of $\int_0^1 \frac{1}{1+x^2}dx$ by quadrature. Multiply the result by 4. Do you recognize this number?

203

The Exponential Function $\exp(x) = e^x$

The need for mathematical skills is greater than ever, but it is widely recognized that, as a consequence of computer developments, there is a need for a shift in emphasis in the teaching of mathematics to students studying engineering. This shift is away from the simple mastery of solution techniques and towards development of a greater understanding of mathematical ideas and processes together with efficiency in applying this understanding to the formulation and analysis of physical phenomena and engineering systems. (Glyn James, in Preface to Modern Engineering Mathematics, 1992)

Because of the limitations of human imagination, one ought to say: everything is possible - and a bit more. (Horace Engdahl)

203.1 Introduction

In this chapter we return to study of the *exponential function* $\exp(x)$, which we have met above in Chapter *A very short course in Calculus* and Chapter *Galileo, Newton, Hooke, Malthus and Fourier*, and which is one of the basic functions of Calculus, see Fig. 203.1. We have said that $\exp(x)$ for $x > 0$ is the solution to the following initial value problem: Find a function $u(x)$ such that

$$\begin{aligned} u'(x) &= u(x) \quad \text{for } x > 0, \\ u(0) &= 1. \end{aligned} \tag{203.1}$$

Evidently, the equation $u'(x) = u(x)$ states that the rate of growth $u'(x)$ is equal to the quantity $u(x)$ itself, that is, the exponential function $\exp(x) =$

e^x is characterized by the property that its derivative is equal to itself: $D\exp(x) = \exp(x)$. What a wonderful almost divine property! We also denote the exponential function by e^x, that is, $e^x = \exp(x)$ and $De^x = e^x$.

In this chapter, we give a constructive proof of the *existence* of a unique solution to the initial value problem (203.1), that is, we *prove* the existence of the exponential function $\exp(x) = e^x$ for $x > 0$. Note that above, we just *claimed* the existence of solutions. As always, a constructive proof also shows how we may *compute* $\exp(x)$ for different values of x.

Below we extend $\exp(x)$ to $x < 0$ by setting $\exp(x) = (\exp(-x))^{-1}$ for $x < 0$, and show that $\exp(-x)$ solves the initial value problem $u'(x) = -u(x)$ for $x > 0$, $u(0) = 1$. We plot the functions $\exp(x)$ and $\exp(-x)$ for $x \geq 0$ in Fig. 203.1. We notice that $\exp(x)$ is increasing and $\exp(-x)$ is decreasing with increasing x, and that $\exp(x)$ is positive for all x. Combining $\exp(x)$ and $\exp(-x)$ for $x \geq 0$ defines $\exp(x)$ for $-\infty < x < \infty$. Below we show that $D\exp(x) = \exp(x)$ for $-\infty < x < \infty$.

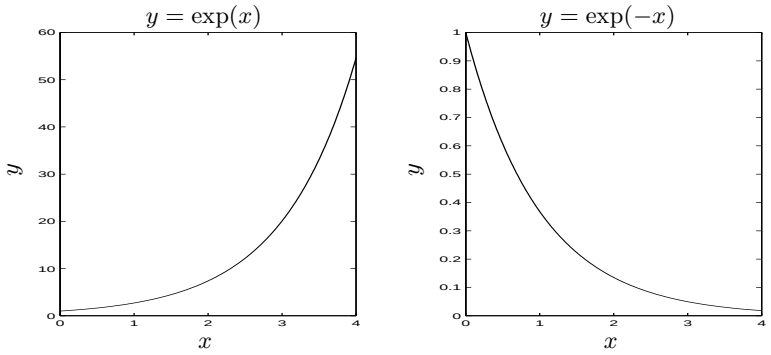


FIGURE 203.1. The exponential functions $\exp(x)$ and $\exp(-x)$ for $x \geq 0$.

The problem (203.1) is a special case of the Malthus population model (**??**), which also models a large variety of phenomena in e.g. physics and economy:

$$\begin{cases} u'(x) = \lambda u(x) & \text{for } x > 0, \\ u(0) = u_0. \end{cases} \tag{203.2}$$

where λ is a constant and u_0 is a given initial value. The solution of this problem can be expressed in terms of the exponential function as

$$u(x) = \exp(\lambda x)u_0 \quad \text{for } x \geq 0. \tag{203.3}$$

This follows directly from the fact that by the Chain rule, $D\exp(\lambda x) = \exp(\lambda x)\lambda$, where we used that $D\exp(x) = \exp(x)$. Assuming $u_0 > 0$ so that $u(x) > 0$, evidently the sign of λ determines if u decreases ($\lambda < 0$) or increases ($\lambda > 0$). In Fig. 203.1, we plotted the solutions of (203.2) with $\lambda = \pm 1$ and $u_0 = 1$.

Before going into the construction of the exponential function $\exp(x)$, we recall two of the key applications of (203.2): population dynamics and banking. Here x represents time and we change notation, replacing x by t.

EXAMPLE 203.1. We consider a population with constant birth and death rates β and δ, which are the numbers of births and deaths per individual creature per unit time. With $u(t)$ denoting the population at time t, there will be during the time interval from t to $t + \Delta t$ with Δt a small increment, approximately $\beta u(t)\Delta t$ births and $\delta u(t)\Delta t$ deaths. Hence the change in population over the time interval is approximately

$$u(t + \Delta t) - u(t) \approx \beta u(t)\Delta t - \delta u(t)\Delta t$$

and therefore

$$\frac{u(t + \Delta t) - u(t)}{\Delta t} \approx (\beta - \delta)u(t),$$

where the approximation improves as we decrease Δt. Taking the limit as $\Delta t \to 0$, assuming $u(t)$ is a differentiable function, we obtain the model $u'(t) = (\beta - \delta)u(t)$. Assuming the initial population at $t = 0$ is equal to u_0, leads to the model (203.2) with $\lambda = \beta - \delta$, with solution $u(x) = \exp(\lambda x)u_0$.

EXAMPLE 203.2. An investment u in a saving account earning 5% interest compounded continuously and beginning with \$2000 at time $t = 0$, satisfies

$$\begin{cases} u' = 1.05u, & t > 0, \\ u(0) = 2000, \end{cases}$$

and thus $u(t) = \exp(1.05t)2000$ for $t \geq 0$.

203.2 Construction of the Exponential exp(x) for $x \geq 0$

In the proof of the Fundamental Theorem, we constructed the solution $u(x)$ of the initial value problem

$$\begin{cases} u'(x) = f(u(x), x) & \text{for } 0 < x \leq 1, \\ u(0) = u_0, \end{cases} \tag{203.4}$$

in the case that $f(u(x), x) = f(x)$ depends only on x and not on $u(x)$. We constructed the solution $u(x)$ as the limit of a sequence of functions $\{U^n(x)\}_{n=1}^{\infty}$, where $U^n(x)$ is a piecewise linear function defined at a set of nodes $x_i^n = ih_n$, $i = 0, 1, 2, ..., N = 2^n$, $h_n = 2^{-n}$, by the relations

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(x_{i-1}^n) \quad \text{for } i = 1, 2, ..., N, \quad U^n(0) = u_0. \tag{203.5}$$

We shall now apply the same technique to construct the solution of (203.1), which has the form (203.4) with $f(u(x), x) = u(x)$ and $u_0 = 1$. We carry out the proof in a form which generalizes in a straight-forward way to any system of equations of the form (203.4), which really includes a very wide range of problems. We hope this will motivate the reader to carefully follow every step of the proof, to get properly prepared for the highlight Chapter *The general initial value problem*.

We construct the solution $u(x)$ of (203.1) for $x \in [0, 1]$ as the limit of a sequence of piecewise linear functions $\{U^n(x)\}_{n=1}^{\infty}$ defined at the nodes by the formula

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n U^n(x_{i-1}^n) \quad \text{for } i = 1, 2, ..., N, \tag{203.6}$$

with $U^n(0) = 1$, which is an analog of (203.5) obtained by replacing $f(x_{i-1}^n)$ by $U^n(x_{i-1}^n)$ corresponding to replacing $f(x)$ by $f(x, u(x)) = u(x)$. Using the formula we can compute the values $U^n(x_i^n)$ one after the other for $i = 1, 2, 3, ...$, starting from the initial value $U^n(0) = 1$, that is marching forward in time with x representing time.

We can write (203.6) in the form

$$U^n(x_i^n) = (1 + h_n)U^n(x_{i-1}^n) \quad \text{for } i = 1, 2, ..., N, \tag{203.7}$$

and conclude since $U^n(x_i^n) = (1 + h_n)U^n(x_{i-1}^n) = (1 + h_n)^2 U^n(x_{i-2}^n) = (1 + h_n)^3 U^n(x_{i-3}^n)$ and so on, that the nodal values of $U^n(x)$ are given by the formula

$$U^n(x_i^n) = (1 + h_n)^i, \quad \text{for } i = 0, 1, 2, ..., N, \tag{203.8}$$

where we also used that $U^n(0) = 1$. We illustrate in Fig. 203.2. We may

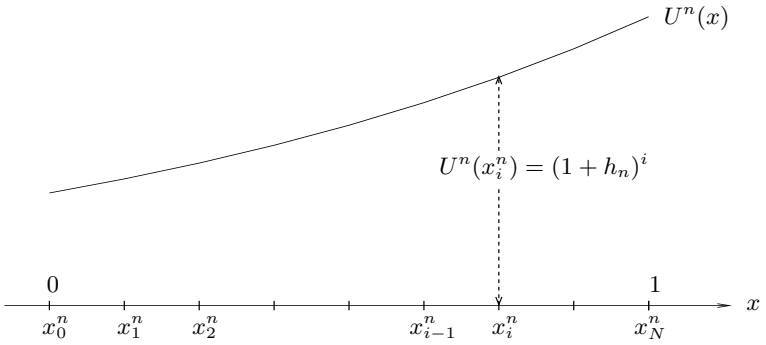


FIGURE 203.2. The piecewise linear approximate solution $U^n(x) = (1 + h_n)^i$.

view $U^n(x_i^n)$ as the capital obtained at time $x_i^n = ih_n$ starting with a unit capital at time zero, if the interest rate at each capitalization is equal to h_n.

To analyze the convergence of $U^n(x)$ as $n \to \infty$, we first prove a bound on the nodal values $U^n(x_i^n)$, by taking the logarithm of (203.8) and using the inequality $\log(1+x) \leq x$ for $x > 0$ from Problem 201.4, to obtain

$$\log(U^n(x_i^n) = i\log(1+h_n) \leq ih_n = x_i^n \leq 1 \quad \text{for } i = 1, 2, .., N.$$

It follows that

$$U^n(x_i^n) = (1+h_n)^i \leq 4 \quad \text{for } i = 1, 2, .., N, \tag{203.9}$$

since $\log(4) > 1$ according to Problem 201.1, and $\log(x)$ is increasing. Since $U^n(x)$ is linear between the nodes, and obviously $U^n(x) \geq 1$, we find that $1 \leq U^n(x) \leq 4$ for all $x \in [0,1]$.

We now show that $\{U^n(x)\}_{n=1}^{\infty}$ is a Cauchy sequence for each fixed $x \in [0,1]$. To see this, we first estimate $|U^n(x) - U^{n+1}(x)|$ at the node points $x = x_i^n = ih_n = 2ih_{n+1} = x_{2i}^{n+1}$ for $i = 0, 1, ..., N$, see Fig. 203.3. Notice that $h_{n+1} = h_n/2$ so that two steps with mesh size h_{n+1} corresponds to one step with mesh size h_n. We start by subtracting

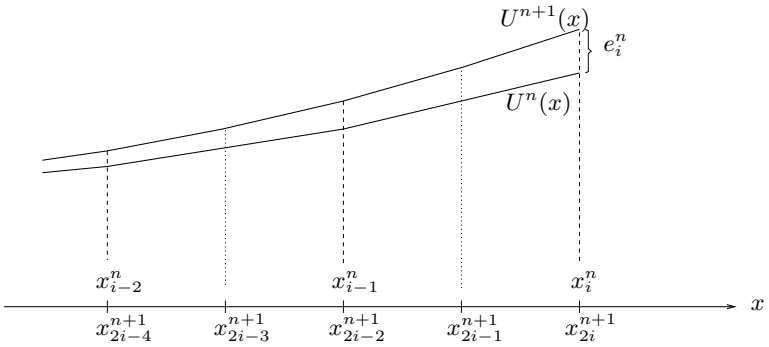


FIGURE 203.3. $U^n(x)$ and $U^{n+1}(x)$.

$$U^{n+1}(x_{2i}^{n+1}) = (1+h_{n+1})U^{n+1}(x_{2i-1}^{n+1}) = (1+h_{n+1})^2 U^{n+1}(x_{2i-2}^{n+1}),$$

from (203.6), using that $x_i^n = x_{2i}^{n+1}$, and setting $e_i^n = U^n(x_i^n) - U^{n+1}(x_i^n)$, to get

$$e_i^n = (1+h_n)U^n(x_{i-1}^n) - (1+h_{n+1})^2 U^{n+1}(x_{i-1}^n),$$

which we may rewrite using that $(1+h_{n+1})^2 = 1 + 2h_{n+1} + h_{n+1}^2$ and $2h_{n+1} = h_n$, as

$$e_i^n = (1+h_n)e_{i-1}^n - h_{n+1}^2 U^{n+1}(x_{i-1}^n).$$

It follows using the bound $1 \leq U^{n+1}(x) \leq 4$ for $x \in [0,1]$, that

$$|e_i^n| \leq (1+h_n)|e_{i-1}^n| + 4h_{n+1}^2.$$

Inserting the corresponding estimate for e^n_{i-1}, we get

$$\begin{aligned}|e^n_i| &\le (1+h_n)((1+h_n)|e^n_{i-2}| + 4h^2_{n+1}) + 4h^2_{n+1}\\ &= (1+h_n)^2|e^n_{i-2}| + 4h^2_{n+1}(1+(1+h_n)).\end{aligned}$$

Continuing this way and using that $e^n_0 = 0$, we obtain for $i = 1, ..., N$,

$$|e^n_i| \le 4h^2_{n+1}\sum_{k=0}^{i-1}(1+h_n)^k = h^2_n\sum_{k=0}^{i-1}(1+h_n)^k.$$

Using now the fact that

$$\sum_{k=0}^{i-1} z^k = \frac{z^i - 1}{z-1} \tag{203.10}$$

with $z = 1 + h_n$, we thus obtain for $i = 1, ..., N$,

$$|e^n_i| \le h^2_n\frac{(1+h_n)^i - 1}{h_n} = h_n((1+h_n)^i - 1) \le 3h_n,$$

where we again used that $(1+h_n)^i = U^n(x^n_i) \le 4$. We have thus proved that for $\bar{x} = x^n_j$, $j = 1, ..., N$,

$$|U^n(\bar{x}) - U^{n+1}(\bar{x})| = |e^n_j| \le 3h_n,$$

which is analogous to the central estimate (199.24) in the proof of the Fundamental Theorem.

Iterating this estimate over n as in the proof of (199.25), we get for $m > n$,

$$|U^n(\bar{x}) - U^m(\bar{x})| \le 6h_n, \tag{203.11}$$

which shows that $\{U^n(\bar{x})\}_{n=1}^{\infty}$ is a Cauchy sequence and thus converges to a real number $u(\bar{x})$, which we choose to denote by $\exp(\bar{x}) = e^{\bar{x}}$. As in the proof of the Fundamental Theorem we can extend to a function $u(x) = \exp(x) = e^x$ defined for $x \in [0, 1]$. Letting m tend to infinity in (203.11), we see that

$$|U^n(x) - \exp(x)| \le 6h_n \quad \text{for } x \in [0, 1]. \tag{203.12}$$

By the construction, we have if $\bar{x} = jh_n$ so that $h_n = \frac{\bar{x}}{j}$, noting that $j \to \infty$ as $n \to \infty$:

$$\exp(\bar{x}) = \lim_{n\to\infty}(1+h_n)^j = \lim_{j\to\infty}(1+\frac{\bar{x}}{j})^j,$$

that is,

$$\exp(x) = \lim_{j\to\infty}(1+\frac{x}{j})^j \qquad \text{for } x \in [0, 1]. \tag{203.13}$$

In particular, we define the number e by

$$e \equiv \exp(1) = \lim_{j\to\infty} (1 + \frac{1}{j})^j. \tag{203.14}$$

We refer to e as the *base of the exponential function*. We will prove below that $\log(e) = 1$.

It remains to verify that the function $u(x) = \exp(x) = e^x$ constructed above, indeed satisfies (203.1) for $0 < x \leq 1$. We note that choosing $\bar{x} = jh_n$ and summing over i in (203.6), we get

$$U^n(\bar{x}) = \sum_{i=1}^{j} U^n(x_{i-1}^n)h_n + 1,$$

which we can write as

$$U^n(\bar{x}) = \sum_{i=1}^{j} u(x_{i-1}^n)h_n + 1 + E_n,$$

where $u(x) = \exp(x)$, and using (203.12),

$$|E_n| = |\sum_{i=1}^{j}(U^n(x_{i-1}^n) - u(x_{i-1}^n))h_n| \leq 6h_n \sum_{i=1}^{j} h_n \leq 6h_n,$$

since obviously $\sum_{i=1}^{j} h_n \leq 1$. Letting n tend to infinity and using $\lim_{n\to\infty} E_n = 0$, we see that $u(\bar{x}) = \exp(\bar{x})$ satisfies

$$u(\bar{x}) = \int_0^{\bar{x}} u(x)\,dx + 1.$$

Differentiating this equality with respect to $\bar{x}$, we get $u'(\bar{x}) = u(\bar{x})$ for $\bar{x} \in [0, 1]$, and we have now proved that the constructed function $u(x)$ indeed solves the given initial value problem.

We conclude the proof by showing uniqueness. Thus, assume that we have two uniformly differentiable functions $u(x)$ and $v(x)$ such that $u'(x) = u(x)$ and $v'(x) = v(x)$ for $x \in (0, 1]$, and $u(0) = v(0) = 1$. The $w = u - v$ satisfies $w'(x) = w(x)$ and $w(0) = 0$, and thus by the Fundamental Theorem,

$$w(x) = \int_0^x w'(y)\,dy = \int_0^x w(y)\,dy \quad \text{for } x \in [0, 1].$$

Setting $a = \max_{0\leq x\leq 0.5} |w(x)|$, we thus have

$$a \leq \int_0^{0.5} a\,dy = 0.5a$$

which is possible only if $a = 0$ showing uniqueness for $0 \le x \le 0.5$. Repeating the argument on $[0.5, 1]$ proves that $w(x) = 0$ for $x \in [0, 1]$ and the uniqueness follows.

The proof immediately generalizes to $x \in [0, b]$ where b is any positive real number. We now summarize:

Theorem 203.1 *The initial value problem $u'(x) = u(x)$ for $x > 0$, and $u(0) = 1$, has a unique solution $u(x) = \exp(x)$ given by (203.13).*

203.3 Extension of the Exponential $\exp(x)$ to $x < 0$

If we define

$$\exp(-x) = \frac{1}{\exp(x)} \quad \text{for } x \ge 0,$$

then we find that

$$D\exp(-x) = D\frac{1}{\exp(x)} = -\frac{D\exp(x)}{(\exp(x))^2} = -\frac{1}{\exp(x)} = -\exp(-x). \tag{203.15}$$

We conclude that $\exp(-x)$ solves the initial value problem

$$u'(x) = -u(x) \quad \text{for } x > 0,\ u(0) = 1.$$

203.4 The Exponential Function $\exp(x)$ for $x \in \mathbb{R}$

Piecing together the functions $\exp(x)$ and $\exp(-x)$ with $x \ge 0$, we obtain the function $u(x) = \exp(x)$ defined for $x \in \mathbb{R}$, which satisfies $u'(x) = u(x)$ for $x \in \mathbb{R}$ and $u(0) = 1$, see Fig. 203.4 and Fig. 203.5.

To see that $\frac{d}{dx}\exp(x)$ for $x < 0$, we set $y = -x > 0$ and compute $\frac{d}{dx}\exp(x) = \frac{d}{dy}\exp(-y)\frac{dy}{dx} = -\exp(-y)(-1) = \exp(x)$, where we used (203.15).

203.5 An Important Property of $\exp(x)$

We now prove the basic property of the exponential function $\exp(x)$ using the fact that $\exp(x)$ satisfies the differential equation $D\exp(x) = \exp(x)$. We start considering the initial value problem

$$u'(x) = u(x) \quad \text{for } x > a, \quad u(a) = u_a, \tag{203.16}$$

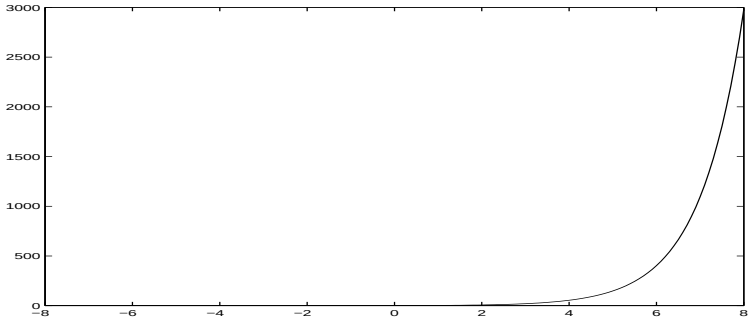


FIGURE 203.4. The exponential $\exp(x)$ for $x \in [-2.5, 2.5]$

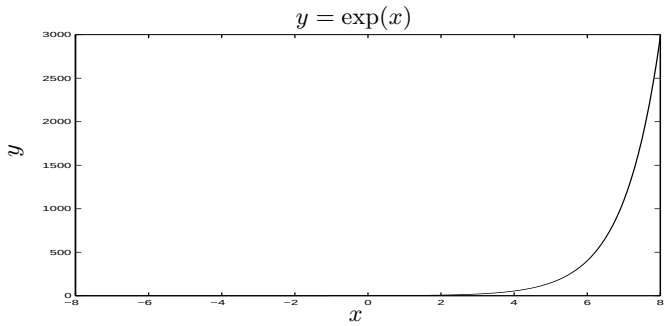


FIGURE 203.5. The exponential $\exp(x)$ for $x \in [-8, 8]$

with initial value at some point a other than zero. Setting $x = y + a$ and $v(y) = u(y + a) = u(x)$, we obtain by the Chain rule

$$v'(y) = \frac{d}{dy}u(y + a) = u'(y + a)\frac{d}{dy}(y + a) = u'(x),$$

and thus $v(y)$ satisfies the differential equation

$$v'(y) = v(y) \quad \text{for } y > 0, \quad v(0) = u_a.$$

This means that

$$v(y) = \exp(y)u_a \quad \text{for } y > 0.$$

Going back to the original variables, using that $y = x - a$, we find that the solution of (203.16) is given by

$$u(x) = \exp(x - a)u_a \quad \text{for } x \geq a. \tag{203.17}$$

We now prove that for $a, b \in \mathbb{R}$,

$$\exp(a + b) = \exp(a)\exp(b) \quad \text{or } e^{a+b} = e^a e^b, \tag{203.18}$$

which is the basic property of the exponential function. We do this by using the fact that $u(x) = \exp(x)$ satisfies the differential equation $u'(x) = u(x)$ and $\exp(0) = 1$. We have on the one hand that $u(a+b) = \exp(a+b)$ is the value of the solution $u(x)$ for $x = a+b$. We may reach to $x = a+b$, assuming $0 < a, b$ to start with, by first computing the solution $u(x) = \exp(x)$ from $x = 0$ up to $x = a$, which gives $u(a) = \exp(a)$. We next consider the following problem

$$v'(x) = v(x) \quad \text{for } x > a, \quad v(a) = \exp(a)$$

with solution $v(x) = \exp(x-a)\exp(a)$ for $x \geq a$. We have $v(x) = u(x)$ for $x \geq a$, since $u(x)$ also solves $u'(x) = u(x)$ for $x > a$, and $u(a) = \exp(a)$. Thus $v(b+a) = u(a+b)$, which translates into the desired equality $\exp(b)\exp(a) = \exp(a+b)$. The proof extends to any $a, b \in \mathbb{R}$.

203.6 The Inverse of the Exponential Is the Logarithm

We shall now prove that

$$\log(\exp(x)) = x \quad \text{for } x \in \mathbb{R}, \tag{203.19}$$

and conclude that

$$y = \exp(x) \quad \text{if and only if } x = \log(y), \tag{203.20}$$

which states that the inverse of the exponential is the logarithm.

We prove (203.19) by differentiation to get by the Chain rule for $x \in \mathbb{R}$,

$$\frac{d}{dx}(\log(\exp(x)) = \frac{1}{\exp(x)}\frac{d}{dx}(\exp(x)) = \frac{1}{\exp(x)}\exp(x) = 1,$$

and noting that $\log(\exp(0)) = \log(1) = 0$, which gives (203.19). Setting $x = \log(y)$ in (203.19), we have $\log(\exp(\log(y))) = \log(y)$, that is

$$\exp(\log(y)) = y \quad \text{for } y > 0. \tag{203.21}$$

We note in particular that

$$\exp(0) = 1 \text{ and } \log(e) = 1 \tag{203.22}$$

since $0 = \log(1)$ and $e = \exp(1)$ respectively.

In many Calculus books the exponential function $\exp(x)$ is defined as the inverse of the logarithm $\log(x)$ (which is defined as an integral). However, we prefer to directly prove the existence of $\exp(x)$ through its defining initial value problem, since this prepares the construction of solutions to general initial value problems.

203.7 The Function a^x with $a > 0$ and $x \in \mathbb{R}$

We recall that in Chapter *The function* $y = x^r$ we defined the function x^r for $r = p/q$ rational with p and $q \neq 0$ integers, and x is a positive real number, as the solution y to the equation $y^q = x^p$.

We thus are familiar with a^x with $a > 0$ and x rational, and we may extend to $x \in \mathbb{R}$ by defining:

$$a^x = \exp(x \log(a)). \tag{203.23}$$

We now prove the basic properties of a^x with $x \in \mathbb{R}$, that is, the positive number a raised to the power $x \in \mathbb{R}$, extending our previous experience with x rational. We note that by the Chain rule the function $u(x) = a^x$ satisfies the differential equation

$$u'(x) = \log(a)u(x)$$

and $u(0) = 1$. In particular, choosing $a = e = \exp(1)$, we find that $a^x = e^x = \exp(x)$, and we thus conclude that the exponential function $\exp(x)$ indeed equals the number e raised to the power x. Note that before we just used e^x just as a another notation for $\exp(x)$.

Using now the exponential law (203.18) for $\exp(x)$, we obtain with a direct computation using the definition (203.23) the following analog for a^x:

$$a^{x+y} = a^x a^y. \tag{203.24}$$

The other basic rule for a^x reads:

$$(a^x)^y = a^{xy}, \tag{203.25}$$

which follows from the following computation:

$$(a^x)^y = \exp(y \log(a^x)) = \exp(y \log(\exp(x \log(a)))) = \exp(yx \log(a)) = a^{xy}.$$

As indicated, the rules (203.24) and (203.25) generalize the corresponding rules with x and y rational proved above.

We conclude computing the derivative of the function a^x from the definition (203.23) using the Chain rule:

$$\frac{d}{dx}a^x = \log(a)a^x. \tag{203.26}$$

Chapter 203 Problems

203.1. Define $U^n(x_i^n)$ alternatively by $U^n(x_i^n) = U^n(x_{i-1}^n) \pm h_n U^n(x_i^n)$, and prove that the corresponding sequence $\{U^n(x)\}$ converges to $\exp(\pm x)$.

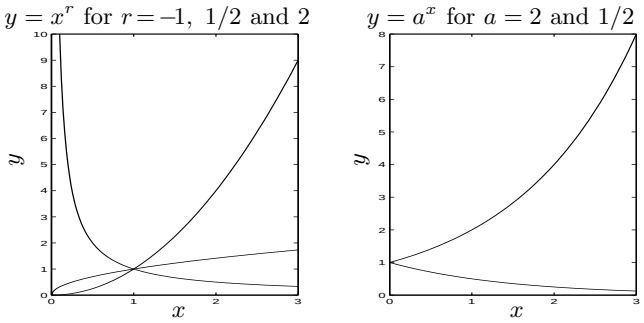


FIGURE 203.6. Examples of functions x^r and a^x.

203.2. Prove that for $x > 0$

$$(1 + \frac{x}{n})^n < \exp(x) \quad \text{for } n = 1, 2, 3 \ldots . \tag{203.27}$$

Hint: Take logarithm and use that $\log(1+x) < x$ for $x > 0$, and that the logarithm is increasing.

203.3. Prove directly the existence of a unique solution of $u'(x) = -u(x)$ for $x > 0$, $u(0) = 1$, that is construct $\exp(-x)$ for $x \geq 0$.

203.4. Show that the more often the bank capitalizes your interest, the better off you are, that is verify that

$$(1 + \frac{a}{n})^n \leq (1 + \frac{a}{n+1})^{n+1}. \tag{203.28}$$

Hint: Use the Binomial theorem.

203.5. Assume a bank offers "continuous capitalization" of the interest, corresponding to the (annual) interest rate a. What is then the "effective annual interest rate"?

203.6. Prove the differentiation formula $\frac{d}{dx}x^r = rx^{r-1}$ for $r \in \mathbb{R}$.

203.7. Prove the basic properties of the exponential function using that it is the inverse of the logarithm and use properties of the logarithm.

203.8. Given that the equation $u'(x) = u(x)$ has a solution for $x \in [0, 1]$ with $u(0) = 1$, construct a solution for all $x \geq 0$. Hint: use that if $u(x)$ satisfies $u'(x) = u(x)$ for $0 < x \leq 1$, then $v(x) = u(x - 1)$ satisfies $v'(x) = v(x)$ for $1 < x \leq 2$ and $v(1) = u(0)$.

203.9. Give the Taylor polynomial of order n with error term for $\exp(x)$ at $x = 0$.

203.10. Find a primitive function to (a) $x\exp(-x^2)$, (b) $x^3\exp(-x^2)$.

203.11. Compute the derivatives of the following functions: a) $f(x) = a^x$, $a > 0$, b) $f(x) = \exp(x + 1)$, c) $f(x) = x \exp(x^2)$, d) $f(x) = x^3 \exp(x^2)$, e) $f(x) = \exp(-x^2)$.

203.12. Compute the integrals $\int_0^1 f(x)dx$ of the functions in the previous exercise, except for the one in e), $f(x) = \exp(-x^2)$. Why do you think we left this one out?

203.13. Try to find the value of $\int_{-\infty}^{\infty} \exp(-x^2)dx$ numerically by quadrature. Square the result. Do you recognize this number?

203.14. Show that $\exp(x) \geq 1 + x$ for all x, not just for $x > -1$.

203.15. Show, by induction, that

$$\frac{d^n}{dx^n}\left(\mathrm{e}^x f(x)\right) = \mathrm{e}^x \left(1 + \frac{d}{dx}\right)^n f(x).$$

203.16. Prove (203.24) using the basic property (203.18) of the exponential and the definition (203.23).

203.17. Construct directly, without using the exponential function, the solution to the initial value problem $u'(x) = au(x)$ for $x \geq 0$ with $u(0) = 1$, where a is a real constant. Call the solution $\mathrm{aexp}(x)$. Prove that the function $\mathrm{aexp}(x)$ satisfies $\mathrm{aexp}(x + y) = \mathrm{aexp}(x)\mathrm{aexp}(y)$ for $x, y \geq 0$.

203.18. Define with $a > 0$ given, the function $y = \log_a(x)$ for $x > 0$ as the solution y to the equation $a^y = x$. With $a = e$ we get $\log_e(x) = \log(x)$, the natural logarithm. With $a = 10$ we get the so-called 10-logarithm. Prove that (i) $\log_a(xy) = \log_a(x) + \log_a(y)$ for $x, y > 0$, (ii) $\log_a(x^r) = r\log_a(x)$ for $x > 0$ and $r \in \mathbb{R}$, and (iii) $\log_a(x)\log(a) = \log(x)$ for $x > 0$.

203.19. Give the details of the proof of (203.26).

204
Trigonometric Functions

When I get to the bottom, I go back to the top of the slide where I stop and I turn and I go for a ride 'til I get to the bottom and I see you again. (Helter Skelter, Lennon-McCartney, 1968)

204.1 The Defining Differential Equation

In this chapter, we shall study the following *initial value problem for a second order differential equation*: Find a function $u(x)$ defined for $x \geq 0$ satisfying

$$u''(x) = -u(x) \quad \text{for } x > 0, \quad u(0) = u_0,\ u'(0) = u_1, \tag{204.1}$$

where u_0 and u_1 are given *initial data*. We here require two initial conditions because the problem involves a second order derivative. We may compare with the first order initial value problem: $u'(x) = -u(x)$ for $x > 0$, $\quad u(0) = u_0$, with the solution $u(x) = \exp(-x)$, which we studied in the previous chapter.

We shall demonstrate below, in Chapter *The general initial value problem*, that (204.1) has a unique solution for any given values of u_0 and u_1, and we shall in this chapter show that the solution with initial data $u_0 = 0$ and $u_1 = 1$ is an old friend, namely, $u(x) = \sin(x)$, and the solution with $u_0 = 1$ and $u_1 = 0$ is $u(x) = \cos(x)$. Here $\sin(x)$ and $\cos(x)$ are the usual trigonometric functions defined geometrically in Chapter *Pythagoras and Euclid*, with the change that we measure the angle x in the unit of *radians*

instead of degrees, with one radian being equal to $\frac{180}{\pi}$ degrees. In particular, we shall explain why one radian equals $\frac{180}{\pi}$ degrees.

We may thus define the trigonometric functions $\sin(x)$ and $\cos(x)$ as the solutions of (204.1) with certain initial data if we measure angles in the unit of radian. This opens a fresh route to understanding properties of the trigonometric functions by studying properties of solutions the differential equation (204.1), and we shall now explore this possibility.

We start by rewriting (204.1) changing the independent variable from x to t, since to aid our intuition we will use a mechanical interpretation of (204.1), where the independent variable represents time. We denote the derivative with respect to t with a dot, so that $\dot{u} = \frac{du}{dt}$, and $\ddot{u} = \frac{d^2u}{dt^2}$. We thus rewrite (204.1) as

$$\ddot{u}(t) = -u(t) \quad \text{for } t > 0, \quad u(0) = 0,\ \dot{u}(0) = 1, \tag{204.2}$$

where we chose $u_0 = 0$ and $u_1 = 1$ anticipating that we are looking for $\sin(t)$.

We now recall that (204.2) is a model of the motion of unit mass along a friction-less horizontal x-axis with the mass connected to one end of a Hookean spring with spring constant equal to 1 and with the other end connected to the origin, see Fig. **??**. We let $u(t)$ denotes the position ($x-$coordinate) of the mass at time t, and we assume that the mass is started at time $t = 0$ at the origin with speed $\dot{u}(0) = 1$, that is, $u_0 = 0$ and $u_1 = 1$. The spring exerts a force on the mass directed towards the origin, which is proportional to the length of the spring, since the spring constant is equal to 1, and the equation (204.2) expresses Newton's law: the acceleration $\ddot{u}(t)$ is equal to the spring force $-u(t)$. Because there is no friction, we would expect the mass to oscillate back and forth across the equilibrium position at the origin. We plot the solution $u(t)$ to (204.2) in Fig. 204.1, which clearly resembles the plot of the $\sin(t)$ function.

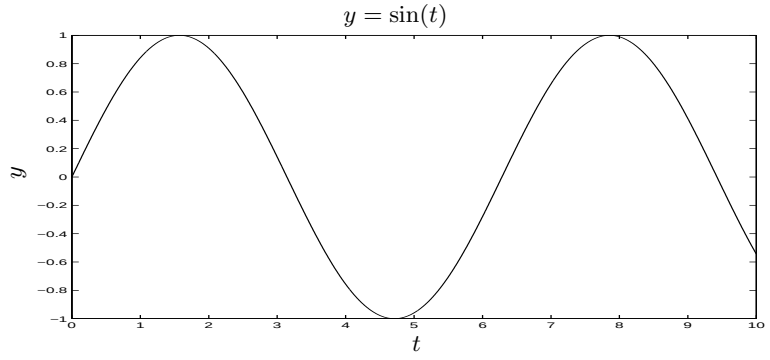


FIGURE 204.1. The solution of (204.2). Is it the function $\sin(t)$?

Let's now prove that our intuitive feeling indeed is correct, that is, let us prove that the solution of (204.2) indeed is our old friend $\sin(t)$. The key step then turns out to be to multiply our equation $\ddot{u} + u = 0$ by $\dot{u}$, to get

$$\frac{d}{dt}(\dot{u}^2 + u^2) = 2\dot{u}\ddot{u} + 2u\dot{u} = 2\dot{u}(\ddot{u} + u) = 0.$$

We conclude that $\dot{u}^2(t)+u^2(t)$ is constant for all t, and since $\dot{u}^2(0)+u^2(0) = 1 + 0 = 1$, we have found that the solution $u(t)$ of (204.2) satisfies the *conservation property*

$$\dot{u}^2(t) + u^2(t) = 1 \quad \text{for } t > 0, \tag{204.3}$$

which states that the point $(\dot{u}(t), u(t)) \in \mathbb{R}^2$ lies on the unit circle in $\mathbb{R}^2$, see Fig. 204.2.

We remark that in mechanical terms, the relation (204.3) expresses that the *total energy*

$$E(t) \equiv \frac{1}{2}\dot{u}^2(t)) + \frac{1}{2}u^2(t), \tag{204.4}$$

is preserved ($= 1/2$) during the motion. The total energy at time t is the sum of the *kinetic energy* $\dot{u}^2(t)/2$, and the *potential energy* $u^2(t)/2$. The potential energy is the energy stored in the spring, which is equal to the *work* $W(u(t))$ to stretch the spring the distance $u(t)$:

$$W(u(t)) = \int_0^{u(t)} v\, dv = \frac{1}{2}u^2(t),$$

where we used the principle that to stretch the spring from v to $v+\Delta v$, the work is $v\Delta v$ since the spring force is v. At the extreme points with $\dot{u}(t) = 0$, the kinetic energy is zero and all energy occurs as potential energy, while all energy occurs as kinetic energy when the body passes the origin with $u(t) = 0$. During the motion of the body, the energy is thus periodically transferred from kinetic energy to potential energy an back again.

Going now back to (204.3), we thus see that the point $(\dot{u}(t), u(t)) \in \mathbb{R}^2$ moves on the unit circle and the *velocity* of the motion is given by $(\ddot{u}(t), \dot{u}(t))$, which we obtain by differentiating each coordinate function with respect to t. We will return to this issue in Chapter *Curves* below. Using the differential equation $\ddot{u} + u = 0$, we see that

$$(\ddot{u}(t), \dot{u}(t)) = (-u(t), \dot{u}(t)),$$

and conclude recalling (204.3) that the modulus of the velocity is equal to 1 for all t. We conclude that the point $(\dot{u}(t), u(t))$ moves around the unit circle with unit velocity and at time $t = 0$ the point is at position $(1, 0)$. But this directly connects with the usual geometrical definition of $(\cos(t), \sin(t))$ as the coordinates of a point on the unit circle at the angle t, see Fig. 204.2, so that we should have $(\dot{u}(t), u(t)) = (\cos(t), \sin(t))$. To

make this connection straight, we of course need to measure angles properly, and the proper measure is *radians* with 2π radians corresponding to 360 degrees. This is because the time for one revolution with speed 1 should be equal to 2π, that is the length of the circumference of the unit circle.

In fact, we can use the solution $\sin(t)$ of the initial value problem (204.2) to define the number π as the smallest positive root $\bar{t}$ of $\sin(t)$, corresponding to one half revolution with $u(\bar{t}) = 0$ and $\dot{u}(\bar{t}) = -1$.

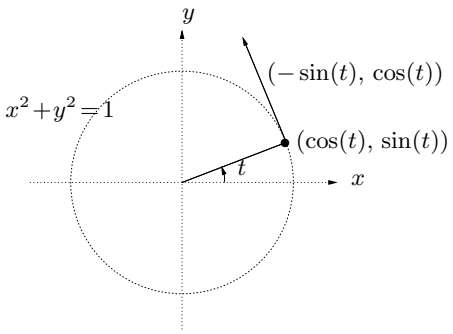


FIGURE 204.2. Energy conservation.

We may now conclude that the solution $u(t)$ of (204.2) satisfies $(\dot{u}(t), u(t)) = (\cos(t), \sin(t))$, so that in particular $u(t) = \sin(t)$ and $\frac{d}{dt}\sin(t) = \cos(t)$, where $(\cos(t), \sin(t))$ is defined geometrically as the point on the unit circle of angle t radians.

We can now turn the argument around, and simply define $\sin(t)$ as the solution $u(t)$ to (204.2) with $u_0 = 0$ and $u_1 = 1$, and then define $\cos(t) = \frac{d}{dt}\sin(t)$. Alternatively, we can define $\cos(t)$ as the solution of the problem

$$\ddot{v}(t) = -v(t) \quad \text{for } t > 0, \quad v(0) = 1,\ \dot{v}(0) = 0, \tag{204.5}$$

which we obtain by differentiation of (204.2) with respect to t and using the initial conditions for $\sin(t)$. Differentiating once more, we see that $\frac{d}{dt}\cos(t) = -\sin(t)$.

Both $\sin(t)$ and $\cos(t)$ will be *periodic with period* 2π, because the point $(\dot{u}(t), u(t))$ moves around the unit circle with velocity one and comes back the same point after a time period of 2π. As we said, we may in particular define π as the first value of $t > 0$ for which $\sin(t) = 0$, which corresponds the point $(\dot{u}, u) = (-1, 0)$, and 2π will then be time it takes for the point $(\dot{u}, u)$ to make one complete revolution starting at $(1, 0)$, moving to $(-1, 0)$ following the upper semi-circle and then returning to $(1, 0)$ following the lower semi-circle. The periodicity of $u(t)$ with period 2π is expressed as

$$u(t + 2n\pi) = u(t) \quad \text{for } t \in \mathbb{R},\ n = 0, \pm 1, \pm 2, ... \tag{204.6}$$

The energy conservation (204.3) translates into the most well known of all trigonometric formulas:

$$\sin^2(t) + \cos^2(t) = 1 \quad \text{for } t > 0. \tag{204.7}$$

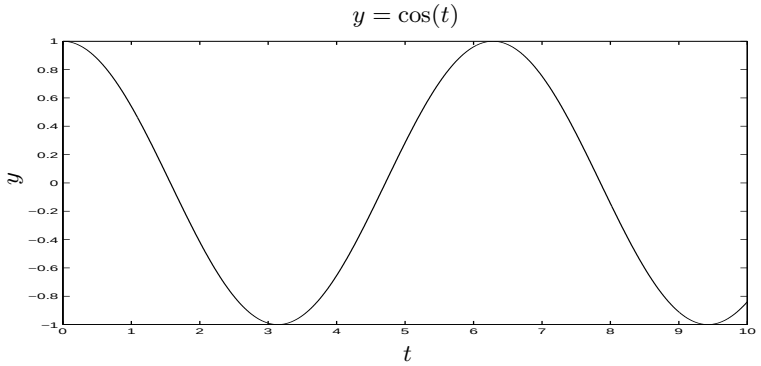


FIGURE 204.3. The function $\cos(t)$!

To compute the values of $\sin(t)$ and $\cos(t)$ for a given t, we may compute the solution to the corresponding defining differential initial value problem. We return to this topic below.

We summarize:

Theorem 204.1 *The initial value problem* $u''(x)+u(x) = 0$ *for* $x > 0$ *with* $u_0 = 0$ *and* $u_1 = 1$, *has a unique solution, which is denoted by* $\sin(x)$. *The initial value problem* $u''(x) + u(x) = 0$ *for* $x > 0$ *with* $u_0 = 1$ *and* $u_1 = 0$, *has a unique solution, which is denoted by* $\cos(x)$. *The functions* $\sin(x)$ *and* $\cos(x)$ *extend to* $x < 0$ *as solutions of* $u''(x) + u(x) = 0$ *and are periodic with period* 2π, *and* $\sin(\pi) = 0$, $\cos(\frac{\pi}{2}) = 0$. *We have* $\frac{d}{dx}\sin(x) = \cos(x)$ *and* $\frac{d}{dx}\cos(x) = -\sin(x)$. *Further* $\cos(-x) = \cos(x)$, $\cos(\pi - x) = -\cos(x)$, $sin(\pi - x) = \sin(x)$, $sin(-x) = -\sin(x)$, $\cos(x) = \sin(\frac{\pi}{2} - x)$, $\sin(x) = \cos(\frac{\pi}{2} - x)$, $\sin(\frac{\pi}{2} + x) = \cos(x)$, *and* $\cos(\frac{\pi}{2} + x) = -\sin(x)$.

204.2 Trigonometric Identities

Using the defining differential equation $u''(x) + u(x) = 0$, we can verify the following basic trigonometric identities for $x, y \in \mathbb{R}$:

$$\sin(x + y) = \sin(x)\cos(y) + \cos(x)\sin(y) \tag{204.8}$$

$$\sin(x - y) = \sin(x)\cos(y) - \cos(x)\sin(y) \tag{204.9}$$

$$\cos(x + y) = \cos(x)\cos(y) - \sin(x)\sin(y) \tag{204.10}$$

$$\cos(x - y) = \cos(x)\cos(y) + \sin(x)\sin(y). \tag{204.11}$$

For example, to prove (204.8), we note that both the right hand and left hand side satisfy the equation $u''(x) + u(x) = 0$, and the initial conditions $u(0) = \sin(y)$, $u'(0) = \cos(y)$, with y acting as a parameter, and thus are equal.

We note the particular special cases:

$$\sin(2x) = 2\sin(x)\cos(x) \tag{204.12}$$

$$\cos(2x) = \cos^2(x) - \sin^2(x) = 2\cos^2(x) - 1 = 1 - 2\sin^2(x). \tag{204.13}$$

Adding (204.8) and (204.9), we obtain

$$\sin(x+y) + \sin(x-y) = 2\sin(x)\cos(y).$$

Setting $\bar{x} = x + y$ and $\bar{y} = x - y$ we obtain the first of the following set of formulas, all proved similarly,

$$\sin(\bar{x}) + \sin(\bar{y}) = 2\sin(\frac{\bar{x}+\bar{y}}{2})\cos(\frac{\bar{x}-\bar{y}}{2}) \tag{204.14}$$

$$\sin(\bar{x}) - \sin(\bar{y}) = 2\cos(\frac{\bar{x}+\bar{y}}{2})\sin(\frac{\bar{x}-\bar{y}}{2}) \tag{204.15}$$

$$\cos(\bar{x}) + \cos(\bar{y}) = 2\cos(\frac{\bar{x}+\bar{y}}{2})\cos(\frac{\bar{x}-\bar{y}}{2}) \tag{204.16}$$

$$\cos(\bar{x}) - \cos(\bar{y}) = -2\sin(\frac{\bar{x}+\bar{y}}{2})\sin(\frac{\bar{x}-\bar{y}}{2}). \tag{204.17}$$

204.3 The Functions $\tan(x)$ and $\cot(x)$ and Their Derivatives

We define

$$\tan(x) = \frac{\sin(x)}{\cos(x)}, \quad \cot(x) = \frac{\cos(x)}{\sin(x)}, \tag{204.18}$$

for values of x such that the denominator is different from zero. We compute the derivatives

$$\frac{d}{dx}\tan(x) = \frac{\cos(x)\cos(x) - \sin(x)(-\sin(x))}{\cos^2(x)} = \frac{1}{\cos^2(x)}, \tag{204.19}$$

and similarly

$$\frac{d}{dx}\cot(x) = -\frac{1}{\sin^2(x)}. \tag{204.20}$$

Dividing (204.8) by (204.10), and dividing both numerator and denominator by $\cos(x)\cos(y)$, we obtain

$$\tan(x+y) = \frac{\tan(x) + \tan(y)}{1 - \tan(x)\tan(y)}, \tag{204.21}$$

and similarly,

$$\tan(x-y) = \frac{\tan(x) - \tan(y)}{1 + \tan(x)\tan(y)}. \tag{204.22}$$

204.4 Inverses of Trigonometric Functions

Inverses of the basic trigonometric functions $\sin(x)$, $\cos(x)$, $\tan(x)$ and $\cot(x)$, are useful in applications. We now introduce and give names to these inverses and derive their basic properties.

The function $f(x) = \sin(x)$ is strictly increasing from -1 to 1 on $[\frac{-\pi}{2}, \frac{\pi}{2}]$, because the derivative $f'(x) = \cos(x)$ is positive on $(\frac{-\pi}{2}, \frac{\pi}{2})$. Thus the function $y = f(x) = \sin(x)$ with $D(f) = [\frac{-\pi}{2}, \frac{\pi}{2}]$ and $R(f) = [-1, 1]$, therefore has an inverse $x = f^{-1}(y)$, which we denote by

$$x = f^{-1}(y) = \arcsin(y), \tag{204.23}$$

and $D(f^{-1}) = D(\arcsin) = [-1, 1]$ and $R(f^{-1}) = R(\arcsin) = [\frac{-\pi}{2}, \frac{\pi}{2}]$, see Fig. 204.4.

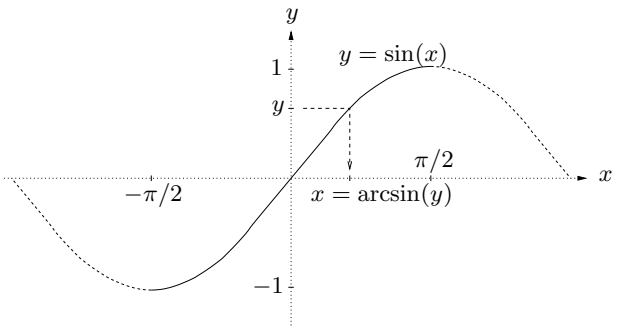


FIGURE 204.4. The function $x = \arcsin(y)$

We thus have

$$\sin(\arcsin(y)) = y \quad \arcsin(\sin(x)) = x \quad \text{for } x \in [\frac{-\pi}{2}, \frac{\pi}{2}], y \in [-1, 1]. \tag{204.24}$$

We next compute the derivative of $\arcsin(y)$ with respect to y:

$$\frac{d}{dy}\arcsin(y) = \frac{1}{\frac{d}{dx}\sin(x)} = \frac{1}{\cos(x)} = \frac{1}{\sqrt{1-\sin^2(x)}} = \frac{1}{\sqrt{1-y^2}}.$$

Similarly, the function $y = f(x) = \tan(x)$ is strictly increasing on $D(f) = (\frac{-\pi}{2}, \frac{\pi}{2})$ and $R(f) = \mathbb{R}$, and thus has an inverse, which we denote by

$$x = f^{-1}(y) = \arctan(y),$$

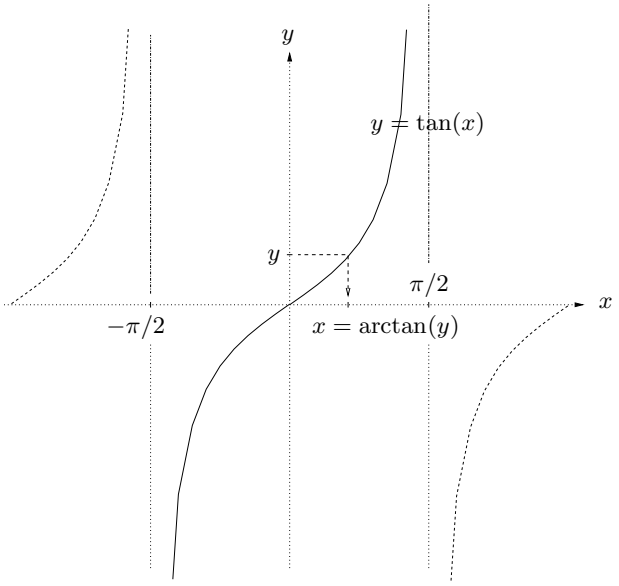


FIGURE 204.5. The function $x = \arctan(y)$

with $D(\arctan) = \mathbb{R}$ and $R(\arctan) = (\frac{-\pi}{2}, \frac{\pi}{2})$, see Fig. 204.5.

We compute the derivative of $\arctan(y)$:

$$\frac{d}{dy}\arctan(y) = \frac{1}{\frac{d}{dx}\tan(x)} = \cos^2(x)$$
$$= \frac{\cos^2(x)}{\cos^2(x) + \sin^2(x)} = \frac{1}{1 + \tan^2(x)} = \frac{1}{1 + y^2}.$$

We define similarly the inverse of $y = f(x) = \cos(x)$ with $D(f) = [0, \pi]$ and denote the inverse by $x = f^{-1}(y) = \arccos(y)$ with $D(\arccos) = [-1, 1]$ and $R(\arccos) = [0, \pi]$. We have

$$\frac{d}{dy}\arccos(y) = \frac{1}{\frac{d}{dx}\cos(x)} = -\frac{1}{\sin(x)} = -\frac{1}{\sqrt{1 - \cos^2(x)}} = -\frac{1}{\sqrt{1 - y^2}}.$$

Finally, we define the inverse of $y = f(x) = \cot(x)$ with $D(f) = (0, \pi)$ and denote the inverse by $x = f^{-1}(y) = \operatorname{arccot}(y)$ with $D(\operatorname{arccot}) = \mathbb{R}$ and $R(\operatorname{arccot}) = (0, \pi)$. We have

$$\frac{d}{dy}\operatorname{arccot}(y) = \frac{1}{\frac{d}{dx}\cot(x)} = -\sin^2(x) = -\frac{\sin^2(x)}{\cos^2(x) + \sin^2(x)}$$
$$= -\frac{1}{1 + \cot^2(x)} = -\frac{1}{1 + y^2}.$$

We summarize:

$$\begin{aligned}
\frac{d}{dx}\arcsin(x) &= \frac{1}{\sqrt{1-x^2}} \quad \text{for } x \in (-1,1)\\
\frac{d}{dx}\arccos(x) &= -\frac{1}{\sqrt{1-x^2}} \quad \text{for } x \in (-1,1)\\
\frac{d}{dx}\arctan(x) &= \frac{1}{1+x^2} \quad \text{for } x \in \mathbb{R}\\
\frac{d}{dx}\text{arccot}(x) &= -\frac{1}{1+x^2} \quad \text{for } x \in \mathbb{R}.
\end{aligned} \tag{204.25}$$

In other words,

$$\begin{aligned}
\arcsin(x) &= \int_0^x \frac{1}{\sqrt{1-y^2}}\,dy \quad \text{for } x \in (-1,1)\\
\arccos(x) &= \frac{\pi}{2} - \int_0^x \frac{1}{\sqrt{1-y^2}}\,dy \quad \text{for } x \in (-1,1)\\
\arctan(x) &= \int_0^x \frac{1}{1+y^2}\,dy \quad \text{for } x \in \mathbb{R}\\
\text{arccot}(x) &= \frac{\pi}{2} - \int_0^x \frac{1}{1+y^2}\,dy \quad \text{for } x \in \mathbb{R}.
\end{aligned} \tag{204.26}$$

We also note the following analog of (204.21) obtained by setting $x = \arctan(u)$ and $y = \arctan(v)$, so that $u = \tan(x)$ and $v = \tan(y)$, and assuming that $x + y \in (-\frac{\pi}{2}, \frac{\pi}{2})$:

$$\arctan(u) + \arctan(v) = \arctan(\frac{u+v}{1-uv}). \tag{204.27}$$

204.5 The Functions sinh(x) and cosh(x)

We define for $x \in \mathbb{R}$

$$\sinh(x) = \frac{e^x - e^{-x}}{2} \quad \text{and} \quad \cosh(x) = \frac{e^x + e^{-x}}{2}. \tag{204.28}$$

We note that

$$D\sinh(x) = \cosh(x) \quad \text{and} \quad D\cosh(x) = \sinh(x). \tag{204.29}$$

We have $y = f(x) = \sinh(x)$ is strictly increasing and thus has an inverse $x = f^{-1}(y) = \text{arcsinh}(y)$ with $D(\text{arcsinh}) = \mathbb{R}$ and $R(\text{arcsinh}) = \mathbb{R}$. Further, $y = f(x) = \cosh(x)$ is strictly increasing on $[0,\infty)$, and thus has an inverse $x = f^{-1}(y) = \text{arccosh}(y)$ with $D(\text{arccosh}) = [1,\infty)$ and $R(\text{arccosh}) = [0,\infty)$. We have

$$\frac{d}{dy}\text{arcsinh}(y) = \frac{1}{\sqrt{y^2+1}}, \quad \frac{d}{dy}\text{arccosh}(y) = \frac{1}{\sqrt{y^2-1}}. \tag{204.30}$$

204.6 The Hanging Chain

Consider a hanging chain fixed at $(-1,0)$ and $(1,0)$ in a coordinate system with the x-axis horizontal and y-axis vertical. Let us seek the curve $y = y(x)$ described by the chain. Let $(F_h(x), F_v(x))$ be the two components of the *force* in the chain at x. Vertical and horizontal equilibrium of the element of the chain between x and $x + \Delta x$ gives

$$F_h(x + \Delta x) = F_h(x), \quad F_v(x) + m\Delta s = F_v(x + \Delta x),$$

where $\Delta s \approx \sqrt{(\Delta x)^2 + (\Delta y)^2} \approx \sqrt{1 + (y'(x))^2}\Delta x$, and m is the weight of the chain per unit length. We conclude that $F_h(x) = F_h$ is constant, and

$$F_v'(x) = m\sqrt{1 + (y'(x))^2}.$$

Momentum equilibrium around the midpoint of the element of the chain between x and $x + \Delta x$, gives

$$F_h\Delta y = \frac{1}{2}F_v(x + \Delta x)\Delta x + \frac{1}{2}F_v(x)\Delta x \approx F_v(x)\Delta x,$$

which leads to

$$y'(x) = \frac{F_v(x)}{F_h}. \tag{204.31}$$

Assuming that $F_h = 1$, we are thus led to the differential equation

$$F_v'(x) = m\sqrt{1 + (F_v(x))^2}.$$

We can check by direct differentiation that this differential equation is satisfied if $F_v(x)$ solves the equation

$$\operatorname{arcsinh}(F_v(x)) = mx,$$

and we also have $F_v(0) = 0$. Therefore

$$F_v(x) = \sinh(mx),$$

and thus by (204.31),

$$y(x) = \frac{1}{m}\cosh(mx) + c$$

with the constant c to be chosen so that $y(\pm 1) = 0$. We thus obtain the following solution

$$y(x) = \frac{1}{m}(\cosh(mx) - \cosh(m)). \tag{204.32}$$

The curve $y(x) = \cosh(mx)+c$ with m and c constants, is called the *hanging chain curve*, or the *catenaria*.

204.7 Comparing $u'' + k^2u(x) = 0$ and $u'' - k^2u(x) = 0$

We summarize some experience from above. The solutions of the equation $u'' + k^2u(x) = 0$ are linear combinations of $\sin(kx)$ and $\cos(kx)$. The solutions of $u'' - k^2u(x) = 0$ are linear combinations of $\sinh(kx)$ and $\cosh(kx)$.

Chapter 204 Problems

204.1. Show that the solution of $\ddot{u}(t) + u(t) = 0$ for $t > 0$ with $u(0) = \sin(\alpha)$ and $u'(0) = \cos(\alpha)$ is given by $u(t) = \cos(t)\sin(\alpha) + \sin(t)\cos(\alpha) = \sin(t + \alpha)$.

204.2. Show that the solution of $\ddot{u}(t)+u(t) = 0$ for $t > 0$ with $u(0) = r\cos(\alpha)$ and $u'(0) = r\sin(\alpha)$ is given by $u(t) = r(\cos(t)\cos(\alpha) + \sin(t)\sin(\alpha)) = r\cos(t - \alpha)$.

204.3. Show that the solution to $\ddot{u}(t) + ku(t) = 0$ for $t > 0$ with $u(0) = r\cos(\alpha)$ and $u'(0) = r\sin(\alpha)$, where k is a given positive constant, is given by $r\cos(\sqrt{k}(t - \alpha))$. Give a mechanical interpretation of this model.

204.4. Show that the function $\sin(nx)$ solves the boundary value problem $u''(x)+n^2u(x) = 0$ for $0 < x < \pi$, $u(0) = u(\pi) = 0$.

204.5. Solve $u'(x) = \sin(x), \quad x > \pi/4$, $u(\pi/4) = 2/3$.

204.6. Show that (a) $\sin(x) < x$ for $x > 0$, (b) $x < \tan(x)$ for $0 < x < \frac{\pi}{2}$.

204.7. Show that $\lim_{x\to 0} \frac{\sin(x)}{x} = 1$.

204.8. Show the following relations from the definition, i.e. from the differential equation defining $\sin(x)$ and $\cos(x)$: (a) $\sin(-x) = -\sin(x)$, (b) $\cos(-x) = \cos(x)$, (c) $\sin(\pi - x) = \sin(x)$, (d) $\cos(\pi - x) = -\cos(x)$, (e) $\sin(\pi/2 - x) = \cos(x)$, (f) $\cos(\pi/2 - x) = \sin(x)$.

204.9. Prove the product formulas show that

$$\sin(x)\sin(y) = \frac{1}{2}\left(\cos(x - y) - \cos(x + y)\right),$$
$$\cos(x)\cos(y) = \frac{1}{2}\left(\cos(x - y) + \cos(x + y)\right),$$
$$\sin(x)\cos(y) = \frac{1}{2}\left(\sin(x - y) + \sin(x + y)\right).$$

204.10. Compute the following integrals by integrating by parts:
(a) $\int_0^1 x^3\sin(x)dx$, (b) $\int_0^1 \exp(x)\sin(x)dx$, (c) $\int_0^1 x^2\cos(x)dx$.

204.11. Determine Taylor's formula for $\arctan(x)$ at $x = 0$ and use your result to calculate approximations of π. Hint: $\arctan(1) = \frac{\pi}{4}$.

204.12. Show that $\arctan(1) = \arctan(1/2) + \arctan(1/3)$. Try to find other rational numbers a and b such that $\arctan(1) = \arctan(a) + \arctan(b)$. In particular seek to find a and b as small as possible.

204.13. Combine your results from the previous two exercises to construct a better algorithm for computing π. Even more efficient methods may be obtained using the identity $\pi/4 = 4\arctan(1/5) - \arctan(1/239)$. Compare the two algorithms and explain why the second is more efficient.

204.14. Show that: (a) $\arcsin(-x) = -\arcsin(x)$, (b) $\arccos(-x) = \pi - \arccos(x)$, (c) $\arctan(-x) = -\arctan(x)$, (d) $\operatorname{arccot}(-x) = \pi - \operatorname{arccot}(x)$, (e) $\arcsin(x) + \arccos(x) = \pi/2$, (f) $\arctan(x) + \operatorname{arccot}(x) = \pi/2$.

204.15. Calculate analytically: (a) $\arctan(\sqrt{2}-1)$, (b) $\arctan(1/8)+\arctan(7/9)$, (c) $\arcsin(1/7) + \arcsin(11/4)$, (d) $\tan(\arcsin(3/5)/2)$, (e) $\sin(2\arcsin(0.8))$, (f) $\arctan(2) + \arcsin(3/\sqrt{10})$.

204.16. Solve the equation: (a) $\arccos(2x) = \arctan(x)$, (b) $\arcsin(\cos(x)) = x\sqrt{3}$.

204.17. Calculate the derivative, if possible, of

(a) $\arctan(\sqrt{x} - x^5)$, (b) $\arcsin(1/x^2)\arcsin(x^2)$,

(c) $\tan(\arcsin(x^2))$, (d) $1/\arctan(\sqrt{x})$.

204.18. Compute numerically for different values of x, (a) $\arcsin(x)$, (b) $\arccos(x)$, (c) $\arctan(x)$, (d) $\operatorname{arccot}(x)$.

204.19. Prove (204.30).

204.20. Verify that $\cosh^2(x) - \sinh^2(x) = 1$.

204.21. (a) Find the inverse $x = \operatorname{arcsinh}(y)$ of $y = sinh(x) = \frac{1}{2}(e^x - e^{-x})$ by solving for x in terms of y. Hint: Multiply by e^x and solve for $z = e^x$. Then take logarithms. (b) Find a similar formula for $\operatorname{arccosh}(y)$.

204.22. Compute analytically the area of a disc of radius 1 by computing the integral

$$\int_{-1}^{1} \sqrt{1 - x^2}\, dx.$$

How do you handle the fact that $\sqrt{1-x^2}$ is not Lipschitz continuous on $[-1, 1]$? Hint: Use the substitution $x = \sin(y)$ and the fact the $\cos^2(y) = \frac{1}{2}(1 + \cos(2y))$.

205

The Functions $\exp(z)$, $\log(z)$, $\sin(z)$ and $\cos(z)$ for $z \in \mathbb{C}$

The shortest path between two truths in the real domain passes through the complex domain. (Hadamard 1865-1963)

205.1 Introduction

In this chapter we extend some of the elementary functions to complex arguments. We recall that we can write a complex number z in the form $z = |z|(\cos(\theta) + i\sin(\theta))$ with $\theta = \arg z$ the argument of z, and $0 \leq \theta = \text{Arg } z < 2\pi$ the principal argument of z.

205.2 Definition of $\exp(z)$

We define, writing $z = x + iy$ with $x, y \in \mathbb{R}$,

$$\exp(z) = e^z = e^x(\cos(y) + i\sin(y)), \tag{205.1}$$

which extends the definition of e^z with $z \in \mathbb{R}$ to $z \in \mathbb{C}$. We note that in particular for $y \in \mathbb{R}$,

$$e^{iy} = \cos(y) + i\sin(y), \tag{205.2}$$

which is also referred to as *Euler's formula*. We note that

$$\sin(y) = \frac{e^{iy} - e^{-iy}}{2i}, \quad \cos(y) = \frac{e^{iy} + e^{-iy}}{2}, \quad \text{for } y \in \mathbb{R}, \tag{205.3}$$

and

$$|e^{iy}| = 1 \quad \text{for } y \in \mathbb{R}. \tag{205.4}$$

We can now express a complex number $z = r(\cos(\theta) + i\sin(\theta))$ in the form

$$z = re^{i\theta} \tag{205.5}$$

with $\theta = \arg z$ and $r = |z|$.

One can prove (using the basic trigonometric formulas) that $\exp(z)$ satisfies the usual law for exponentials so that in particular for $z, \zeta \in \mathbb{C}$,

$$e^z e^\zeta = e^{z+\zeta}. \tag{205.6}$$

In particular, the rule for multiplication of two complex numbers $z = |z|e^{i\theta}$ and $\zeta = |\zeta|e^{i\varphi}$ can be expressed as follows:

$$z\zeta = |z|e^{i\theta}|\zeta|e^{i\varphi} = |z||\zeta|e^{i(\theta+\varphi)}. \tag{205.7}$$

205.3 Definition of sin(z) and cos(z)

We define for $z \in \mathbb{C}$

$$\sin(z) = \frac{e^{iz} - e^{-iz}}{2i}, \quad \cos(z) = \frac{e^{iz} + e^{-iz}}{2}, \tag{205.8}$$

which extends (205.3) to $\mathbb{C}$.

205.4 de Moivres Formula

We have for $\theta \in \mathbb{R}$ and n an integer

$$(e^{i\theta})^n = e^{in\theta},$$

that is,

$$(\cos(\theta) + i\sin(\theta))^n = \cos(n\theta) + i\sin(n\theta), \tag{205.9}$$

which is referred to as *de Moivres formula*. In particular,

$$(\cos(\theta) + i\sin(\theta))^2 = \cos(2\theta) + i\sin(2\theta),$$

from which follows separating into real and complex parts

$$\cos(2\theta) = \cos^2(\theta) - \sin^2(\theta), \quad \sin(2\theta) = 2\cos(\theta)\sin(\theta).$$

Using de Moivres formula gives a quick way of deriving some of the basic trigonometric formulas (in case one has forgotten these formulas).

205.5 Definition of log(z)

We have defined above log(x) for $x > 0$ and we now pose the problem of defining log(z) for $z \in \mathbb{C}$. We recall that $w = \log(x)$ can be viewed as the unique solution to the equation $e^w = x$, where $x > 0$. We consider therefore the equation

$$e^w = z,$$

with $z = |z|(\cos(\theta) + i\sin(\theta)) \in \mathbb{C}$ being given assuming $z \neq 0$, and we seek $w = \text{Re } w + i\text{Im } w \in \mathbb{C}$, with the intention to call a solution $w = \log(z)$. Here Re w and Im w denote the real and imaginary parts of w. Equating the modulus of both sides of the equation $e^w = z$, we get

$$e^{\text{Re } w} = |z|,$$

and thus

$$\text{Re } w = \log(|z|).$$

Further, equating the argument of both sides, we get

$$\text{Im } w = \theta = \arg z,$$

and thus

$$w = \log(|z|) + i \arg z.$$

We are thus led to define

$$\log(z) = \log(|z|) + i \arg z, \tag{205.10}$$

which extends the definition of the natural logarithm from the positive real numbers to non-zero complex numbers. We see that the imaginary part log(z) is not uniquely defined up to multiples of 2π, since arg z is not, and thus log(z) is *multi-valued*: the imaginary part of log(z) is not uniquely defined up to multiples of 2π. Choosing $\theta = \text{Arg } z$ with $0 \leq \text{Arg } z < 2\pi$, we obtain the *principal branch* of log(z) denoted by

$$\text{Log}(z) = \log(|z|) + i\text{Arg } z.$$

We see that if we let arg z increase from 0 beyond 2π, the function Log (z) will be discontinuous at Im $z = 2\pi$. We thus have to remember that the imaginary part of log(z) is not uniquely defined.

Chapter 205 Problems

205.1. Describe in geometrical terms the mappings $f : \mathbb{C} \to \mathbb{C}$ given by (a) $f(z) = \exp(z)$, (b) $f(z) = \text{Log}(z)$, (c) $\sin(z)$.

206
Techniques of Integration

A poor head, having subsidiary advantages,.... can beat the best, just as a child can draw a line with a ruler better than the greatest master by hand. (Leibniz)

206.1 Introduction

It is not generally possible to find an explicit formula for a primitive function of a given arbitrary function in terms of known *elementary functions*, by which we mean the polynomials, rational functions, root functions, exponentials and trigonometric functions along with their inverses and combinations. It is not even true that the primitive function of an elementary function is another elementary function. A famous example is given by the function $f(x) = \exp(-x^2)$, whose primitive function $F(x)$ (with $F(0) = 0$), which exists by the Fundamental Theorem, is known *not* to be an elementary function (by a tricky proof by contradiction). To compute values of $F(x) = \int_0^x \exp(y)\,dy$ for different values of x we therefore have to use numerical quadrature just as in the case of the logarithm. Of course we can give $F(x)$ a *name*, for example we may agree to call it the *error function* $F(x) = \text{erf}(x)$ and add it to our list of known functions that we can use. Nevertheless there will be other functions (such as $\frac{\sin(x)}{x}$) whose primitive function cannot be expressed in the known functions.

The question of how to handle such functions (including also $\log(x)$, $\exp(x)$, $\sin(x)$...) of course arises: should we pre-compute long tables of values of these functions and print them in thick books or store them in

the computer, or should we compute each required value from scratch using numerical quadrature? The first option was favored in earlier times when computing power was sparse, and the second one is favored today (even in the pocket calculator).

Despite the impossibility to reach generality, it is it is possible (and useful) to compute primitive functions analytically in certain cases, and in this chapter, we collect some tricks that have proved useful for doing this. The tricks we present are basically various clever substitutions together with integration by parts. We have no ambition to be encyclopedic. We refer to Mathematics Handbook for Science and Engineering for further development.

We start with rational functions, and then proceed to various combinations of polynomials, logarithms, exponentials and trigonometric functions.

206.2 Rational Functions: The Simple Cases

Integration of rational functions depends on three basic formulas

$$\int_{x_0}^{x} \frac{1}{s-c}\,ds = \log|x-c| - \log|x_0-c|, \quad c \neq 0 \tag{206.1}$$

$$\int_{x_0}^{x} \frac{s-a}{(s-a)^2+b^2}\,dx = \frac{1}{2}\log((x-a)^2+b^2) - \frac{1}{2}\log((x_0-a)^2+b^2) \tag{206.2}$$

and

$$\int_{x_0}^{x} \frac{1}{(s-a)^2+b^2}\,ds = [\frac{1}{b}\arctan(\frac{x-a}{b})] - [\frac{1}{b}\arctan(\frac{x_0-a}{b})], \quad b \neq 0. \tag{206.3}$$

These formulas can be verified by differentiation. Using the formulas can be straightforward as in

Example 206.1.

$$\int_6^8 \frac{ds}{s-4} = \log 4 - \log 2 = \log 2.$$

Or more complicated as in

Example 206.2.

$$\begin{aligned}\int_2^4 \frac{ds}{2(s-2)^2+6} &= \frac{1}{2}\int_2^4 \frac{ds}{(s-2)^2+3} \\ &= \frac{1}{2}\int_2^4 \frac{ds}{(s-2)^2+(\sqrt{3})^2} \\ &= \frac{1}{2}\left(\frac{1}{\sqrt{3}}\arctan\left(\frac{4-2}{\sqrt{3}}\right) - \frac{1}{\sqrt{3}}\arctan\left(\frac{2-2}{\sqrt{3}}\right)\right).\end{aligned}$$

Of course we may combine these formulas with substitution:

EXAMPLE 206.3.

$$\int_0^x \frac{\cos(s)\,ds}{\sin(s)+2} = \int_0^{\sin(x)} \frac{du}{u+2} = \log|\sin(x)+2| - \log 2.$$

Using (206.2) and (206.3) may require *completing the square*, as we now show in

EXAMPLE 206.4. For example, consider

$$\int_0^3 \frac{ds}{s^2-2s+5}.$$

We want to get s^2-2s+5 into the form $(s-a)^2+b^2$ if possible. We set

$$(s-a)^2+b^2 = s^2-2as+a^2+b^2 = s^2-2s+5.$$

Equating the coefficients of s on both sides gives $a=1$. Equating the constant terms on both sides gives $b^2 = 5-1 = 4$ and therefore we may take $b=2$. After a little practice with completing the square, we can often argue directly, as

$$s^2-2s+5 = s^2-2s+1^2-1^2+5 = (s-1)^2+2^2.$$

Returning to the integral, we have

$$\begin{aligned}\int_0^3 \frac{ds}{s^2-2s+5} &= \int_0^3 \frac{ds}{(s-1)^2+2^2} \\ &= \frac{1}{2}\arctan\left(\frac{3-2}{2}\right) - \frac{1}{2}\arctan\left(\frac{0-2}{2}\right).\end{aligned}$$

206.3 Rational Functions: Partial Fractions

We now investigate a systematic method for computing integrals of *rational* functions $f(x)$, i.e. functions of the form $f(x) = p(x)/q(x)$, where $p(x)$ and $q(x)$ are polynomials. The method is based manipulating the integrand so that the basic formulas (206.1)–(206.3) can be used. The manipulation is based on the observation that it is possible to write a complicated rational function as a sum of relatively simple rational functions.

EXAMPLE 206.5. Consider the integral

$$\int_4^5 \frac{s^2+s-2}{s^3-3s^2+s-3}\,ds.$$

The integrand can be expanded

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

which we can verify by adding the two fractions on the right after computing a common denominator,

$$\begin{aligned}\frac{1}{s^2+1} + \frac{1}{s-3} &= \frac{s-3}{s-3} \times \frac{1}{s^2+1} + \frac{s^2+1}{s^2+1} \times \frac{1}{s-3} \\ &= \frac{s-3+s^2+1}{(s^2+1)(s-3)} = \frac{s^2+s-2}{s^3-3s^2+s-3}.\end{aligned}$$

Therefore we can integrate

$$\begin{aligned}\int_4^5 \frac{s^2+s-2}{s^3-3s^2+s-3}\,ds &= \int_4^5 \frac{1}{s^2+1}\,ds + \int_4^5 \frac{1}{s-3}\,ds \\ &= (\arctan(5) - \arctan(4)) + (\log(5-3) - \log(4-3)).\end{aligned}$$

The general technique of *partial fractions* is based on a systematic method for writing a rational function as a sum of simple rational functions that can be integrated with the basic formulas (206.1)–(206.3). The method is analogous to "reversing" the addition of rational functions by finding a common denominator.

Applying the technique of partial fractions to a general rational function has several steps, which we explain in "reverse" order. So we begin by assuming that the numerator $p(x)$ of the rational function $p(x)/q(x)$ has smaller degree than the denominator $q(x)$, i.e. $\deg p(x) < \deg q(x)$, and that $q(x)$ has the form

$$\frac{p(x)}{q(x)} = \frac{p(x)}{k(x-c_1)\cdots(x-c_n)((x-a_1)^2+b_1^2)\cdots((x-a_m)^2+b_m^2)}, \tag{206.4}$$

where k is a number, the c_i are the real roots of $q(x)$, and the second degree factors $(x-a_j)^2+b_j^2$ correspond to the complex roots $a_j \pm ib_j$ of $q(x)$ that necessarily come in pairs of complex conjugates. We call polynomials of the form $(x-a_j)^2+b_j^2$ *irreducible* because we cannot factor them as a product of linear polynomials with *real* coefficients.

In the first instance, we assume that the zeroes $\{c_i\}$ and $\{a_j \pm ib_j\}$ are distinct. In this case, we rewrite $p(x)/q(x)$ as the sum of partial fractions

$$\frac{p(x)}{q(x)} = \frac{C_1}{x-c_1} + \cdots + \frac{C_n}{x-c_n} + \frac{A_1(x-a_1)+B_1}{(x-a_1)^2+b_1^2} + \cdots + \frac{A_m(x-a_m)+B_m}{(x-a_m)^2+b_m^2}, \tag{206.5}$$

for some constants C_i, $1 \le j \le n$, and A_j, B_j, $1 \le j \le m$ that we have to determine. The motivation to rewrite $p(x)/q(x)$ in this way is that we can

then compute an integral of $p(x)/q(x)$ by applying the formulas (206.1)–(206.3) to integrate the individual terms on the right-hand side of (206.5) as in the example above.

EXAMPLE 206.6. For $p(x) = q(x) = (x-1)/(x^2-x-2)$ with $q(x) = (x-2)(x+1)$ we have

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

and thus

$$\begin{aligned}\int_{x_0}^{x} \frac{s-1}{s^2-s-2}\,ds &= \frac{1}{3}\int_{x_0}^{x} \frac{1}{s-2}\,ds + \frac{2}{3}\int_{x_0}^{x} \frac{1}{s+1}\,ds \\ &= \frac{1}{3}[\log(s-2)]_{s=x_0}^{s=x} + \frac{2}{3}[\log(s+1)]_{s=x_0}^{s=x}.\end{aligned}$$

The rationale for the expansion (206.5) is simply that if we ask for the most general sum of rational functions with denominators of degrees 1 and 2 that can yield $p(x)/q(x)$, where $q(x)$ is the common denominator for the sum, then we get precisely the right- hand side of (206.5). In particular if the terms on the right had numerators of any higher degree, then $p(x)$ would have to have degree greater than $q(x)$.

The constants C_i, A_j and B_j in (206.5) can be found by rewriting the right-hand side of (206.5) with a common denominator.

EXAMPLE 206.7. In the last example with $q(x) = (x-2)(x+1)$, we find that

$$\frac{C_1}{x-2} + \frac{C_2}{x+1} = \frac{C_1(x+1) + C_2(x-2)}{(x-2)(x+1)} = \frac{(C_1+C_2)x + (C_1-2C_2)}{(x-2)(x+1)},$$

which equals

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

if and only if

$$C_1 + C_2 = 1 \quad \text{and} \quad C_1 - 2C_2 = -1,$$

that is if $C_1 = 1/3$ and $C_2 = 2/3$.

Since it is cumbersome to compute the constants by dealing with the fractions, we usually rewrite the problem by multiplying both sides of (206.5) by the common denominator.

EXAMPLE 206.8. We multiply both sides of

$$\frac{x-1}{(x-2)(x+1)} = \frac{C_1}{x-2} + \frac{C_2}{x+1}$$

by $(x-2)(x+1)$ to get

$$x-1 = C_1(x+1)C_2(x-2) = (C_1+C_2)x + (C_1-2C_2).$$

Equating coefficients, we find $C_1+C_2=1$ and $C_1-2C_2=-1$, which yields $C_1=1/3$ and $C_2=2/3$.

EXAMPLE 206.9. To integrate $f(x) = (5x^2-3x+6)/((x-2)((x+1)^2+2^2))$, we begin by writing the partial fraction expansion

$$\frac{5x^2-3x+6}{(x-2)((x+1)^2+2^2)} = \frac{C}{x-2} + \frac{A(x+1)+B}{(x+1)^2+2^2}.$$

To determine the constants, we multiply both sides by $(x-2)((x+1)^2+2^2))$ to obtain

$$\begin{aligned} 5x^2-3x+6 &= C((x+1)^2+2^2) + (A(x+1)+B)(x-2) \\ &= (C+A)x^2 + (2C-2A+B)x + (4C-2A-2B). \end{aligned}$$

Equating coefficients, we find that $C+A=0$, $2C-2A+B=1$ and $5C-2A-2B=0$, that is $C=2$, $A=3$ and $B=-1$. Therefore we find that

$$\begin{aligned} &\int_{x_0}^{x} \frac{5s^2-3s+6}{(s-2)((s+1)^2+2^2))}\,ds \\ &\quad = 2\int_{x_0}^{x} \frac{1}{s-2}\,ds + \int_{x_0}^{x} \frac{3(s+1)-1}{(s+1)^2+2^2}\,ds \\ &\quad = 2\int_{x_0}^{x} \frac{1}{s-2}\,ds + 3\int_{x_0}^{x} \frac{s+1}{(s+1)^2+2^2}\,ds - \int_{x_0}^{x} \frac{1}{(s+1)^2+2^2}\,ds \\ &\quad = 2\big(\log|x-2| - \log|x_0-2|\big) \\ &\qquad\qquad + \frac{3}{2}\big(\log((x+1)^2+4) - \log((x_0+1)^2+4)\big) \\ &\qquad\qquad - \frac{1}{2}\big(\arctan(\frac{x+1}{2}) - \arctan(\frac{x_0+1}{2})\big). \end{aligned}$$

In the case that some of the factors in the factorization of the denominator (206.4) are repeated, i.e. some of the roots have multiplicity greater than one, then we have to modify the partial sum expansion (206.5). We do not write out a general case because it is a mess and nearly unreadable, we just note that the principle for determining the correct partial fractions is always to write down the most general sum that can give the indicated common denominator.

EXAMPLE 206.10. The general partial fraction expansion of $f(x) = x^2/((x-2)(x+1)^2)$ has the form

$$\frac{x^2}{(x-2)(x+1)^2} = \frac{C_1}{x-2} + \frac{C_{2,1}}{x+1} + \frac{C_{2,2}}{(x+1)^2},$$

for constants C_1, $C_{2,1}$ and $C_{2,2}$ because all of the terms on the right-hand will yield the common denominator $(x-2)(x+1)^2$. Multiplying both sides by the common denominator and equating coefficients as usual, we find that $C_1 = 4/9$, $C_{2,1} = 5/9$ and $C_{2,2} = -3/9$.

In general if $q(x)$ has the multiple factor $(x-c_i)^L$ the term $\frac{C_i}{x-c_i}$ in the partial fraction expansion (206.5) should be replaced by the *sum* of fractions $\sum_{l=1}^{l=L} \frac{C_{i,l}}{(x-c)^l}$. There is a corresponding procedure for multiple factors of the form $((x-a)^2+b^2)^L$.

We have discussed how to integrate rational functions $p(x)/q(x)$ where $\deg p < \deg q$ and q is factored into a product of linear and irreducible quadratic polynomials. Now we discuss removing these restrictions. First we deal with the factorization of the denominator $q(x)$. The Fundamental Theorem of Algebra says that a polynomial q of degree n with real coefficients has exactly n roots and hence it can be factored into a product of n linear polynomials with possibly complex coefficients. However, because the polynomial q has real coefficients, the complex roots always come in complex conjugate pairs, i.e. if r is a root of q then so is $\bar{r}$. This means that there are an even number of linear factors of q corresponding to complex roots and furthermore we can combine the factors corresponding to conjugate roots to get quadratic factors with real coefficients. For example, $(x-3+i)(x-3-i) = (x-3)^2+1$. Therefore every polynomial $q(x)$ can theoretically be factored into a product $k(x-c_1)\cdots(x-c_n)((x-a_1)^2+b_1^2)\cdots((x-a_m)^2+b_m^2)$.

However, we caution that this theoretical result does not carry over in practice to situations in which the degree of q is large. To determine the factorization of q, we must determine the roots of q. In the problems and examples, we stick to cases in which the roots are simple, relatively small integers. But in general we know that the roots can be any kind of algebraic number which we can only approximate. Unfortunately it turns out that it is extremely difficult to determine the roots of a polynomial of high degree, even using Newton's method. So the method of partial fractions is used only for low degree polynomials in practice, though it is a very useful theoretical tool.

Finally we remove the restriction that $\deg p < \deg q$. When the degree of the numerator polynomial $p(x)$ is $\geq$ the degree of the denominator polynomial $q(x)$, we first use polynomial division to rewrite $f(x)$ as the sum of a polynomial $s(x)$ and a rational function $\frac{r(x)}{q(x)}$ for which the degree of the numerator $r(x)$ is *less* than the degree of the denominator $q(x)$.

EXAMPLE 206.11. For $f(x) = (x^3-x)/(x^2+x+1)$, we divide to get $f(x) = x-1+(1-x)/(x^2+x+1)$, so that

$$\int_0^{\bar{x}} \frac{x^3}{x^2+x+1}\,dx = [\frac{1}{2}x^2 - x]_{x=0}^{x=\bar{x}} + \int_0^{\bar{x}} \frac{1-x}{x^2+x+1}\,dx.$$

206.4 Products of Polynomial and Trigonometric or Exponential Functions

To integrate the product of a polynomial and a trigonometric or exponential function, we use integration by parts repeatedly to reduce the polynomial to an constant.

EXAMPLE 206.12. To compute a primitive function of $x\cos(x)$, we integrate by parts once

$$\int_0^x y\cos(y)\,dy = [y\sin(y)]_{y=0}^{y=x} - \int_0^x \sin(y)\,dy = x\sin(x) + \cos(x) + 1.$$

To handle higher order polynomials, we use integration by parts several times.

EXAMPLE 206.13. We have

$$\begin{aligned}\int_0^x s^2e^s\,ds &= s^2(e^s)_{s=0}^{s=x} - 2\int_0^x se^s\,ds\\ &= [s^2e^s]_{s=0}^{s=x} - 2\big([se^s]_{s=0}^{s=x} - \int_0^x e^s\,ds\big)\\ &= [s^2e^s)_{s=0}^{s=x} - 2\big([se^s]_{s=0}^{s=x} - [e^s]_{s=0}^{s=x}\big)\\ &= x^2e^x - 2xe^x + 2e^x - 2.\end{aligned}$$

206.5 Combinations of Trigonometric and Root Functions

To compute a primitive function of $\sin(\sqrt{y})$ for $x > 0$, we set $y = t^2$ and obtain by using partial integration

$$\begin{aligned}\int_0^x \sin(\sqrt{y})\,dy &= \int_0^{\sqrt{x}} 2t\sin(t)\,dt = [-2t\cos(t)]_{t=0}^{t=\sqrt{x}} + 2\int_0^{\sqrt{x}}\cos(t)\,dt\\ &= -2\sqrt{x}\cos(\sqrt{x}) + 2\sin(\sqrt{x}).\end{aligned}$$

206.6 Products of Exponential and Trigonometric Functions

To compute a primitive function of $e^y\sin(y)$, we use repeated integration by parts as follows

$$\begin{aligned}\int_0^x e^y\sin(y)\,dy &= [e^y\sin(y)]_{y=0}^{y=x} - \int_0^x e^y\cos(y)\,dy\\ &= e^x\sin(x) - [e^y\cos(y)]_{y=0}^{y=x} - \int_0^x e^y\sin(y)\,dy,\end{aligned}$$

which shows that

$$\int_0^x e^y \sin(y)\,dy = \frac{1}{2}(e^x \sin(x) - e^x \cos(x) + 1)$$

206.7 Products of Polynomials and Logarithm Functions

To compute a primitive function of $x^2 \log(x)$, we integrate by parts:

$$\int_1^x y^2 \log(y)\,dy = [\frac{y^3}{3}\log(y)]_{y=1}^{y=x} - \int_1^x \frac{y^3}{3}\frac{1}{y}\,dy = \frac{x^3}{3}\log(x) - \frac{x^3}{9} + \frac{1}{9}.$$

Chapter 206 Problems

206.1. Compute

(a) $\int_0^x t\sin(2t)\,dt$ (b) $\int_0^x t^2\cos(t)\,dt$ (c) $int_0^x t\exp(-2t)\,dt$. Hint: Integrate by parts.

206.2. Compute (a) $\int_1^x y\log(y)\,dy$ (b) $\int_1^x \log(y)\,dy$ (c) $\int_0^x \arctan(t)\,dt$ (d) $\int_0^x$
$exp(-t)\cos(2t)\,dt$. Hint: Integrate by parts.

206.3. Compute using the formula $\int_0^x \frac{g'(y)}{g(y)}\,dy = \log(g(x)) - \log(g(0))$ the following integrals. (a) $\int_0^x \frac{y}{y^2+1}\,dy$ (b) $\int_0^x \frac{e^t}{e^t+1}\,dt$.

206.4. Compute by a suitable change of variable

(a) $\int_0^x y\exp(y^2)\,dy$ (b) $\int_0^x y\sqrt{y-1}\,dy$ (c) $\int_0^x \sin(t)\cos^2(t)\,dt$.

206.5. Compute (a) $\int_0^x \frac{dy}{y^2-y-2}\,dy$ (b) $\int_0^x \frac{y^3}{y^2+2y-3}\,dy$ (c) $\int_0^x \frac{dy}{y^2+2y+5}\,dy$ (d) $\int_0^x \frac{x-x^2}{(y-1)(y^2+2y+5)}\,dy$ (e) $\int_0^x \frac{x^4}{(x-1)(x^2+x-6)}\,dy$.

206.6. Recalling that a function is called *even* if $f(-x) = f(x)$ and *odd* if $f(-x) = -f(x)$ for all x, (a) give examples of even and odd functions (b) sketch their graphs, and (c) show that

$$\int_{-a}^a f(x)\,dx = 2\int_0^a f(x)\,dx \text{ if f is even,} \quad \int_{-a}^a f(x)\,dx = 0 \text{ if f is odd.} \quad (206.6)$$

206.7. Compute (a) $\int_{-\pi}^{\pi} |x|\cos(x)\,dx$ (b) $\int_{-\pi}^{\pi} \sin^2(x)\,dx$ (c) $\int_{-\pi}^{\pi} x\sin^2(x)\,dx$ (d) $\int_{-\pi}^{\pi} \arctan(x+3x^3)\,dx$.

207

Solving Differential Equations Using the Exponential

...he climbed a little further... and further...and then just a little further. (Winnie-the-Pooh)

207.1 Introduction

The exponential function plays a fundamental role in modeling and analysis because of its basic properties. In particular it can be used to solve a variety of differential equations analytically as we show in this chapter. We start with generalizations of the initial value problem (203.2) from Chapter *The exponential function*:

$$u'(x) = \lambda u(x) \quad \text{for } x > a, \quad u(a) = u_a, \tag{207.1}$$

where $\lambda \in \mathbb{R}$ is a constant, with solution

$$u(x) = \exp(\lambda(x-a))u_a \quad \text{for } x \geq a. \tag{207.2}$$

Analytic solutions formulas may give very important information and help the intuitive understanding of different aspects of a mathematical model, and should therefore be kept as valuable gems in the scientist and engineer's tool-bag. However, useful analytical formulas are relatively sparse and must be complemented by numerical solutions techniques. In the Chapter *The General Initial Value Problem* we extend the constructive numerical method for solving (207.1) to construct solutions of general initial value problems for systems of differential equations, capable of modeling a very

large variety of phenomena. We can thus numerically compute the solution to just about any initial value problem, with more or less computational work, but we are limited to computing one solution for each specific choice of data, and getting qualitative information for a variety of different data may be costly. On the other hand, an analytical solution formula, when available, may contain this qualitative information for direct information.

An analytical solution formula for a differential equation may thus be viewed as a (smart and beautiful) short-cut to the solution, like evaluating an integral of a function by just evaluating two values of a corresponding primitive function. On the other hand, numerical solution of a differential equation is like a walk along a winding mountain road from point A to point B, without any short-cuts, similar to computing an integral by numerical quadrature. It is useful to be able to use both approaches.

207.2 Generalization to $u'(x) = \lambda(x)u(x) + f(x)$

The first problem we consider is a model in which the rate of change of a quantity $u(x)$ is proportional to the quantity with a variable factor of proportionality $\lambda(x)$, and moreover in which there is an external "forcing" function $f(x)$. The problem reads:

$$u'(x) = \lambda(x)u(x) + f(x) \quad \text{for } x > a, \quad u(a) = u_a, \tag{207.3}$$

where $\lambda(x)$ and $f(x)$ are given functions of x, and u_a is a given initial value. We first describe a couple physical situations being modeled by (207.3).

EXAMPLE 207.1. Consider for time $t > 0$ the population $u(t)$ of rabbits in West Virginia with inital value $u(0) = u_0$ given, which we assume has time dependent known birth rate $\beta(t)$ and death rate $\delta(t)$. In general, we would expect that rabbits will migrate quite freely back and forth across the state border and that the rates of the migration would vary with the season, i.e. with time t. We let $f_i(t)$ and $f_o(t)$ denote the rate of migration into and out of the state respectively at time t, which we assume to be known (realistic?). Then the population $u(t)$ will satisfy

$$\dot{u}(t) = \lambda(t)u(t) + f(t), \quad \text{for } t > a, \quad u(a) = u_a, \tag{207.4}$$

with $\lambda(t) = \beta(t) - \delta(t)$ and $f(t) = f_i(t) - f_o(t)$, which is of the form (207.3). Recall that $\dot{u} = \frac{du}{dt}$.

EXAMPLE 207.2. We model the amount of solute such as salt in a solvent such as water in a tank in which there is both inflow and outflow, see Fig. 207.1. We let $u(t)$ denote the amount of solute in the tank at time t and suppose that we know the initial amount u_0 at $t = 0$. We suppose that a mixture of solute/solvent, of concentration C_i in say

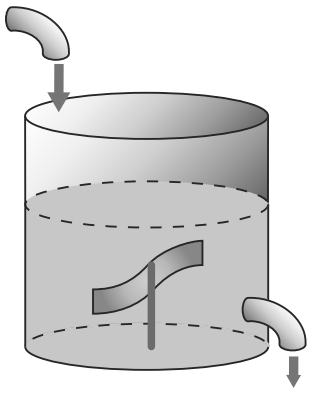


FIGURE 207.1. An illustration of a chemical mixing tank.

grams per liter, flows into the tank at a rate σ_i liters per second. We assume there is also outflow at a rate of σ_o liters per second, and we assume that the mixture in the tank is well mixed with a uniform concentration $C(t)$ at any time t.

To get a differential equation for $u(t)$, we compute the change $u(t+\Delta t)-u(t)$ during the interval $[t, t+\Delta t]$. The amount of solute that flows into the tank during that time interval is $\sigma_i C_i \Delta t$, while the amount of solute that flows out of the tank during that time equals $\sigma_o C(t)\Delta t$, and thus

$$u(t+\Delta t) - u(t) \approx \sigma_i C_i \Delta t - \sigma_o C(t)\Delta t, \tag{207.5}$$

where the approximation improves when we decrease Δt. Now the concentration at time t will be $C(t) = u(t)/V(t)$ where $V(t)$ is the volume of fluid in the tank at time t. Substituting this into (207.5) and dividing by Δt gives

$$\frac{u(t+\Delta t) - u(t)}{\Delta t} \approx \sigma_i C_i - \sigma_o \frac{u(t)}{V(t)}$$

and taking the limit $\Delta t \to 0$ assuming $u(t)$ is differentiable gives the following differential equation for u,

$$\dot{u}(t) = -\frac{\sigma_o}{V(t)} u(t) + \sigma_i C_i.$$

The volume $V(t)$ is determined simply by the flow rates of fluid in and out of the tank. If there is initially V_0 liters in the tank then at time t, $V(t) = V_0 + (\sigma_i - \sigma_o)t$ because the flow rates are assumed to be constant. This gives again a model of the form (207.3):

$$\dot{u}(t) = -\frac{\sigma_o}{V_0 + (\sigma_i - \sigma_o)t} u(t) + \sigma_i C_i \quad \text{for } t > 0, \quad u(0) = u_0. \tag{207.6}$$

The Method of Integrating Factor

We now return to derive an analytical solution formula for (207.3), using the method of *integrating factor*. To work out the solution formula, we begin with the special case

$$u'(x) = \lambda(x)u(x) \quad \text{for } x > a, \quad u(a) = u_a, \tag{207.7}$$

where $\lambda(x)$ is a given function of x. We let $\Lambda(x)$ be a primitive function of $\lambda(x)$ such that $\Lambda(a) = 0$, assuming that $\lambda(x)$ is Lipschitz continuous on $[a, \infty)$. We now multiply the equation $0 = u'(x) - \lambda(x)u(x)$ by $\exp(-\Lambda(x))$, and we get

$$0 = u'(x)\exp(-\Lambda(x)) - u(x)\exp(-(x))\lambda(x) = \frac{d}{dx}(u(x)\exp(-\Lambda(x))),$$

where we refer to $\exp(-\Lambda(x))$ as an integrating factor because it brought the given equation to the form $\frac{d}{dx}$ of something, namely $u(x)\exp(-\Lambda(x))$, equal to zero. We conclude that $u(x)\exp(-\Lambda(x))$ is constant and is therefore equal to u_a since $u(a)\exp(-\Lambda(a)) = u(a) = u_a$. In other words, the solution to (207.7) is given by the formula

$$u(x) = \exp(\Lambda(x))u_a = e^{\Lambda(x)}u_a \quad \text{for } x \geq a. \tag{207.8}$$

We can check by differentiation that this function satisfies (207.7), and thus by uniqueness is the solution. To sum up, we have derived a solution formula for (207.7) in terms of the exponential function and a primitive function $\Lambda(x)$ of the coefficient $\lambda(x)$.

EXAMPLE 207.3. If $\lambda(x) = \frac{r}{x}$ and $a = 1$ then $\Lambda(x) = r\log(x) = \log(x^r)$, and the solution of

$$u'(x) = \frac{r}{x}u(x) \quad \text{for } x \neq 1, \quad u(1) = 1, \tag{207.9}$$

is according to (207.8) given by $u(x) = \exp(r\log(x)) = x^r$. We may define x^r for

Duhamel's Principle

We now continue with the general problem to (207.3). We multiply by $e^{-\Lambda(x)}$, where again $\Lambda(x)$ is the primitive function of $\lambda(x)$ satisfying $\Lambda(a) = 0$, and get

$$\frac{d}{dx}(u(x)e^{-\Lambda(x)}) = f(x)e^{-\Lambda(x)}.$$

Integrating both sides, we see that the solution $u(x)$ satisfying $u(a) = u_a$ can be expressed as

$$u(x) = e^{\Lambda(x)}u_a + e^{\Lambda(x)}\int_a^x e^{-\Lambda(y)}f(y)\,dy. \tag{207.10}$$

This formula for the solution $u(x)$ of (207.3), expressing $u(x)$ in terms of the given data u_a and the primitive function $\Lambda(x)$ of $\lambda(x)$ satisfying $\Lambda(a) = 0$, is referred to as *Duhamel's principle* or the *variation of constants formula*.

We can check the validity of (207.10) by directly computing the derivative of $u(x)$:

$$\begin{aligned} u'(x) &= \lambda e^{\Lambda(x)}u_a + f(x) + \int_0^x (\lambda(x)e^{\Lambda(x)-\Lambda(y)}f(y)\,dy \\ &= \lambda(x)\left(e^{\Lambda(x)}u_a + \int_0^x e^{\Lambda(x)-\Lambda(y)}f(y)\,dy\right) + f(x). \end{aligned}$$

EXAMPLE 207.4. If $\lambda(x) = \lambda$ is constant, $f(x) = x$, $a = 0$ and $u_0 = 0$, the solution of (207.3) is given by

$$\begin{aligned} u(x) &= \int_0^x e^{\lambda(x-y)}y\,dy = e^{\lambda x}\int_0^x ye^{-\lambda y}\,dy \\ &= e^{\lambda x}\big([-\frac{y}{\lambda}e^{-\lambda y}]_{y=0}^{y=x} + \int_0^x \frac{1}{\lambda}e^{-\lambda y}\,dy\big) = -\frac{x}{\lambda} + \frac{1}{\lambda^2}(e^{\lambda x} - 1). \end{aligned}$$

EXAMPLE 207.5. In the model of the rabbit population (207.4), consider a situation with an initial population of 100, the death rate is greater than the birth rate by a constant factor 4, so $\lambda(t) = \beta(t) - \delta(t) = -4$, and there is a increasing migration into the state, so $f(t) = f_i(t) - f_o(t) = t$. Then (207.10) gives

$$\begin{aligned} u(t) &= e^{-4t}100 + e^{-4t}\int_0^t e^{4s}s\,ds \\ &= e^{-4t}100 + e^{-4t}\left(\frac{1}{4}se^{4s}|_0^t - \frac{1}{4}\int_0^t e^{4s}\,ds\right) \\ &= e^{-4t}100 + e^{-4t}\left(\frac{1}{4}te^{4t} - \frac{1}{16}e^{4t} + \frac{1}{16}\right) \\ &= 100.0625e^{-4t} + \frac{t}{4} - \frac{1}{16}. \end{aligned}$$

Without the migration into the state, the population would decrease exponentially, but in this situation the population decreases only for a short time before beginning to increase at a linear rate.

EXAMPLE 207.6. Consider a mixing tank in which the input flow at a rate of $\sigma_i = 3$ liters/sec has a concentration of $C_i = 1$ grams/liter, and the outflow is at a rate of $\sigma_o = 2$ liters/sec, the initial volume is $V_0 = 100$ liters with no solute dissolved, so $u_0 = 0$. The equation is

$$\dot{u}(t) = -\frac{2}{100 + t}u(t) + 3.$$

We find $\Lambda(t) = 2\ln(100+t)$ and so

$$\begin{aligned} u(t) &= 0 + e^{2\ln(100+t)} \int_0^t e^{-2\ln(100+s)} 3\, ds \\ &= (100+t)^2 \int_0^t (100+s)^{-2} 3\, ds \\ &= (100+t)^2 \left(\frac{-3}{100+t} + \frac{3}{100} \right) \\ &= \frac{3}{100} t(100+t). \end{aligned}$$

As expected from the conditions, the concentration increases steadily until the tank is full.

207.3 The Differential Equation $u''(x) - u(x) = 0$

Consider the second order initial value problem

$$u''(x) - u(x) = 0 \quad \text{for } x > 0,\ u(0) = u_0,\ u'(0) = u_1, \tag{207.11}$$

with two initial conditions. We can write the differential equation $u''(x) - u(x) = 0$ formally as

$$(D+1)(D-1)u = 0,$$

where $D = \frac{d}{dx}$, since $(D+1)(D-1)u = D^2u - Du + Du - u = D^2u - u$. Setting $w = (D-1)u$, we thus have $(D+1)w = 0$, which gives $w(x) = ae^{-x}$ with $a = u_1 - u_0$, since $w(0) = u'(0) - u(0)$. Thus, $(D-1)u = (u_1 - u_0)e^{-x}$, so that by Duhamel's principle

$$\begin{aligned} u(x) &= e^x u_0 + \int_0^x e^{x-y}(u_1 - u_0)e^{-y}\, dy \\ &= \frac{1}{2}(u_0 + u_1)e^x + \frac{1}{2}(u_0 - u_1)e^{-x}. \end{aligned}$$

We conclude that the solution $u(x)$ of $u''(x) - u(x) = 0$ is a linear combination of e^x and e^{-x} with coefficients determined by the initial conditions. The technique of "factoring" the differential equation $(D^2 - 1)u = 0$ into $(D+1)(D-1)u = 0$, is very powerful and we now proceed to follow up this idea.

207.4 The Differential Equation $\sum_{k=0}^{n} a_k D^k u(x) = 0$

In this section, we look for solutions of the *linear differential equation with constant coefficients*:

$$\sum_{k=0}^{n} a_k D^k u(x) = 0 \quad \text{for } x \in I, \tag{207.12}$$

where the coefficients a_k are given real numbers, and I is a given interval. Corresponding to the *differential operator* $\sum_{k=0}^{n} a_k D^k$, we define the polynomial $p(x) = \sum_{k=0}^{n} a_k x^k$ in x of degree n with the same coefficients a_k as the differential equation. This is called the *characteristic polynomial* of the differential equation. We can now express the differential operator formally as

$$p(D)u(x) = \sum_{k=0}^{n} a_k D^k u(x).$$

For example, if $p(x) = x^2 - 1$ then $p(D)u = D^2u - u$.

The technique for finding solutions is based on the observation that the exponential function $\exp(\lambda x)$ has the following property:

$$p(D)\exp(\lambda x) = p(\lambda)\exp(\lambda x), \tag{207.13}$$

which follows from repeated use of the Chain rule. This translates the differential operator $p(D)$ acting on $\exp(\lambda x)$ into the simple operation of multiplication by $p(\lambda)$. Ingenious, right?

We now seek solutions of the differential equation $p(D)u(x) = 0$ on an interval I of the form $u(x) = \exp(\lambda x)$. This leads to the equation

$$p(D)\exp(\lambda x) = p(\lambda)\exp(\lambda x) = 0, \quad \text{for } x \in I,$$

that is, λ should be a root of the polynomial equation

$$p(\lambda) = 0. \tag{207.14}$$

This algebraic equation is called the *characteristic equation* of the differential equation $p(D)u = 0$. To find the solutions of a differential equation $p(D)u = 0$ on the interval I, we are thus led to search for the roots $\lambda_1, ...\lambda_n$, of the algebraic equation $p(\lambda) = 0$ with corresponding solutions $\exp(\lambda_1 x), ..., \exp(\lambda_n x)$. Any linear combination

$$u(x) = \alpha_1 \exp(\lambda_1 x) + ... + \alpha_n \exp(\lambda_n x), \tag{207.15}$$

with α_i real (or complex) constants, will then be a solution of the differential equation $p(D)u = 0$ on I. If there are n distinct roots λ_1,...,λ_n, then the

general solution of $p(D)u = 0$ has this form. The constants α_i will be determined from initial or boundary conditions in a specific situation.

If the equation $p(\lambda) = 0$ has a multiple roots λ_i of multiplicity r_i, then the situation is more complicated. It can be shown that the solution is a sum of terms of the form $q(x)\exp(\lambda_i x)$, where $q(x)$ is a polynomial of degree at most $r_i - 1$. For example, if $p(D) = (D-1)^2$, then the general solution of $p(D)u = 0$ has the form $u(x) = (\alpha_0 + \alpha_1 x)\exp(x)$. In the Chapter *N-body systems* below we study the the constant coefficient linear second order equation $a_0 + a_1 Du + a_2 D^2 u = 0$ in detail, with interesting results!

The translation from a differential equation $p(D)u = 0$ to an algebraic equation $p(\lambda) = 0$ is very powerful, but requires the coefficients a_k of $p(D)$ to be independent of x and is thus not very general. The whole branch of *Fourier analysis* is based on the formula (207.13).

EXAMPLE 207.7. The characteristic equation for $p(D) = D^2 - 1$ is $\lambda^2 - 1 = 0$ with roots $\lambda_1 = 1, \lambda_2 = -1$, and the corresponding general solution is given by $\alpha_1 \exp(x) + \alpha_2 \exp(-x)$. We already met this example just above.

EXAMPLE 207.8. The characteristic equation for $p(D) = D^2 + 1$ is $\lambda^2 + 1 = 0$ with roots $\lambda_1 = i, \lambda_2 = -i$, and the corresponding general solution is given by

$$\alpha_1 \exp(ix) + \alpha_2 \exp(-ix).$$

with the α_i complex constants. Taking the real part, we get solutions of the form

$$\beta_1 \cos(x) + \beta_2 \sin(x)$$

with the β_i real constants.

207.5 The Differential Equation $\sum_{k=0}^{n} a_k D^k u(x) = f(x)$

Consider now the nonhomogeneous differential equation

$$p(D)u(x) = \sum_{k=0}^{n} a_k D^k u(x) = f(x), \tag{207.16}$$

with given constant coefficients a_k, and a given right hand side $f(x)$. Suppose $u_p(x)$ is any solution of this equation, which we refer to as a *particular solution*. Then any other solution $u(x)$ of $p(D)u(x) = f(x)$ can be written

$$u(x) = u_p(x) + v(x)$$

where $v(x)$ is a solution of the corresponding homogeneous differential equation $p(D)v = 0$. This follows from linearity and uniqueness since $p(D)(u - u_p) = f - f = 0$.

EXAMPLE 207.9. Consider the equation $(D^2 - 1)u = f(x)$ with $f(x) = x^2$. A particular solution is given by $u_p(x) = -x^2 - 2$, and thus the general solution is given by

$$u(x) = -x^2 - 2 + \alpha_1 \exp(x) + \alpha_2 \exp(-x).$$

207.6 Euler's Differential Equation

In this section, we consider Euler's equation

$$a_0 u(x) + a_1 x u'(x) + a_2 x^2 u''(x) = 0, \tag{207.17}$$

which has variable coefficients $a_i x^i$ of a very particular form. Following a grand mathematical tradition, we guess, or make an *Ansatz* on the form of the solution, and assume that $u(x) = x^m$ for some m to be determined. Substituting into the differential equation, we get

$$a_0 x^m + a_1 x (x^m)' + a_2 x^2 (x^m)'' = (a_0 + (a_1 - 1)m + a_2 m^2) x^m,$$

and we are thus led to the auxiliary algebraic equation

$$a_0 + (a_1 - 1)m + a_2 m^2 = 0$$

in m. Letting the roots of this equation be m_1 and m_2, assuming the roots are real, any linear combination

$$\alpha_1 x^{m_1} + \alpha_2 x^{m_2}$$

is a solution of (207.17). In fact the general solution of (207.17) has this form if m_1 and m_2 are distinct and real.

EXAMPLE 207.10. The auxiliary equation for the differential equation $x^2 u'' - \frac{3}{2} x u' - 2u = 0$ is $m^2 - \frac{7}{2} m - 2 = 0$ with roots $m_1 = -\frac{1}{2}$ and $m_2 = 4$ and thus the general solution takes the form

$$u(x) = \alpha_1 \frac{1}{\sqrt{x}} + \alpha_2 x^4.$$

Leonard Euler (1707-83) is the mathematical genius of the 18th century, with an incredible production of more than 800 scientific articles half of them written after he became completely blind in 1766, see Fig. 179.1.

Chapter 207 Problems

207.1. Solve the initial value problem (207.7) with $\lambda(x) = x^r$, where $r \in \mathbb{R}$, and $a = 0$.

207.2. Solve the following initial value problems: a) $u'(x) = 8xu(x), \quad u(0) = 1, \quad x > 0$, b) $\frac{(15x+1)u(x)}{u'(x)} = 3x, \quad u(1) = \mathrm{e}, \quad x > 1$, c) $u'(x) + \frac{x}{(1-x)(1+x)}u = 0, \quad u(0) = 1, \quad x > 0$.

207.3. Make sure that you got the correct answer in the previous problem, part c). Will your solution hold for $x > 1$ as well as $x < 1$?

207.4. Solve the following initial value problems: a) $xu'(x) + u(x) = x, \quad u(1) = \frac{3}{2}, \quad x > 1$, b) $u'(x) + 2xu = x, \quad u(0) = 1, \quad x > 0$, c) $u'(x) = \frac{x+u}{2}, \quad u(0) = 0, \quad x > 0$.

207.5. Describe the behavior of the population of rabbits in West Virginia in which the birth rate exceeds the death rate by 5, the initial population is 10000 rabbits, and (a) there is a net migration out of the state at a rate of $5t$ (b) there is a net migration out of the state at a rate of $\exp(6t)$.

207.6. Describe the concentration in a mixing tank with an initial volume of 50 liters in which 20 grams of solute are dissolved, there is an inflow of 6 liters/sec with a concentration of 10 grams/liter and an outflow of 7 liters/sec.

208
Improper Integrals

All sorts of funny thoughts, run around my head. (When We Were Very Young, Milne)

208.1 Introduction

In some applications, it is necessary to compute integrals of functions that are unbounded at isolated points or to compute integrals of functions over unbounded intervals. We call such integrals *improper*, or sometimes (more properly) *generalized* integrals. We compute these integrals using the basic results on convergence of sequences that we have already developed.

We now consider these two kinds of improper integrals: integrals over unbounded intervals and integrals of unbounded functions.

208.2 Integrals Over Unbounded Intervals

We start considering the following example of an integral over the unbounded interval $[0, \infty)$:

$$\int_0^\infty \frac{1}{1+x^2}\, dx.$$

The integrand $f(x) = (1+x^2)^{-1}$ is a smooth (positive) function that we can integrate over any finite interval $[0, n]$ to get,

$$\int_0^n \frac{1}{1+x^2}\,dx = \arctan(n). \tag{208.1}$$

Now we consider what happens as n increases, that is we integrate f over increasingly longer intervals. Since $\lim_{n\to\infty} \arctan(n) = \pi/2$, we may write

$$\lim_{n\to\infty} \int_0^n \frac{1}{1+x^2}\,dx = \frac{\pi}{2},$$

and we are thus led to *define*

$$\int_0^\infty \frac{1}{1+x^2}\,dx = \lim_{n\to\infty} \int_0^n \frac{1}{1+x^2}\,dx = \frac{\pi}{2}.$$

We generalize in the obvious way to an arbitrary (Lipschitz continuous) function $f(x)$ defined for $x > a$, and thus define

$$\int_a^\infty f(x)\,dx = \lim_{n\to\infty} \int_a^n f(x)\,dx \tag{208.2}$$

granted that the limit is defined and is finite. In this case, we say the improper integral is *convergent* (or is *defined*) and that the function $f(x)$ is *integrable* over $[a, \infty)$. Otherwise, we say the integral is *divergent* (or is *undefined*), and that $f(x)$ is *not* integrable over $[a, \infty)$.

If the function $f(x)$ is positive, then in order for the integral $\int_a^\infty f(x)\,dx$ to be convergent, the integrand $f(x)$ has to get sufficiently small for large values of x, since otherwise $\lim_{n\to\infty} \int_a^n f(x)\,dx = \infty$ and the integral is divergent. We saw above that the function $\frac{1}{1+x^2}$ was decaying to zero sufficiently quickly for large values of x to be integrable over $[a, \infty)$.

Consider now the function $\frac{1}{1+x}$ with a less quick decay as $x \to \infty$. Is it integrable on $[0, \infty)$? Well, we have

$$\int_0^n \frac{1}{1+x}\,dx = \Big[\log(1+x)\Big]_0^n = \log(1+n),$$

and since

$$\log(1+n) \to \infty \quad \text{as } n \to \infty$$

although the divergence is slow, we understand that $\int_0^\infty \frac{1}{1+x}\,dx$ is divergent.

EXAMPLE 208.1. The improper integral

$$\int_1^\infty \frac{dx}{x^\alpha}$$

is convergent for $\alpha > 1$, since

$$\lim_{n\to\infty} \int_1^n \frac{dx}{x^{-\alpha}} = \lim_{n\to\infty} [-\frac{x^{-(\alpha-1)}}{\alpha-1}]_1^n = \frac{1}{\alpha-1}.$$

We can sometimes show that an improper integral exists even when we can not compute its value.

EXAMPLE 208.2. Consider the improper integral

$$\int_1^\infty \frac{e^{-x}}{x}\,dx.$$

Since $f(x) = \frac{e^{-x}}{x} > 0$ for $x > 1$, we see that the sequence $\{I_n\}_{n=1}^\infty$, with

$$I_n = \int_1^n \frac{e^{-x}}{x}\,dx$$

is increasing. By Chapter *Optimization* we know that $\{I_n\}_{n=1}^\infty$ will have a limit if we only can show that $\{I_n\}_{n=1}^\infty$ is bounded above. Since trivially $1/x \leq 1$ if $x \leq 1$, we have for all $n \geq 1$

$$I_n \leq \int_1^n e^{-x}\,dx = e^{-1} - e^{-n} \leq e^{-1}.$$

We conclude that $\int_1^\infty \frac{e^{-x}}{x}\,dx$ converges. Note that we may restrict n to take integer values because the integrand e^{-x}/x tends to zero as x tends to infinity.

We may also compute integrals of the form

$$\int_{-\infty}^\infty f(x)\,dx.$$

We do this by choosing an arbitrary point $-\infty < a < \infty$ and defining

$$\begin{aligned}\int_{-\infty}^\infty f(x)\,dx &= \int_{-\infty}^a f(x)\,dx + \int_a^\infty f(x)\,dx \\ &= \lim_{m\to-\infty}\int_m^a f(x)\,dx + \lim_{n\to\infty}\int_a^n f(x)\,dx,\end{aligned} \tag{208.3}$$

where we compute the two limits independently and both must be defined and finite for the integral to exist.

208.3 Integrals of Unbounded Functions

We begin this section by considering the integral

$$\int_a^b f(x)\,dx,$$

where $f(x)$ is unbounded at a, i.e. $\lim_{x\downarrow a} f(x) = \pm\infty$. We consider the following example:

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx.$$

The function $\frac{1}{\sqrt{x}}$ is unbounded on $(0,1]$, but bounded and Lipschitz continuous on $[\epsilon,1]$ for any $1 \geq \epsilon > 0$. This means that the integrals

$$I_\epsilon = \int_\epsilon^1 \frac{1}{\sqrt{x}}\,dx = 2 - 2\sqrt{\epsilon} \tag{208.4}$$

are defined for any $1 \geq \epsilon > 0$, and evidently

$$\lim_{\epsilon\downarrow 0 \to \infty} I_\epsilon = 2,$$

where we recall that $\epsilon \downarrow 0$ means that ϵ tends to zero through positive values. It is thus natural to define

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx = \lim_{\epsilon\downarrow 0} \int_\epsilon^1 \frac{1}{\sqrt{x}}\,dx = 2.$$

In general if $f(x)$ is unbounded close to a, then we define

$$\int_a^b f(x)\,dx = \lim_{s\downarrow a} \int_s^b f(x)\,dx, \tag{208.5}$$

and if $f(x)$ is unbounded at b then we define

$$\int_a^b f(x)\,dx = \lim_{s\uparrow a} \int_a^s f(x)\,dx \tag{208.6}$$

when these limits are defined and finite. As above, we say the improper integrals are convergent and defined if the limits exist and are finite, and otherwise say the integrals are divergent and not defined.

We may naturally extend this definition to the case when $f(x)$ is unbounded at a point $a < c < b$ by defining

$$\begin{aligned}\int_a^b f(x)\,dx &= \lim \int_a^c f(x)\,dx + \lim \int_c^b f(x)\,dx \\ &= \lim_{s\uparrow c} \int_a^s f(x)\,dx + \lim_{t\downarrow c} \int_t^b f(x)\,dx\end{aligned} \tag{208.7}$$

where the two limits are computed independently and must both be defined and finite for the integral to converge.

Chapter 208 Problems

208.1. If possible, compute the following integrals

1. $\displaystyle\int_0^\infty \frac{x}{(1+x^2)^2}\,dx$
2. $\displaystyle\int_{-\infty}^\infty xe^{-x^2}\,dx$
3. $\displaystyle\int_0^1 \frac{1}{\sqrt{1-x}}\,dx$
4. $\displaystyle\int_0^\pi \frac{\cos(x)}{(1-\sin(x))^{1/3}}\,dx$

208.2. Prove that if $\int_0^\infty |f(x)|\,dx$ is convergent, then so is $\int_0^\infty f(x)\,dx$, that is, absolute convergence implies convergence.

208.3. Prove that $\int_B \|x\|^{-\alpha}\,dx$, where $B = \{x \in \mathbb{R}^d : \|x\| < 1\}$, is convergent if $\alpha < d$ for $d = 1, 2, 3$.

209
Series

If you disregard the very simplest cases, there is in all of mathematics not a single series whose sum has been rigorously determined. In other words, the most important part of mathematics stand without a foundation.(Abel 1802-1829)

209.1 Introduction

In this chapter we consider the concept of *series*, which is a sum of numbers. We distinguish between a *finite* series, where the sum has a finite number of terms, and an *infinite series* with an infinite number of terms. A finite series does not pose any mysteries; we can, at least in principle, compute the sum of a finite series by adding the terms one-by-one, given enough time. The concept of an infinite series requires some explanation, since we cannot actually add an infinite number of terms one-by-one, and we thus need to define what we mean by an "infinite sum".

The concept of infinite series has a central role in Calculus, because a basic idea has been to seek to express "arbitrary" functions in terms of series as sums of simple terms. This was the grand idea of Fourier who thought of representing general functions as sums of trigonometric functions in the form of Fourier series, and Weierstrass who tried to do the same with monomials or polynomials in the form of power series. There are limitations to both Fourier and power series and the role of such series is today largely being taken over by computational methods. We therefore do not go into

any excessive treatment of series, but we do present some important basic facts, which are useful to know.

We recall that we already met one infinite series, namely the *geometric series*

$$\sum_{i=0}^{\infty} a^i = 1 + a + a^2 + a^3 + \cdots ,$$

where a is a real number. We determined the sum of this infinite series in the case $|a| < 1$ by first computing the *partial sum of order* n:

$$s_n = \sum_{i=0}^{n} a^i = 1 + a + a^2 + \cdots + a^n = \frac{1 - a^{n+1}}{1 - a}.$$

by summing the terms a^i with $i \leq n$. We then made the observation that if $|a| < 1$, then

$$\lim_{n\to\infty} s_n = \lim_{n\to\infty} \frac{1 - a^{n+1}}{1 - a} = \frac{1}{1 - a},$$

and so we defined for $|a| < 1$ the sum of the infinite geometric series to be

$$\sum_{i=0}^{\infty} a^i = \lim_{n\to\infty} \sum_{i=0}^{n} a^i = \frac{1}{1 - a}.$$

We note that if $|a| \geq 1$, then we had to leave the sum of the geometric series $\sum_{i=0}^{\infty} a^i$ undefined. If $|a| \geq 1$, then $|s_n - s_{n-1}| = |a^n| \geq 1$, and therefore $\{s_n\}_{n=0}^{\infty}$ is not a Cauchy sequence, and thus $\lim_{n\to\infty} s_n = \lim_{n\to\infty} \sum_{i=0}^{n} a^i$ does not exist. Evidently, a necessary condition for convergence is that the terms a^i tend to zero as i tends to infinity.

209.2 Definition of Convergent Infinite Series

We now generalize these ideas to arbitrary infinite series. Thus let $\{a_n\}_{n=0}^{\infty}$ denote a sequence of real numbers and consider the *sequence of partial sums* $\{s_n\}_{n=0}^{\infty}$, where

$$s_n = \sum_{i=0}^{n} a_i = a_0 + a_1 + \cdots + a_n \tag{209.1}$$

is the *partial sum of order* n. We now say that the series $\sum_{i=0}^{\infty} a_i$ is *convergent* if the corresponding sequence of partial sums $\{s_n\}_{n=0}^{\infty}$ converges, and we then write

$$\sum_{i=0}^{\infty} a_i = \lim_{n\to\infty} s_n = \lim_{n\to\infty} \sum_{i=0}^{n} a_i, \tag{209.2}$$

which we refer to as the sum of the series. The convergence of a series $\sum_{i=1}^{\infty} a_i$ is thus reduced to the convergence of the sequence of it's partial

sums. All convergence issues for a series are handled in this way by reduction to convergence of sequences. This chapter therefore may be viewed as a direct gives a direct continuation of Chapters *Sequences and limits* and *Real numbers*. In particular, we understand as in the case of a geometric series, that a necessary condition for convergence of a series $\sum_{i=0}^{\infty} a_i$ is that the terms a_i tend to zero as i tends to infinity. However, this condition is not sufficient, as we should know from our previous experience with sequences, and as we will see again below.

Note that we can similarly consider series of the form $\sum_{i=1}^{\infty} a_i$ or $\sum_{i=m}^{\infty} a_i$ for any integer m.

Note that in a few special cases like the geometric series, we can actually find an analytic formula for the sum of the series. However, for most series $\sum_{i=0}^{\infty} a_i$ this is not possible, or may be so be tricky that we can't make it. Of course, we can then usually compute an approximation by directly computing a partial sum $s_n = \sum_{i=0}^{n} a_i$ for some appropriate n, that is, if n is not too big and the terms a_i not too difficult to evaluate. To then estimate the error, we are led to estimate the remainder $\sum_{i=n+1}^{\infty} a_i$. Thus we see a need to be able to analytically estimate the sum of a series, which may be easier than to analytically compute the exact sum.

In particular, such estimation may be used to decide if a series is convergent or not, which of course is an important issue because playing around with divergent series cannot have any meaning. In this pursuit, it is natural to distinguish between series in which all of the terms have the same sign and those in which the terms can have different signs. It may be more difficult to determine convergence for a series in which the terms can have different signs because of the possibility of cancellation between the terms.

Further, if we bound a series remainder $\sum_{i=n+1}^{\infty} a_i$ by using the triangle inequality, we get

$$|\sum_{i=n+1}^{\infty} a_i| \leq \sum_{i=n+1}^{\infty} |a_i|,$$

where the series on the right hand side is positive. So, positive series are of prime importance and we now turn to this topic.

209.3 Positive Series

A series $\sum_{i=1}^{\infty} a_i$ is said to be a *positive series*, if $a_i \geq 0$ for $i = 1, 2, ...$ The important point about a positive series is that the sequence of partial sums is non-decreasing, because

$$s_{n+1} - s_n = \sum_{i=1}^{n+1} a_i - \sum_{i=1}^{n} a_i = a_{n+1} \geq 0. \tag{209.3}$$

In Chapter *Optimization* below we shall prove that a nondecreasing sequence converges if and only if the sequence is bounded above. If we accept this as a fact, we understand that a positive series is convergent if and only if the sequence of partial sums is bounded above, that is there is a constant C such that

$$\sum_{i=1}^{n} a_i \le C \quad \text{for } n = 1, 2, ..., . \tag{209.4}$$

This gives a definite way to check convergence, which we state as a theorem:

Theorem 209.1 *A positive series converges if and only if the sequence of partial sums is bounded above.*

This result does not apply if the series has terms with different signs. For example, the series $\sum_{i=0}^{\infty}(-1)^i = 1 - 1 + 1 - 1 + 1...$ has bounded partial sums, but is not convergent since $(-1)^i$ does not tend to zero as i tends to infinity.

EXAMPLE 209.1. We can sometimes use an integral to bound the partial sums of a positive series and thus to prove convergence or estimate remainders. As an example, consider the positive series $\sum_{i=2}^{\infty}\frac{1}{i^2}$. The partial sum

$$s_n = \sum_{i=2}^{n} \frac{1}{i^2}$$

may be viewed as a quadrature formula for the integral of $\int_1^n x^{-2}\,dx$, see Fig. 209.1.

More precisely, we see that

$$\begin{aligned}\int_1^n x^{-2}\,dx &= \int_1^2 x^{-2}\,dx + \int_2^3 x^{-2}\,dx + \cdots + \int_{n-1}^n x^{-2}\,dx \\ &\ge \int_1^2 \frac{1}{2^2}\,dx + \int_2^3 \frac{1}{3^2}\,dx + \cdots + \int_{n-1}^n \frac{1}{n^2}\,dx \\ &\ge \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2} = s_n.\end{aligned}$$

Since

$$\int_1^n x^{-2}\,dx = \big(1 - \frac{1}{n}\big) \le 1,$$

we conclude that $s_n \le 1$ for all n and therefore the series $\sum_{i=2}^{\infty}\frac{1}{i^2}$ is convergent. To compute an approximation of the sum of the series, we of course compute a partial sum s_n with n sufficiently large. To estimate the remainder we may of course use a similar comparison, see Problem 209.5.

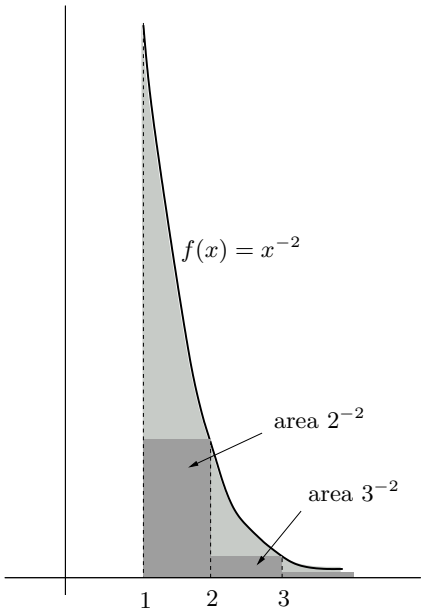


FIGURE 209.1. The relation between $\int_1^n x^{-2}\,dx$ and $\sum_{i=2}^n i^{-2}$.

EXAMPLE 209.2. The positive series $\sum_{i=1}^{\infty} \frac{1}{i+i^2}$ converges because for all n

$$s_n = \sum_{i=1}^{n} \frac{1}{i+i^2} \leq \sum_{i=1}^{n} \frac{1}{i^2} \leq 2$$

by the previous example.

Similarly, a *negative series* with all terms non-positive, converges if and only if its partial sums are bounded *below*.

EXAMPLE 209.3. For the *alternating* series

$$\sum_{i=1}^{\infty} \frac{(-1)^i}{i},$$

we have that the difference between two successive partial sums

$$s_n - s_{n-1} = \frac{(-1)^n}{n}$$

alternates in sign, and thus the sequence of partial sums is not monotone, and therefore we cannot decide convergence or not from the above theorem. We shall return to this series below and prove that it is in fact convergent.

209.4 Absolutely Convergent Series

Now we turn to series with terms of different signs. We begin by first considering series that converge regardless of any cancellation between the terms. We are motivated by the convergence results for positive series. A series $\sum_{i=1}^{\infty} a_i$ is said to be *absolutely convergent* if the series

$$\sum_{i=1}^{\infty} |a_i|$$

converges. By the previous result we know that a series $\sum_{i=1}^{\infty} a_i$ is absolutely convergent if and only if the sequence $\{\hat{s}_n\}$ with

$$\hat{s}_n = \sum_{i=1}^{n} |a_i|, \tag{209.5}$$

is bounded above.

We shall now prove that an absolutely convergent series $\sum_{i=1}^{\infty} a_i$ is convergent. By the triangle inequality we have for $m > n$,

$$|s_m - s_n| = |\sum_{n}^{m} a_i| \leq \sum_{n}^{m} |a_i| = |\hat{s}_m - \hat{s}_n|. \tag{209.6}$$

Now, since we can make $|\hat{s}_m - \hat{s}_n|$ arbitrarily small by taking m and n large, because $\sum_{i=1}^{\infty} a_i$ is absolutely convergent and thus $\{\hat{s}_n\}_{n=1}^{\infty}$ is a Cauchy sequence, we conclude that $\{s_n\}_{n=1}^{\infty}$ is a Cauchy sequence and therefore converges and thus the series $\sum_{i=1}^{\infty} a_i$ is convergent. We state this fundamental result as a theorem:

Theorem 209.2 *An absolutely convergent series is convergent.*

EXAMPLE 209.4. The series $\sum_{i=1}^{\infty} \frac{(-1)^i}{i^2}$ is convergent because $\sum_{i=1}^{\infty} \frac{1}{i^2}$ is convergent.

209.5 Alternating Series

The convergence of a general series with terms of "random" sign may be very difficult to analyze because of cancellation of terms. We now consider a special case with a regular pattern to the signs of the terms:

$$\sum_{i=0}^{\infty} (-1)^i a_i \tag{209.7}$$

where $a_i \geq 0$ for all i. This is called an *alternating series* since the signs of the terms alternate. We shall now prove that if $a_{i+1} \leq a_i$ for $i = 0, 1, 2...$ and

$\lim_{i\to\infty} a_i = 0$, then the alternating series converges. The key observation is that the sequence $\{s_n\}$ of partial sums satisfies

$$s_1 \leq s_3 \leq s_5 \leq ...s_{2j+1} \leq s_{2i} \leq ... \leq s_4 \leq s_2 \leq s_0, \tag{209.8}$$

which shows that both limits $\lim_{j\to\infty} s_{2j+1}$ and $\lim_{i\to\infty} s_{2i}$ exist. Since $a_i \to 0$ as i tends to infinity, $\lim_{j\to\infty} s_{2j+1} = \lim_{i\to\infty} s_{2i}$, and thus $\lim_{n\to\infty} s_n$ exists and convergence of the series $\sum_{i=0}^{\infty}(-1)^i a_i$ follows. We summarize in the following theorem first stated and proved by Leibniz:

Theorem 209.3 *An alternating series with the property that the modulus of its terms tends monotonically to zero, converges.*

EXAMPLE 209.5. The *harmonic series*

$$\sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots$$

converges. We now proceed to show that this series is not absolutely convergent.

209.6 The Series $\sum_{i=1}^{\infty} \frac{1}{i}$ Theoretically Diverges!

We shall now show that the *harmonic series* $\sum_{i=1}^{\infty} \frac{(-1)^i}{i}$ is **not** absolutely convergent, i.e. we shall prove that the series

$$\sum_{i=1}^{\infty} \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$

diverges. We do this by proving that the sequence $\{s_n\}_{n=1}^{\infty}$ of partial sums

$$s_n = \sum_{i=1}^{n} \frac{1}{i}$$

can become arbitrarily large if n is large enough. To see this we group the terms of a partial sum as follows:

$$1 + \overline{\frac{1}{2}} + \overline{\frac{1}{3} + \frac{1}{4}} + \overline{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}} + \overline{\frac{1}{9} + \frac{1}{10} + \frac{1}{11} + \frac{1}{12} + \frac{1}{13} + \frac{1}{14} + \frac{1}{15} + \frac{1}{16}} + \overline{\frac{1}{17} + \cdots + \frac{1}{32}} + \cdots$$

The first "group" is 1/2. The second group is

$$\frac{1}{3}+\frac{1}{4}\geq\frac{1}{4}+\frac{1}{4}=\frac{1}{2}.$$

The third group is

$$\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}\geq\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}=\frac{1}{2}.$$

The fourth group

$$\frac{1}{9}+\frac{1}{10}+\frac{1}{11}+\frac{1}{12}+\frac{1}{13}+\frac{1}{14}+\frac{1}{15}+\frac{1}{16}$$

has 8 terms that are larger than 1/16, so it also gives a sum larger than $8/16 = 1/2$. We can continue in this way, taking the next 16 terms, all of which are larger than 1/32, then the next 32 terms, all of which are larger than 1/64, and so on. Each time we take a group, we get a contribution to the overall sum that is larger than 1/2.

When we take n larger and larger, we can combine more and more terms in this way, making the sum larger in increments of 1/2 each time. The partial sums therefore just become larger and larger as n increases, which means the partial sums diverge to infinity.

Note that by the arithmetic rules, the partial sum s_n should be the same whether we compute the sum in the "forward" direction

$$s_n = 1+\frac{1}{2}+\frac{1}{3}+\cdots\frac{1}{n-1}+\frac{1}{n}$$

or the "backward" direction

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

In Fig. 209.2, we list various partial sums in both the forward and backward directions computed using FORTRAN with single precision variables with about 7 digits of accuracy. Note two things about these results:

First, the computed partial sums s_n all become equal when n is large enough, even though theoretically they should keep increasing to infinity as n increases. This is because in finite precision the new terms we add eventually get so small that they effectively give zero contribution. Thus, although in principle the series diverges, in practice the series appears to converge on the computer. This gives an illustration of idealism vs realism in mathematics!

Second, the backward sum is strictly larger than the forward sum! This is because in the summation a term effectively adds zero when the term is sufficiently small compared to the current partial sum, and the size of the partial sums is vastly different if we add in a forward or backward manner.

n	forward sum	backward sum
10000	9.787612915039062	9.787604331970214
100000	12.090850830078120	12.090151786804200
1000000	14.357357978820800	14.392651557922360
2000000	15.311032295227050	15.086579322814940
3000000	15.403682708740240	15.491910934448240
5000000	15.403682708740240	16.007854461669920
10000000	15.403682708740240	16.686031341552740
20000000		17.390090942382810
30000000		17.743585586547850
40000000		18.257812500000000
50000000		18.807918548583980
100000000	15.403682708740240	18.807918548583980
200000000		18.807918548583980
1000000000		18.807918548583980

FIGURE 209.2. Forward $1 + \frac{1}{2} + \cdots + \frac{1}{n}$ and backward $\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{2} + 1$ partial harmonic sums for various n computed with double precision.

FIGURE 209.3. Niels Henrik Abel (1802-1829):“ The divergent series are the invention of the devil, and it is a shame to base on them any demonstration whatsoever. By using them, one may draw any conclusion he pleases and that is why these series have produced so many fallacies and so many paradoxes ..”.

209.7 Abel

Niels Henrik Abel (1802-1829), the great mathematical genius of Norway, is today world famous for his half-page proof from 1824 of the impossibility of solving polynomial equations of degree larger or equal to five by root-extraction. This settled a famous problem which had haunted many generations of mathematicians. However, Abel's life was short and tragic and his fame came only after his sudden death at the age of 27. Gauss in Göttingen was indifferent to the proof when it was first presented, based on his view expressed in his thesis of 1801 that the algebraic solution of an equation was no better than devising a symbol for the root of the equation and then saying that the equation had a root equal to the symbol (compare the square root of two).

Abel tried also unsuccessfully to convince Cauchy on a trip to Paris 1825, which ended in misery, and he then left for Berlin on borrowed money but succeeded to produce another master-piece now on so called elliptic integrals. After returning to a modest position in Christiania he continued to pour out high quality mathematics while his health was deteriorating. After a sled journey to visit his girl friend for Christmas 1828 he became seriously ill and died quickly after.

FIGURE 209.4. Evariste Galois: (1811-1832):" Since the beginning of the century, computational procedures have become so complicated that any progress by those means has become impossible, without the elegance which modern mathematicians have brought to bear on their research, and by means of which the spirit comprehends quickly and in one step a great many computations. It is clear that elegance, so vaunted and so aptly named, can have no other purpose. ... Go to the roots, of these calculations! Group the operations. Classify them according to their complexities rather than their appearances! This, I believe, is the mission of future mathematicians. This is the road on which I am embarking in this work" (from the preface to Galois' final manuscript).

209.8 Galois

Abel is contemporary with Evariste Galois (1811-32), who independently 1830 proved the same fifth order equation result as Abel, again with no reaction from Cauchy. Galois was refused twice in the entrance exam to Ecole Polytechnique apparently after accusing the examiner for posing questions incorrectly. Galois was imprisoned for a revolutionary speech against King Louis Philippe 1830, was released in 1832 but soon died after wounds from a duel about his girl friend, at the age of 21.

Chapter 209 Problems

209.1. Prove that the series $\sum_{i=1}^{\infty} i^{-\alpha}$ converges if and only if $\alpha > 1$. Hint: Compare with a primitive function of $x^{-\alpha}$.

209.2. Prove that the series $\sum_{i=1}^{\infty}(-1)^i i^{-\alpha}$ converges if and only if $\alpha > 0$.

209.3. Prove that the following series converges: (a) $\displaystyle\sum_{i=1}^{\infty} e^{-i}$. (b) $\displaystyle\sum_{i=1}^{\infty} \frac{1+(-1)^i}{i^2}$. (c) $\displaystyle\sum_{i=1}^{\infty} \frac{e^{-i}}{i}$. (d) $\displaystyle\sum_{i=1}^{\infty} \frac{1}{(i+1)(i+4)}$.

209.4. Prove that $\displaystyle\sum_{i=1}^{\infty} \frac{1}{i^2-i}$ converges. Hint: first show that $\frac{1}{2}i^2 - i \geq 0$ for $i \geq 2$.

209.5. Estimate the remainder $\sum_{i=n}^{\infty} \frac{1}{i^2}$ for different values of n.

209.6. Prove that $\displaystyle\sum_{i=1}^{\infty}(-1)^i \sin(1/i)$ converges. More difficult: prove that it is **not** absolutely convergent.

209.7. Explain in detail why the backward partial sum of the series $\sum_{i=1}^{\infty} \frac{1}{i}$ is larger than the forward sum.

210

Scalar Autonomous Initial Value Problems

He doesn' use long, difficult words, like Owl. (The House at Pooh Corner, Milne)

210.1 Introduction

In this chapter, we consider the initial value problem for a *scalar autonomous non-linear differential equation*: Find a function $u : [0,1] \to \mathbb{R}$ such that

$$u'(x) = f(u(x)) \quad \text{for } 0 < x \leq 1,\ u(0) = u_0, \tag{210.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a given function and u_0 a given initial value. We assume that $f : \mathbb{R} \to \mathbb{R}$ is bounded and Lipschitz continuous, that is, there are constants L_f and M_f such that for all $v, w \in \mathbb{R}$,

$$|f(v) - f(w)| \leq L_f|v - w|, \quad \text{and} \quad |f(v)| \leq M_f. \tag{210.2}$$

For definiteness, we choose the interval $[0,1]$, and we may of course generalize to any interval $[a,b]$.

The problem (210.1) is in general *non-linear*, since $f(v)$ in general is non-linear in v, that is, $f(u(x))$ depends non-linearly on $u(x)$. We have already in Chapter *The exponential function* considered the basic case with f linear, which is the case $f(u(x)) = u(x)$ or $f(v) = v$. Now we pass on to nonlinear functions such as $f(v) = v^2$ and others.

Further, we call (210.1) *autonomous* because $f(u(x))$ depends on the value of the solution $u(x)$, but not directly on the independent variable x.

A *non-autonomous* differential equation has the form $u'(x) = f(u(x), x)$, where $f(u(x), x)$ depends on both $u(x)$ and x. The differential equation $u'(x) = xu^2(x)$ is non-autonomous and non-linear with $f(v, x) = xv^2$, while the equation $u'(x) = u(x)$ defining the exponential is autonomous and linear with $f(v) = v$.

Finally, we refer to (210.1) as a *scalar* problem since $f : \mathbb{R} \to \mathbb{R}$ is a real valued function of one real variable, that is, $v \in \mathbb{R}$ and $f(v) \in \mathbb{R}$, and thus $u(x)$ takes real values or $u : [0, 1] \to \mathbb{R}$. Below we shall consider *systems* of equations with $f : \mathbb{R}^d \to \mathbb{R}^d$ and $u : [0, 1] \to \mathbb{R}^d$, where $d > 1$, which models a very large range of phenomena.

We hope the reader (like Owl) is now at ease with the terminology: In this chapter we thus focus on scalar autonomous non-linear differential equations.

The initial value problem for a scalar autonomous differential equation is the simplest of all initial value problems and the solution (when it exists) can be expressed analytically in terms of a primitive function $F(v)$ of the function $1/f(v)$. In the next chapter we present an extension of this solution formula to a certain class of scalar non-autonomous differential equations referred to as *separable* differential equations. The analytical solution formula does not generalize to an initial value problems for a system of differential equations, and is thus of very very limited use. However, the solution formula is really a beautiful application of Calculus, which may give valuable information in compact form in the special cases when it is applicable.

We also present a direct constructive proof of existence of a solution to the scalar autonomous problem, which generalizes to the very general case of a initial value problems for (autonomous and non-autonomous) systems of differential equations, as presented in Chapter *The general initial value problem* below.

210.2 An Analytical Solution Formula

To derive the analytical solution formula, we let $F(v)$ be a primitive function of the function $1/f(v)$, assuming v takes values so that zeros of $f(v)$ are avoided. Observe that here $F(v)$ is a primitive function of the function $1/f(v)$, and not of $f(v)$. We can then write the equation $u'(x) = f(u(x))$ as

$$\frac{d}{dx}F(u(x)) = 1,$$

since by the Chain rule $\frac{d}{dx}F(u(x)) = F'(u(x))u'(x) = \frac{u'(x)}{f(u(x))}$. We conclude that

$$F(u(x)) = x + C,$$

where the constant C is to be determined by the initial condition by setting $F(u_0) = C$ at $x = 0$. Formally, we can carry out the calculus as follows: We write the differential equation $\frac{du}{dx} = f(u)$ in the form

$$\frac{du}{f(u)} = dx$$

and integrate to get

$$F(u) = x + C,$$

which gives the solution formula

$$u(x) = F^{-1}(x + F(u_0)), \tag{210.3}$$

where F^{-1} is the inverse of F.

The Model $u' = u^n$ for $n > 1$

We use this example to show that the nonlinear nature of (210.1) allows the interesting behavior of *finite-time-blow-up* of the solution. First consider the case $n = 2$, that is, the initial value problem

$$u'(x) = u^2(x) \quad \text{for } x > 0, \quad u(0) = u_0 > 0, \tag{210.4}$$

with $f(v) = v^2$. In this case $F(v) = -1/v$ with $F^{-1}(w) = -1/w$, and we obtain the solution formula

$$u(x) = \frac{1}{u_0^{-1} - x} = \frac{u_0}{1 - u_0 x}.$$

We see that that $u(x) \to \infty$ as $x \to u_0^{-1}$, that is, the solution $u(x)$ of (210.1) with $f(u) = u^2$ tends to infinity as x increases to u_0^{-1} and the solution does not exist beyond this point, see Fig. 210.1. We say that the solution u *blows up* in finite time or exhibits *finite time blow-up*.

If we consider $u'(x) = u^2(x)$ as a model for the growth of a quantity $u(x)$ with time x in which the rate of growth is proportional to $u^2(x)$ and compare with the model $u'(x) = u(x)$ with solution $u_0 \exp(x)$ showing exponential growth. In the model $u'(x) = u^2(x)$ the growth is eventually much quicker than exponential growth since $u^2(x) > u(x)$ as soon as $u(x) > 1$.

We now generalize to

$$u'(x) = u^n(x) \quad \text{for } x > 0, \quad u(0) = u_0,$$

where $n > 1$. In this case $f(v) = v^{-n}$ and $F(v) = -\frac{1}{n-1} v^{-(n-1)}$, and we find the solution formula

$$u(x) = \frac{1}{(u_0^{-n+1} - (n-1)x)^{1/(n-1)}}.$$

Again the solution exhibits finite time blow-up.

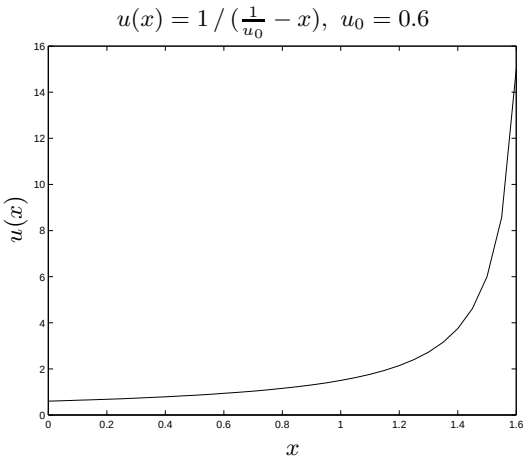


FIGURE 210.1. Solution of the equation $u' = u^2$

The Logistic Equation $u' = u(1-u)$

We now consider the initial value problem for the *logistic equation*

$$u'(x) = u(x)(1-u(x)) \quad \text{for } x > 0, \quad u(0) = u_0,$$

which was derived by the mathematician and biologist Verhulst as a model of a population with the *growth rate* decreasing with the factor $(1-u)$, as compared with the basic model $u' = u$, as the population approaches the value 1. Typically we assume $0 < u_0 < 1$ and expect to have $0 \le u(x) \le 1$.

In this case we have $f(u) = \frac{1}{u(1-u)}$ and using that $f(u) = \frac{1}{u} + \frac{1}{1-u}$, we find that

$$F(u) = \log(u) - \log(1-u) = \log(\frac{u}{1-u}),$$

so that

$$\log(\frac{u}{1-u}) = x + C,$$

or

$$\frac{u}{1-u} = \exp(C)\exp(x).$$

Solving for u and using the initial condition we find that

$$u(x) = \frac{1}{\frac{1-u_0}{u_0}\exp(-x) + 1}.$$

We see that the solution $u(x)$ increases from $u_0 < 1$ to 1 as x increases to infinity, see Fig. 210.2, which gives the famous logistic S-curve modeling growth with decreasing growth rate.

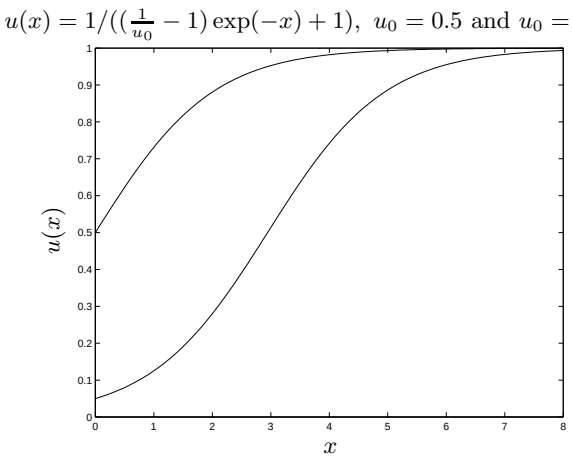


FIGURE 210.2. Solution of the logistic equation

210.3 Construction of the Solution

For the direct construction of a solution of (210.1), we shall use the same technique as that used for the linear problem $f(u(x)) = u(x)$ considered in Chapter *The exponential function*. Of course, one may ask why we should worry about constructing the solution, when we already have the solution formula (210.3). We may reply that the solution formula involves the (inverse of) the primitive function $F(v)$ of $1/f(v)$, which we may have to construct anyway, and then a direct construction of the solution may in fact be preferable. In general, a solution formula when available may give valuable information about qualitative properties of the solution such as dependence of parameters of the problem, even if it is not necessarily the most effective way of actually computing the solution.

To construct the solution we introduce meshes with nodes $x_i^n = ih_n$ for $i = 1, \cdots, N$, where $h_n = 2^{-n}$ and $N = 2^n$, and for $n = 1, 2, \cdots$. we then define an approximate continuous piecewise linear solution $U^n(x)$ for $0 < x \leq 1$ by the formula

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(U^n(x_{i-1}^n)) \quad \text{for } i = 1, \cdots, N, \tag{210.5}$$

with $U^n(0) = u_0$.

We want to prove that $\{U^n(x)\}$ is a Cauchy sequence for $x \in [0, 1]$ and we start by estimating $U^n(x_i^n) - U^{n+1}(x_i^n)$ for $i = 1, \cdots, N$. Taking two steps with step size $h_{n+1} = \frac{1}{2}h_n$ to go from time $x_{i-1}^n = x_{2i-2}^{n+1}$ to $x_i^n = x_{2i}^{n+1}$, we get

$$U^{n+1}(x_{2i-1}^{n+1}) = U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1})),$$
$$U^{n+1}(x_{2i}^{n+1}) = U^{n+1}(x_{2i-1}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-1}^{n+1})).$$

Inserting now the value of $U^{n+1}(x_{2i-1}^{n+1})$ at the intermediate step x_{2i-1}^{n+1} from the first equation into the second equation gives

$$U^{n+1}(x_{2i}^{n+1}) = U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1}f(U^{n+1}(x_{2i-2}^{n+1})) + h_{n+1}f\big(U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1}f(U^{n+1}(x_{2i-2}^{n+1}))\big). \quad (210.6)$$

Setting $e_i^n \equiv U^n(x_i^n) - U^{n+1}(x_{2i}^{n+1})$ and subtracting (210.6) from (210.5), we get

$$\begin{aligned} e_i^n &= e_{i-1}^n + h_n\big(f(U^n(x_{i-1}^n)) - f(U^{n+1}(x_{2i-2}^{n+1}))\big) \\ &\quad + h_{n+1}\Big(f(U^{n+1}(x_{2i-2}^{n+1})) - f\big(U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1}f(U^{n+1}(x_{2i-2}^{n+1}))\big)\Big) \\ &\equiv e_{i-1}^n + F_{1,n} + F_{2,n}, \end{aligned}$$

with the obvious definition of $F_{1,n}$ and $F_{2,n}$. Using the Lipschitz continuity and boundedness (210.2), we have

$$\begin{aligned} |F_{1,n}| &\le L_f h_n |e_{i-1}^n|, \\ |F_{2,n}| &\le L_f h_{n+1}^2 |f(U^{n+1}(x_{2i-2}^{n+1}))| \le L_f M_f h_{n+1}^2. \end{aligned}$$

Thus for $i = 1, \cdots, 2^N$,

$$|e_i^n| \le (1 + L_f h_n)|e_{i-1}^n| + L_f M_f h_{n+1}^2.$$

Iterating this inequality over i and using that $e_0^n = 0$, we get

$$|e_i^n| \le L_f M_f h_{n+1}^2 \sum_{k=0}^{i-1} (1 + L_f h_n)^k \quad \text{for } i = 1, \cdots, N.$$

Now recalling (203.10) and (203.27), we have

$$\sum_{k=0}^{i-1} (1 + L_f h_n)^k \le \frac{\exp(L_f) - 1}{L_f h_n},$$

and thus we have proved that for $i = 1, \cdots, N$,

$$|e_i^n| \le \frac{1}{2} M_f \exp(L_f) h_{n+1},$$

that is, for $\bar{x} = i h_n$ with $i = 0, ..., N$,

$$|U^n(\bar{x}) - U^{n+1}(\bar{x})| \le \frac{1}{2} M_f \exp(L_f) h_{n+1}.$$

Iterating this inequality as in the proof of the Fundamental Theorem, we get for $m > n$ and $\bar{x} = i h_n$ with $i = 0, ..., N$,

$$|U^n(\bar{x}) - U^m(\bar{x})| \le \frac{1}{2} M_f \exp(L_f) h_n.$$

Again as in the proof of the Fundamental Theorem, we conclude that $\{U^n(x)\}$ is a Cauchy sequence for each $x \in [0,1]$, and thus converges to a function $u(x)$, which by the construction satisfies the differential equation $u'(x) = f(u(x))$ for $x \in (0,1]$ and $u(0) = u_0$, and thus the limit $u(x)$ is a solution of the initial value problem (210.1).

It remains to prove uniqueness. Assume that $v(x)$ satisfies $v'(x) = f(v(x))$ for $x \in (0,1]$ and $v(0) = u_0$, and consider the function $w = u - v$. Since $w(0) = 0$,

$$|w(x)| = |\int_0^x w'(y)\,dy| = |\int_0^x f(u(y)) - f(v(y))\,dy|$$
$$\leq \int_0^x |f(u(y)) - f(v(y))|\,dy \leq \int_0^x L_f|w(y)|\,dy.$$

Setting $a = \max_{0\leq x\leq (2L_f)^{-1}} |w(x)|$, we have

$$a \leq \int_0^{(2L_f)^{-1}} L_f a\,dy \leq \frac{1}{2}a$$

which proves that $w(x) = 0$ for $0 \leq x \leq (2L_f)^{-1}$. We now repeat the argument for $x \geq (2L_f)^{-1}$ to get uniqueness for $0 \leq x \leq 1$.

We have now proved:

Theorem 210.1 *The initial value problem* (210.1) *with* $f : \mathbb{R} \to \mathbb{R}$ *Lipschitz continuous and bounded has a unique solution* $u : [0,1] \to \mathbb{R}$*, which is the limit of the sequence of continuous piecewise linear functions* $\{U^n(x)\}$ *constructed from* (210.5) *and satisfying* $|u(x) - U^n(x)| \leq \frac{1}{2}M_f \exp(L_f)h_n$ *for* $x \in [0,1]$.

The attentive reader will note that the existence proof does not seem to apply to e.g. the initial value problem (210.4), because the function $f(v) = v^2$ is not Lipschitz continuous and bounded on $\mathbb{R}$. In fact, the solution $u(x) = \frac{u_0}{1-u_0 x}$ only exists on the interval $[0, u_0^{-1}]$ and blows up at $x = u_0^{-1}$. However, we can argue that *before* blow-up with say $|u(x)| \leq M$ for some (large) constant M, it suffices to consider the function $f(v) = v^2$ on the interval $[-M, M]$ where the assumption of Lipschitz continuity and boundedness is satisfied. We conclude that for functions $f(v)$ which are Lipschitz continuous and bounded on bounded intervals of $\mathbb{R}$, the constructive existence proof applies as longs as the solution does not blow up.

Chapter 210 Problems

210.1. Solve the following initial value problem analytically: $u'(x) = f(u(x))$ for $x > 0$, $u(0) = u_0$, with (a) $f(u) = -u^2$, (b) $f(u) = \sqrt{u}$, (c) $f(u) = u\log(u)$, (d) $f(u) = 1 + u^2$, (e) $f(u) = \sin(u)$, (f) $f(u) = (1+u)^{-1}$, (g) $f(u) = \sqrt{u^2+4}$.

210.2. Verify that the constructed function $u(x)$ satisfies (210.1). Hint: Use that by the construction we have $u(x) = u_0 + \int_0^x f(u(y))\,dy$ for $x \in [0,1]$.

210.3. Find the velocity of a parachute jumper assuming that the air resistance is proportional to the square of the velocity.

210.4. Let $u(t)$ be the position of a body sliding along x-axis with the velocity $\dot{u}(t)$ satisfying $\dot{u}(t) = -\exp(-u)$. How long time does it take for the body to reach the position $u = 0$ starting from $u(0) = 5$

211

Separable Scalar Initial Value Problems

> The search for general methods for integrating ordinary differential equations ended about 1755. (Mathematical Thought, from Ancient to Modern Times, Kline)

211.1 Introduction

We now consider the initial value problem for a scalar non-autonomous differential equation:

$$u'(x) = f(u(x), x) \quad \text{for } 0 < x \leq 1,\ u(0) = u_0, \tag{211.1}$$

in the special case when $f(u(x), x)$ has the form

$$f(u(x), x) = \frac{h(x)}{g(u(x))}, \tag{211.2}$$

where $h : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$. We thus consider the initial value problem

$$u'(x) = \frac{h(x)}{g(u(x))} \quad \text{for } 0 < x \leq 1,\ u(0) = u_0, \tag{211.3}$$

where $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ are given functions, which we refer to as a *separable* problem, because the right hand side $f(u(x), x)$ separates into the quotient of one function $h(x)$ of x only, and one function $g(u(x))$ of $u(x)$ only according to (211.2).

211.2 An Analytical Solution Formula

We shall now derive an analytical solution formula that generalizes the solution formula (210.3) for a scalar autonomous problem (corresponding to the case $h(x) = 1$). Let then $G(v)$ and $H(x)$ be primitive functions of $g(v)$ and $h(x)$ so that $\frac{dG}{dv} = g$ and $\frac{dH}{dx} = h$, and suppose that the function $u(x)$ solves the equation

$$G(u(x)) = H(x) + C, \tag{211.4}$$

for $x \in [0, 1]$, where C is a constant. Differentiating with respect to x using the Chain rule on the left hand side, we then find that $g(u(x))u'(x) = h(x)$, that is $u(x)$ solves the differential equation $u'(x) = h(x)/g(u(x)) = f(u(x), x)$ as desired. Choosing the constant C so that $u(0) = u_0$, we thus obtain a solution $u(x)$ of (211.3), that is the problem (211.1) with $f(u(x), x)$ of the separable form (211.2).

Note that (211.4) is an algebraic equation for the value of the solution $u(x)$ for each value of x. We have thus rewritten the differential equation (211.3) as an algebraic equation (211.4) with x acting as a parameter, and involving primitive functions of $g(y)$ and $h(x)$.

Of course, we may consider (211.1) with x in an interval $[a, b]$ or $[a, \infty)$ with $a, b \in \mathbb{R}$.

EXAMPLE 211.1. Consider the separable initial value problem

$$u'(x) = xu(x), \quad x > 0, \; u(0) = u_0, \tag{211.5}$$

where $f(u(x), x) = h(x)/g(u(x))$ with $g(v) = 1/v$ and $h(x) = x$. The equation $G(u(x)) = H(x) + C$ takes the form

$$\log(u(x)) = \frac{x^2}{2} + C, \tag{211.6}$$

and thus the solution $u(x)$ of (211.5) is given by the formula

$$u(x) = \exp(\frac{x^2}{2} + C) = u_0 \exp(\frac{x^2}{2}),$$

with $\exp(C) = u_0$ chosen so that the initial condition $u(0) = u_0$ is satisfied. We check by differentiation using the Chain rule that indeed $u_0 \exp(\frac{x^2}{2})$ satisfies $u'(x) = xu(x)$ for $x > 0$.

Formally ("multiplying by dx"), we can rewrite (211.5) as

$$\frac{du}{u} = x\,dx$$

and integrate to get

$$\log(u) = \frac{x^2}{2} + C,$$

which corresponds to the equation (211.6).

EXAMPLE 211.2. On the rainy evening of November 11 1675 Leibniz successfully solved the following problem as a first (crucial) test of the power of the Calculus he had discovered on October 29: Find a curve $y = y(x)$ such that the *subnormal* p, see Fig. 211.1, is inversely proportional to y. Leibniz argued as follows: By similarity, see again Fig. 211.1, we have

$$\frac{dy}{dx} = \frac{p}{y},$$

and assuming the subnormal p to be inversely proportional to y, that is,

$$p = \frac{\alpha}{y}$$

with α a positive constant, we get the differential equation

$$\frac{dy}{dx} = \frac{\alpha}{y^2} = \frac{h(x)}{g(y)}, \tag{211.7}$$

which is separable with $h(x) = \alpha$ and $g(y) = y^2$. The solution $y = y(x)$ with $y(0) = 0$ thus is given by, see Fig. 211.1,

$$\frac{y^3}{3} = \alpha x, \quad \text{that is } y = (3\alpha x)^{\frac{1}{3}}, \tag{211.8}$$

The next morning Leibniz presented his solution to a stunned audience of colleagues in Paris, and rocketed to fame as a leading mathematician and Inventor of Calculus.

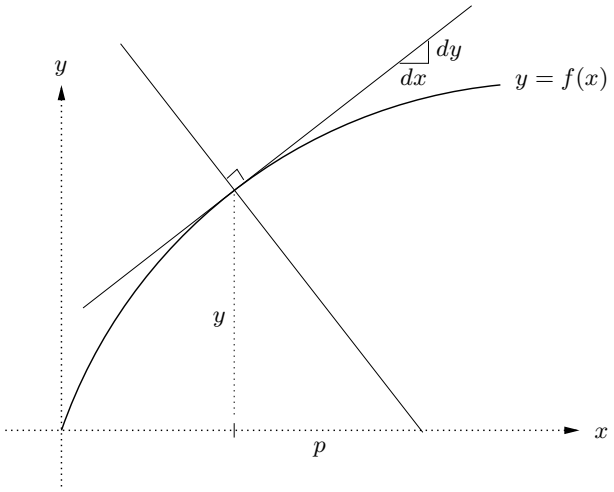


FIGURE 211.1. Leibniz' subnormal problem (change from $y = f(x)$ to $y = y(x)$)

211.3 Volterra-Lotka's Predator-Prey Model

We now consider a biological system consisting of prey and predators like rabbits and foxes which interact. Let $x(t)$ be the density of the prey and $y(t)$ that of the predators at time t and consider *Volterra-Lotka's predator-prey model* for their interaction:

$$\begin{aligned} \dot{x}(t) &= ax(t) - bx(t)y(t), \\ \dot{y}(t) &= -\alpha y(t) + \beta x(t)y(t) \end{aligned} \tag{211.9}$$

where a, b, α and β are positive constants, and $\dot{x} = \frac{dx}{dt}$ and $\dot{y} = \frac{dy}{dt}$. The model includes a growth term $ax(t)$ for the prey corresponding to births and a decay term $bx(t)y(t)$ proportional to the density of prey and predators corresponding to the consumption of prey by the predators, together with corresponding terms for the predators with different signs.

$(x(t), y(t))$ for $0<t<25.5$ with $(a, b, c, d)=(.5, 1, .2, 1)$, $(x_0, y_0)=(.5, .3)$

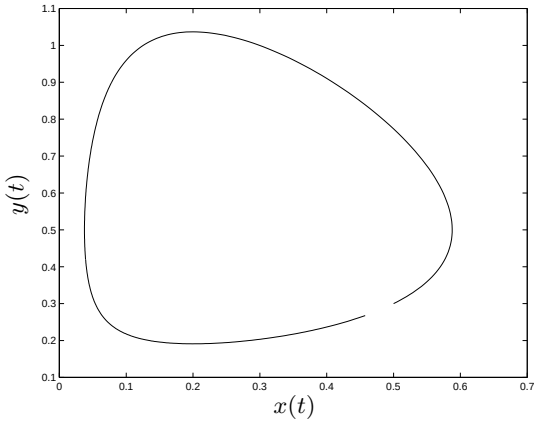


FIGURE 211.2. Phase plane plot of a solution of Volterra-Lotka's equation

This is a system of two differential equations in two unknowns $x(t)$ and $y(t)$ for which analytical solutions are unknown in general. However, we can derive an equation satisfied by the points $(x(t), y(t))$ in an $x-y$ plane, referred to as the $x-y$ *phase plane*, as follows: Dividing the two equations, we get

$$\frac{\dot{y}}{\dot{x}} = \frac{-\alpha y + \beta xy}{ax - bxy}$$

and formally replacing $\frac{\dot{y}}{\dot{x}}$ (by formally dividing out the common dt), we are led to the equation

$$y'(x) = \frac{-\alpha y + \beta xy}{ax - bxy} = \frac{y(-\alpha + \beta x)}{(a - by)x},$$

where $y' = \frac{dy}{dx}$, which is a separable equation with solution $y = y(x)$ satisfying

$$a\log(y) - by = -\alpha\log(x) + \beta x + C,$$

or

$$y^a \exp(-by) = \exp(C)x^{-\alpha}\exp(\beta x)$$

where C is a constant determined by the initial conditions. We plot pairs of (x, y) satisfying this equation in Fig. 211.2 as we let the prey x vary, which traces a *phase plane curve* of the solution $(x(t), y(t)$ of Fig. 211.2 as t varies. We see that the solution is periodic with a variation from (many rabbits, many foxes) to (few rabbits, many foxes) to (few rabbits, few foxes) to (many rabbits, few foxes) and back to (many rabbits, many foxes). Note that the phase plane curve shows the different combinations of rabbits and foxes (x, y), but does *not* give the time evolution $(x(t), y(t))$ of their interaction as a function of time t. We know that for a given t, the point $(x(t), y(t))$ lies on the phase plane curve, but not where.

211.4 A Generalization

We now consider a generalization of the separable differential equation (211.3) with solution $u(x)$ satisfying an equation of the form $G(u(x)) - H(x) = C$, to a differential equation with solution satisfying a more general equation of the form $F(x, u(x)) = C$. This closely couples to Chapter *Potential fields* below, and uses a generalization of the Chain rule, which can be accepted right now by a willing reader, and which we will meet again in Chapter *Vector-valued functions of several variables* below.

We thus consider the scalar initial value problem

$$u'(x) = f(u(x), x) \quad \text{for } 0 < x \leq 1,\ u(0) = u_0, \tag{211.10}$$

in the case $f(u(x), x)$ has the form

$$f(u(x), x) = \frac{h(u(x), x)}{g(u(x)), x)}, \tag{211.11}$$

where $h(v, x)$ and $g(v, x)$ are functions of v and x with the special property that

$$g(v, x) = \frac{\partial F}{\partial v}(v, x), \qquad h(v, x) = -\frac{\partial F}{\partial x}(v, x), \tag{211.12}$$

where $F(v, x)$ is a given function of v and x. Above we considered the case when $g(v, x) = g(v)$ is a function of v only and $h(v, x) = h(x)$ is a function of x only, and $F(v, x) = G(v) - H(x)$ with $G(v)$ and $H(x)$ primitive functions of $g(v)$ and $h(x)$, respectively. Now we allow $F(v, x)$ to have a more general form.

Assume now that $u(x)$ satisfies the equation

$$F(u(x), x) = C \qquad \text{for } 0 < x \leq 1.$$

Differentiating both sides with respect to x, using a generalization of the Chain rule, we then get

$$\frac{\partial F}{\partial u}\frac{du}{dx} + \frac{\partial F}{\partial x}\frac{dx}{dx} = g(x, u(x))u'(x) - h(x, u(x)) = 0,$$

and thus $u'(x)$ solves (211.10) with $f(u(x), x)$ of the form (211.11). Again, we thus have rewritten a differential equation as an algebraic equation $F(x, u(x)) = C$ with x acting as a parameter. We give an example. The reader can construct many other similar examples.

EXAMPLE 211.3. Let $F(v, x) = \frac{x^3}{3} + xv + \frac{v^3}{3}$ so that $g(v, x) = \frac{\partial F}{\partial v} = x + v^2$ and $h(v, x) = -\frac{\partial F}{\partial x} = -x^2 - v$. If $u(x)$ satisfies the algebraic equation $\frac{x^3}{3} + xu(x) + \frac{u^3(x)}{3} = C$ for $x \in [0, 1]$, then $u(x)$ solves the differential equation

$$u'(x) = -\frac{x^2 + u(x)}{x + u^2(x)} \qquad \text{for } 0 < x < 1.$$

To sum up: In this chapter we have given analytical solution formula for some special cases of the scalar initial value problem (211.1), but we were not able to give a solution formula in the case of a general non-autonomous scalar equation.

Chapter 211 Problems

211.1. Prove that solutions $(x(t), y(t)$ of the Volterra-Lotka model satisfies

$$\bar{x} = \frac{1}{T}\int_0^T x(t)\,dt = \frac{c}{d}, \quad \bar{y} = \frac{1}{T}\int_0^T y(t)\,dt = \frac{a}{b},$$

where T is the period of periodic solutions. Investigate the effect on the mean values $\bar{x}$ and $\bar{y}$ of hunting of both predator and prey corresponding to including dissipative terms $-\epsilon x$ and $-\epsilon y$ with $\epsilon > 0$. Hint: Consider the integral of $\dot{x}/x$ over a period.

211.2. Extend the Volterra-Lotka model to the model

$$\begin{aligned} \dot{x}(t) &= ax(t) - bx(t)y(t) - ex^2(t), \\ \dot{y}(t) &= -cy(t) + dx(t)y(t) - fy^2(t), \end{aligned} \tag{211.13}$$

where e and f are positive constants, with the additional terms modeling negative influences from competition within the species as the populations densities increase. Compare the solutions of the two models numerically. Is the extended system separable?

211.3. Consider the spread of an infection modeled by

$$\dot{u} = -auv,$$
$$\dot{v} = auv - bv,$$

where $u(t)$ is the density of the susceptibles and $v(t)$ is that of the infectives at time t, and a and b are positive constants. The term $\pm auv$ models the transfer of susceptibles to infectives at a rate proportional to auv, and $-bv$ models the decay of infectives by death or immunity. Study the qualitative behavior of phase plane curves.

211.4. Extend the previous model by changing the first equation to $\dot{u} = -auv + \mu$, with μ a positive constant modeling a constant growth of the susceptibles. Find the equilibrium point, and study the linearized model linearized at the equilibrium point.

211.5. Motivate the following model for a national economy:

$$\dot{u} = u - av, \quad \dot{v} = b(u - v - w),$$

where u is the national income, v the rate of consumer spending and w the rate of government spending, and $a > 0$ and $b \geq 1$ are constants. Show that if w is constant, then there is an equilibrium state, that is a solution independent of time satisfying $u - av = b(u - v - w) = 0$. Show that the economy oscillates if $b = 1$. Study the stability of solutions. Study a model with $w = w_0 + cu$ with w_0 a constant. Show that there is no equilibrium state in this model if $c \geq (a-1)/a$. Draw some conclusion. Study a model with $w = w_0 + cu^2$.

211.6. Consider a boat being rowed across a river occupying the strip $\{(x, y) : 0 \leq x \leq 1, y \in \mathbb{R}\}$, in such a way that the boat always points in the direction of $(0, 0)$. Assume that the boat moves with the constant speed u relative to the water and that the river flows with constant speed v in the positive y-direction. Show that the equations of motion are

$$\dot{x} = -\frac{ux}{\sqrt{x^2 + y^2}}, \ \dot{y} = -\frac{uy}{\sqrt{x^2 + y^2}}.$$

Show that the phase-plane curves are given by

$$y = \sqrt{x^2 + y^2} = Ax^{1-\alpha}, \quad \text{where } \alpha - \frac{v}{u}.$$

What happens if $v > u$?. Compute solutions.

212
The General Initial Value Problem

Things are as they are because they were as they were. (Thomas Gold)

212.1 Introduction

We now consider the Initial Value Problem or IVP for a *system* of nonlinear differential equations of the form: Find $u : [0,1] \to \mathbb{R}^d$ such that

$$u'(x) = f(u(x), x) \quad \text{for } 0 < x \leq 1,\ u(0) = u^0, \tag{212.1}$$

where $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is a given bounded and Lipschitz continuous function, $u^0 \in \mathbb{R}^d$ is a given initial value, and $d \geq 1$ is the dimension of the system. The reader may assume $d = 2$ or $d = 3$, recalling the chapters on analytic geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$, and extend to the case $d > 3$ after having read the chapter on analytic geometry in $\mathbb{R}^n$ below. The material in Chapter *Vector-valued functions of several real variables* is largely motivated from the need of studying problems of the form (212.1), and there is thus a close connection between this chapter and the present one. We keep this chapter abstract (and a bit philosophical), and present many concrete examples below. Note that for notational convenience we here use superscript index in the initial value u^0 (instead of u_0).

The IVP (212.1) is the non-autonomous vector version of the scalar initial value problem (210.1), and reads as follows in component form: Find

functions $u_i : [0,1] \to \mathbb{R}$, $i = 1, ..., d$, such that

$$\begin{aligned} u_1'(x) &= f_1(u_1(x), u_2(x), ..., u_d(x), x) \quad \text{for } 0 < x \leq 1, \\ u_2'(x) &= f_2(u_1(x), u_2(x), ..., u_d(x), x) \quad \text{for } 0 < x \leq 1, \\ &\cdots\cdots\cdots \\ u_d'(x) &= f_d(u_1(x), u_2(x), ..., u_d(x), x) \quad \text{for } 0 < x \leq 1, \\ u_1(0) &= u_{10},\ u_2(0) = u_{20},\ u_d(0) = u_d^0, \end{aligned} \tag{212.2}$$

where $f_i : \mathbb{R}^d \times [0,1] \to \mathbb{R}$, $i = 1, ..., d$, are given functions and u_i^0, $i = 1, ..., d$, are given initial values. With vector notation writing $u = (u_1, ..., u_d)$, $f = (f_1, ..., f_d)$ and $u^0 = (u_1^0, ..., u_d^0)$, we may write (212.2) in the compact form (212.1). Of course, writing $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$, means that for each vector $v = (v_1, ..., v_d) \in \mathbb{R}^d$ and $x \in [0,1]$ there is assigned a vector $f(v,x) = (f_1(v,x), ..., f_d(v,x)) \in \mathbb{R}^d$, where $f_i(v,x) = f_i(v_1, ..., v_d, x)$.

We assume Lipschitz continuity and boundedness of $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ in the form: There are constants L_f and M_f such that for all $v, w \in \mathbb{R}^d$ and $x, y \in [0,1]$,

$$|f(v,x) - f(w,y)| \leq L_f(|v - w| + |x - y|) \quad \text{and} \quad |f(v,x)| \leq M_f, \tag{212.3}$$

where $|v| = (\sum_{i=1}^d v_i^2)^{1/2}$ is the Euclidean norm of $v = (v_1, ..., v_d) \in \mathbb{R}^d$.

In short, everything looks the same as in the scalar case of (210.1) with the natural extension to a non-autonomous problem, but the vector interpretation makes the actual content of this chapter vastly different from that of Chapter *Scalar autonomous initial value problems*. In particular, there is in general no analytical solution formula if $d > 1$, since the solution formula for $d = 1$ based on the existence of a primitive function of $1/f(v)$, does not generalize to $d > 1$.

We prove the existence of a unique solution of the IVP (212.1) by using a constructive process which is a direct generalization of the method used for the scalar problem (210.1), which was a direct generalization of method used to construct the integral. The result of this chapter is definitely one of the highlights of mathematics (or at least of this book), because of its generality and simplicity: $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ can be *any* bounded Lipschitz continuous function with the dimension d arbitrarily large, and the proof looks exactly the same as in the scalar case. Therefore this chapter has a central role in the book and couples closely to several other chapters below including *Analytic geometry in* $\mathbb{R}^n$, *Solving systems of linear equations*, *Linearization and stability of IVP*, *Adaptive IVP-solvers*, *Vector-valued functions of several real variales* and various chapters on applications including mechanical systems, electrical circuits, chemical reactions, and other phenomena. This means that a full appreciation of this chapter can only be made after digesting all this material. Nevertheless, it should be possible to go through this chapter and understand that the general IVP (212.1) can

be solved through a constructive process requiring more or less work. This chapter also may be used as a basis for a bit of philosophical discussion on constructive aspects of the World, as we now proceed to do (for the interested reader).

212.2 Determinism and Materialism

Before we present the existence proof (which we thus have already seen), we pause to reflect a little on the related *mechanistic/deterministic* view of science and philosophy going back to Descartes and Newton and forming the basis the industrial society leading into our own time. With this view the World is like a big mechanical Clock governed by laws of mechanics, which may be modeled as an initial value problem of the form (212.1) with a certain function f and initial value u^0 at time $x = 0$. The state of this system for positive time is, according to the existence proof, uniquely determined by the function f and u^0, which would support a *deterministic* or *materialistic* view of the World including the mental processes in human beings: everything that will happen in the future is in principle determined by the present state (assuming no blow-up). Of course, this view is in serious conflict with massive everyday experience of unpredictability and our firm belief in the existence of a *free will*, and considerable efforts have gone into resolving this paradox through the centuries without complete success.

Let's see if we can approach this paradox from a mathematical point of view. After all, the determinitic/materialistic view is founded on a proof of existence of a unique solution of an initial value problem of the form (212.1), and thus the roots of the paradox may be hidden in the mathematical proof itself. We will argue that the resolution of the paradox must be coupled to aspects of *predictability* and *computability* of the problem (212.1), which we will now briefly touch upon and return to in more detail below. We hope the reader is open for this type of discussion, seldom met in a Calculus text. We try to point to the necessity of a proper understanding of a mathematical result, which may appear to be very simple and clear, like the existence proof to be presented, but which in fact may require a lot of explanation and qualification to avoid misunderstanding.

212.3 Predictability and Computability

The *predictability* of the problem (212.1) concerns the *sensitivity* of the solution to the given *data*, that is, the function f and the initial value u^0. The sensitivity is a measure of the change of the solution under changes of the data f and u^0. If the solution changes very much even for very small changes of data, then the sensitivity is very high. In such a case we

need to know the data with very high precision to accurately predict the solution. We shall see below that solutions of certain initial value problems are highly sensitive to changes in data and in these problems accurate prediction will be impossible. An example is given by the familiar process of tossing a coin, which can be modeled as an initial value problem. In principle, by repeatedly choosing the same initial value, the person tossing the coin should be able to always get heads, for example. However, we all know that this is impossible in practice, because the process is too sensitive to small changes in the initial value (and the corresponding function f). To handle this type of unpredictability the scientific field of *statistics* has been developed.

Similarly, the *computability* of the problem (212.1) concerns (i) the sensitivity of the solution to errors made in constructing the solution according to the existence proof, and (ii) the amount of computational work needed to construct the solution. Usually, (i) and (ii) go hand in hand: if the sensitivity is high, then a lot of work is required and of course the work also increases with the dimension d. A highly sensitive problem with d very large is thus a computational night-mare. To construct the solution of the initial value problem for even a small part of the Universe will thus be practically impossible with any kind of computer, and claiming that in principle the solution is determined would make little sense.

We will meet this problem with painstaking evidence when we turn into numerical methods. We will see that most systems of the form (212.1) with d small ($d \leq 10$ say) may be solved within fractions of a second on a PC, while some systems (like the famous Lorenz system with $d = 3$ to be studied below) quickly will exhaust even supercomputers because of very high sensitivity. We will further see that many systems of practical interest with d large ($d \approx 10^6 - 10^7$) can be solved within minutes/hours on a PC, while accurate modeling of e.g. turbulent flow requires $d \geq 10^{10}$ and super-computer power. The most powerful super-computer in sight, the *Blue Gene* consisting of 10^6 connected PCs to appear in a couple of years, is designed for initial value problems of molecular dynamics of protein folding for the purpose of medical drug design. A landmark in computing was set in 1997 when the chess computer *Deep Blue* put the the world-champion Gary Kasparov chess mate.

The computational work required to solve (212.1) may thus vary considerably. Below we shall successively uncover a bit of this mystery and identify basic features of problems requiring different amounts of computational work.

We will return to the concepts of predictability and computability of differential equations below. Here we just wanted to give some perspective on the constructive existence proof to be given showing some limits of mathematics as a human activity.

212.4 Construction of the Solution

The construction of the solution $u(x)$ of (212.1) looks identical to the construction of the solution of (210.1), after we interpret $u(x)$ and $f(u(x))$ as vectors instead of scalars and make the natural extension to a non-autonomous problem.

We begin by discretizing $[0,1]$ using a mesh with nodes $x_i^n = ih_n$ for $i = 1, \cdots, N$, where $h_n = 2^{-n}$ and $N = 2^n$. For $n = 1, \cdots, N$, we define an approximate piecewise linear solution $U^n : [0,1] \to \mathbb{R}^d$ by the formula

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(U^n(x_{i-1}^n), x_{i-1}^n), \quad \text{for } i = 1, \cdots, N, \quad (212.4)$$

and setting $U^n(0) = u^0$. Note that $U^n(x)$ is linear on each subinterval $[x_{n-1}^n, x_i^n]$.

We want to prove that for $x \in [0,1]$, $\{U^n(x)\}_{n=1}^{\infty}$ is a Cauchy sequence in $\mathbb{R}^d$. We start by estimating $U^n(x_i^n) - U^{n+1}(x_i^n)$ for $i = 1, \cdots, N$. Taking two steps with step size $h_{n+1} = \frac{1}{2}h_n$ to go from time $x_{i-1}^n = x_{2i-2}^{n+1}$ to $x_i^n = x_{2i}^{n+1}$, we have

$$\begin{aligned} U^{n+1}(x_{2i-1}^{n+1}) &= U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n), \\ U^{n+1}(x_{2i}^{n+1}) &= U^{n+1}(x_{2i-1}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-1}^{n+1}), x_{2i-1}^{n+1}). \end{aligned}$$

Inserting now the value of $U^{n+1}(x_{2i-1}^{n+1})$ at the intermediate step x_{2i-1}^{n+1} from the first equation into the second equation, we get

$$\begin{aligned} U^{n+1}(x_{2i}^{n+1}) = U^{n+1}(x_{2i-2}^{n+1}) &+ h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n) \\ &+ h_{n+1} f\big(U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n), x_{2i-1}^{n+1}\big). \end{aligned} \quad (212.5)$$

Setting $e_i^n \equiv U^n(x_i^n) - U^{n+1}(x_{2i}^{n+1})$ and subtracting (212.5) from (212.4) gives

$$\begin{aligned} e_i^n = e_{i-1}^n &+ h_n \big(f(U^n(x_{i-1}^n), x_{i-1}^n) - f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n)\big) \\ &+ h_{n+1} \Big(f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n) - f\big(U^{n+1}(x_{2i-2}^{n+1}) \\ &+ h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n), x_{2i-1}^{n+1}\big) \Big) \equiv e_{i-1}^n + F_{1,n} + F_{2,n}, \end{aligned}$$

with the obvious definitions of $F_{1,n}$ and $F_{2,n}$. Using (212.3), we have

$$\begin{aligned} &|F_{1,n}| \le L_f h_n |e_{i-1}^n|, \\ &|F_{2,n}| \le L_f h_{n+1}^2 (|f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n)| + 1) \le L_f \bar{M}_f h_{n+1}^2, \end{aligned}$$

where $\bar{M}_f = M_f + 1$, and so for $i = 1, \cdots, N$,

$$|e_i^n| \le (1 + L_f h_n)|e_{i-1}^n| + L_f \bar{M}_f h_{n+1}^2.$$

Iterating this inequality over i and using that $e_0^n = 0$, we get

$$|e_i^n| \le L_f \bar{M}_f h_{n+1}^2 \sum_{k=0}^{i-1} (1 + L_f h_n)^i \quad \text{for } i = 1, \cdots, N.$$

Recalling (203.10) and (203.27), we have

$$\sum_{k=0}^{i-1} (1 + L_f h_n)^k = \frac{(1 + L_f h_n)^i - 1}{L_f h_n} \le \frac{\exp(L_f) - 1}{L_f h_n},$$

and thus we have proved that for $i = 1, \cdots, N$,

$$|e_i^n| \le \frac{1}{2} \bar{M}_f \exp(L_f) h_{n+1},$$

that is, for $\bar{x} = i h_n$ with $i = 0, ..., N$,

$$|U^n(\bar{x}) - U^{n+1}(\bar{x})| \le \frac{1}{2} \bar{M}_f \exp(L_f) h_{n+1}.$$

Iterating this inequality as in the proof of the Fundamental Theorem, we get for $m > n$ and $\bar{x} = i h_n$ with $i = 0, ..., N$,

$$|U^n(\bar{x}) - U^m(\bar{x})| \le \frac{1}{2} \bar{M}_f \exp(L_f) h_n. \tag{212.6}$$

Again as in the proof of the Fundamental Theorem, we conclude that $\{U^n(x)\}$ is a Cauchy sequence for each $x \in [0, 1]$, and thus converges to a function $u(x)$, which by the construction satisfies the differential equation $u'(x) = f(u(x))$ for $x \in (0, 1]$ and $u(0) = u^0$, and thus the limit $u(x)$ is a solution of the initial value problem (212.1). Uniqueness of a solution follows as in the scalar case considered in Chapter *Scalar autonomous initial value problems*. We have now proved the following basic result:

Theorem 212.1 *The initial value problem* (212.1) *with* $f : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ *bounded and Lipschitz continuous, has a unique solution* $u(x)$*, which is the limit of the sequence of continuous piecewise linear functions* $\{U^n(x)\}$ *constructed from* (212.4) *and satisfying*

$$|u(x) - U^n(x)| \le (M_f + 1) \exp(L_f) h_n \qquad \text{for } x \in [0, 1]. \tag{212.7}$$

212.5 Computational Work

The convergence estimate (212.7) indicates that the work required to compute a solution $u(x)$ of (212.1) to a given accuracy is proportional to $\exp(L_f)$ and to $\exp(L_f T)$ if we consider a time interval $[0, T]$ instead of

$[0,1]$. With $L_f = 10$ and $T = 10$, which would seem to be a very innocent case, we would have $\exp(L_f T) = \exp(10^2)$ and we would thus have to choose h_n smaller than $\exp(-10^2) \approx 10^{-30}$, and the number of computational operations would be of the order 10^{30} which would be at the limit of any practical possibility. Already moderately large constants such as $L_f = 100$ and $T = 100$, would give an exponential factor $\exp(10^4)$ way beyond any comprehension. We conclude that the appearance of the exponential factor $\exp(L_f T)$, which corresponds to a worst possible case, seems to limit the interest of the existence proof. Of course, the worst possible case does not necessarily have to occur always. Below we will present problems with special features for which the error is actually smaller than worst possible, including the important class of *stiff problems* where large Lipschitz constants cause quick exponential decay instead of exponential growth, and the Lorenz system where the error growth turns out to be of order $\exp(T)$ instead of $\exp(L_f T)$ with $L_f = 100$.

212.6 Extension to Second Order Initial Value Problems

Consider a second order initial value problem

$$\ddot{v}(t) = g(v(t), \dot{v}(t)) \text{ for } 0 < t \leq 1,\ v(0) = v_0,\ \dot{v}(0) = \dot{v}_0, \tag{212.8}$$

with initial conditions for $v(0)$ and $\dot{v}(0)$, where $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous, $v : [0,1] \to \mathbb{R}^d$ and $\dot{v} = \frac{dv}{dt}$. In mechanics, initial value problems often come in such second order form as they express Newton's Law with $\ddot{v}(t)$ representing acceleration and $g(v(t), \dot{v}(t))$ force. This problem can be reduced to a first order system of the form (212.1) by introducing the new variable $w(t) = \dot{v}(t)$ and writing (212.8) as

$$\begin{aligned} \dot{w}(t) &= g(v(t), w(t)) \quad \text{for } 0 < t \leq 1, \\ \dot{v}(t) &= w(t) \quad \text{for } 0 < t \leq 1, \\ v(0) &= v_0,\ w(0) = \dot{v}_0. \end{aligned} \tag{212.9}$$

Setting $u = (u_1, ..., u_{2d}) = (v_1, ..., v_d, w_1, ..., w_d)$ and $f(u) = (g_1(u), \cdots, g_d(u), u_{d+1}, ..., u_{2d})$, the system (212.9) takes the form $\dot{u}(t) = f(u(t))$ for $0 < t \leq 1$, and $u(0) = (v_0, \dot{v}_0)$.

In particular, we can rewrite the second order scalar equations $\ddot{v} + v = 0$ as a first order system and obtain existence of the trigonometric functions via the general existence result for first order systems as solutions of the corresponding initial value problem with appropriate data.

212.7 Numerical Methods

The computational solution of differential equations is an important subject with many aspects. The overall objective may be viewed to be to compute approximate solutions with as little work as possible per digit of accuracy. So far we have discussed only the simplest method for constructing approximate solutions. In this section, we give a brief glimpse of other methods. In Chapter *Adaptive IVP solvers*, we continue this study.

The computational method we have used so far, in which

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(U^n(x_{i-1}^n), x_{i-1}^n), \quad \text{for } i = 1, \cdots, N, \quad (212.10)$$

with $U^n(0) = u^0$, is called the *forward Euler* method. The forward Euler method is an *explicit* method because we can directly compute $U^n(x_i^n)$ from $U^n(x_{i-1}^n)$ without solving a system of equations.

In contrast, the *backward Euler* method in which the approximate solution is computed via the equation

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(U^n(x_i^n), x_i^n), \quad \text{for } i = 1, \cdots, N, \quad (212.11)$$

with $U^n(0) = u^0$, is an *implicit* method. At each step we need to solve the system

$$V = U^n(x_{i-1}^n) + h_n f(V, x_i^n), \quad (212.12)$$

to compute $U^n(x_i^n)$ from $U^n(x_{i-1}^n)$. Another implicit method is the *midpoint method*

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(\frac{1}{2}(U^n(x_{i-1}^n) + U^n(x_i^n)), \bar{x}_{i-1}^n), \quad i = 1, \cdots, N, \quad (212.13)$$

with $\bar{x}_{i-1}^n = \frac{1}{2}(x_{i-1}^n + x_i^n)$, where we have to solve the system

$$V = U^n(x_{i-1}^n) + h_n f(\frac{1}{2}(U^n(x_{i-1}^n) + V), \bar{x}_{i-1}^n) \quad (212.14)$$

at each step. Note that both (212.12) and (212.14) are nonlinear equations when f is nonlinear. We may use Fixed Point Iteration or Newton's method to solve them, see Chapter *Vector-valued functions of several real variables* below.

We also present the following variant of the midpoint method, which we call the cG(1), *continuous Galerkin method with trial functions of order* 1:. The approximate solution is computed via

$$U^n(x_i^n) = U^n(x_{i-1}^n) + \int_{x_{i-1}^n}^{x_i^n} f(U(x), x)\, dx, \quad i = 1, \cdots, N, \quad (212.15)$$

and $U^n(0) = u^0$, where $U^n(x)$ is continuous piecewise linear function with the values $U^n(x_i^n)$ at the nodes x_i^n. If we evaluate the integral in (212.15)

with the midpoint quadrature rule, we obtain the midpoint method. We can of course use other quadrature formulas to get different methods.

We shall see that cG(1) is the first in a family of methods cG(q) with $q = 1, 2, ...$, where the solution is approximated by continuous piecewise polynomials of order q. The Galerkin feature of cG(1) is expressed by the fact that the method can be formulated as

$$\int_{x_{i-1}^n}^{x_i^n} (\frac{dU^n}{dx}(x) - f(U^n(x), x))\, dx = 0,$$

stating that the mean-value over each subinterval of the *residual* $\frac{dU^n}{dx}(x) - f(U(x), x)$ of the continuous piecewise linear approximate solution $U^n(x)$, is equal to zero (or that the residual is orthogonal to the set of constant functions on each subinterval with a terminology to be used below).

We can prove convergence of the backward Euler and midpoint methods in the same way as for the forward Euler method. The forward and backward Euler methods are *first order accurate* methods in the sense that the error $|u(x) - U^n(x)|$ is proportional to the step size h_n, while the midpoint method is *second order accurate* with the error proportional to h_n^2 and thus in general is more accurate. The computational work per step is generally smaller for an explicit method than for an implicit method, since no system of equations has to be solved at each step. For so-called stiff problems, explicit methods may require very small time steps compared to implicit methods, and then implicit methods can give a smaller total cost. We will return to these issues in Chapter *Adaptive IVP solvers* below.

Note that all of the methods discussed so far generalize to allow non-uniform meshes $0 = x_0 < x_1 < x_2 < ... < x_N = 1$ with possibly varying steps $x_i - x_{i-1}$. We will below return to the problem of *automatic step-size control* with the purpose of keeping the error $|u(x_i) - U(x_i)| \leq TOL$ for $i = 1, \cdots, N$, where TOL is a given tolerance, while using as few time steps as possible by varying the mesh steps, cf. the Chapter *Numerical Quadrature*.

Chapter 212 Problems

212.1. Prove existence of a solution of the initial value problem (212.1) using the backward Euler method or the midpoint method.

212.2. Complete the proof of existence for (212.1) by proving that the constructed limit function $u(x)$ solves the initial value problem. Hint: use that $u_i(x) = \int_0^x f_i(u(y))\, dy$ for $x \in [0, 1]$, $i = 1, ..., d$.

212.3. Give examples of problems of the form (212.1).

213

Lagrange and the Principle of Least Action*

Dans les modifications des mouvements, l'action devient ordinairement un Maximum ou un Minimum. (Leibniz)

Whenever any action occurs in nature, the quantity of action employed by this change is the least possible. (Maupertuis 1746)

From my earliest recollection I have had an irresistible liking for mechanics and the physical laws on which mechanics as a science is based. (Reynolds)

213.1 Introduction

Lagrange (1736-1813), see Fig. 213.1, found a way to formulate certain dynamical problems in mechanics using a *Principle of Least Action*. This principle states that the *state* $u(t)$ of a system changes with time t over a given time interval $[t_1, t_2]$, so that the *action integral*

$$I(u) = \int_{t_1}^{t_2} (T(\dot{u}(t)) - V(u(t))\, dt \tag{213.1}$$

is *stationary*, where $T(\dot{u}(t))$ with $\dot{u} = \frac{du}{dt}$ is the *kinetic energy*, and $V(u(t))$ is the *potential energy* of the *state* $u(t)$. We here assume that the state $u(t)$ is a function $u : [t_1, t_2] \to \mathbb{R}$ satisfying $u(t_1) = u_1$ and $u(t_2) = u_2$, where u_1 and u_2 are given initial and final values. For example, $u(t)$ may be the position of a moving mass at time t. The action integral of a state is thus

the difference between the kinetic and potential energies integrated in time along the state.

We shall now get acquainted with Lagrange's famous Principle of Least Action and we shall see that it may be interpreted as a reformulation of Newton's law stating that mass times acceleration equals force. To this end, we first need to explain what is meant by the statement that the *action integral is stationary* for the actual solution $u(t)$. Our tool is Calculus, at its best!

FIGURE 213.1. Lagrange, Inventor of the Principle of Least Action: "I regard as quite useless the reading of large treatises of pure analysis: too large a number of methods pass at once before the eyes. It is in the works of applications that one must study them; one judges their ability there and one apprises the manner of making use of them".

Following in the foot-steps of Lagrange, consider a *perturbation* $v(t) = u(t)+\epsilon w(t) = (u+\epsilon w)(t)$ of the state $u(t)$, where $w(t)$ is a function on $[t_1, t_2]$ satisfying $w(t_1) = w(t_2) = 0$ and ϵ is a small parameter. The function $v(t)$ corresponds to changing $u(t)$ with the function $\epsilon w(t)$ inside (t_1, t_2) while keeping the values $v(t_1) = u_1$ and $v(t_2) = u_2$. The Principle of Least Action states that the actual path $u(t)$ has the property that for all such functions $w(t)$, we have

$$\frac{d}{d\epsilon} I(u + \epsilon w) = 0 \quad \text{for } \epsilon = 0. \tag{213.2}$$

The derivative $\frac{d}{d\epsilon} I(u + \epsilon w)$ at $\epsilon = 0$, measures the rate of change with respect to ϵ at $\epsilon = 0$ of the value of the action integral with $u(t)$ replaced by $v(t) = u(t) + \epsilon w(t)$. The Principle of Least Action says this rate of change is zero if u is the actual solution, which expresses the stationarity of the action integral.

We now present a couple of basic applications illustrating the use of the Principle of Least Action.

213.2 A Mass-Spring System

We consider a system of a mass m sliding on a on a horizontal friction-less x-axis and being connected to the origin with a weight-less Hookean spring with spring constant k, see the Chapter Galileo, Newton et al. We know that this system may be described by the equation $m\ddot{u} + ku = 0$, where $u(t)$ is the length of the spring at time t. We derive this model by using the Principle of Least Action. In this case,

$$T(\dot{u}(t)) = \frac{m}{2}\dot{u}^2(t) \quad \text{and} \quad V(u(t)) = \frac{k}{2}u^2(t),$$

and thus

$$I(u) = \int_{t_1}^{t_2} \big(\frac{m}{2}\dot{u}^2(t) - \frac{k}{2}u^2(t)\big)\, dt.$$

To motivate the expression $V(u(t)) = \frac{k}{2}u^2(t)$ for the potential energy, we use the definition of the potential energy as the total work required to move the mass from position $u = 0$ to position $u(t)$. The work to move the mass from position v to $v + \Delta v$ is equal to $kv\Delta v$ following the principle that work = force $\times$ displacement. The total work is thus

$$V(u(t)) = \int_0^{u(t)} kv\, dv = \frac{k}{2}u^2(t),$$

as announced.

To see how the equation $m\ddot{u}+ku = 0$ arises, we compute the derivative of $I(u+\epsilon w)$ with respect to ϵ and then set $\epsilon = 0$, where $w(x)$ is a perturbation satisfying $w(t_1) = w(t_2) = 0$. Direct computation based on moving $\frac{d}{d\epsilon}$ inside the integral, which is allowed since the limits of integration are fixed,

$$\frac{d}{d\epsilon}I(u + \epsilon w) =$$
$$\int_{t_1}^{t_2} \frac{d}{d\epsilon}\big(\frac{m}{2}\dot{u}\dot{u} + \epsilon m\dot{u}\dot{w} + \frac{m}{2}\epsilon^2\dot{w}\dot{w} - \frac{k}{2}u^2 - k\epsilon uw - \frac{k}{2}\epsilon^2 w^2\big)\, dt$$
$$= \int_{t_1}^{t_2} \big(m\dot{u}\dot{w} - kuw\big)\, dt \quad \text{for } \epsilon = 0.$$

Integrating by parts in the term $m\dot{u}\dot{w}$, we get

$$\int_{t_1}^{t_2} \big(m\ddot{u} + ku\big)w\, dt = 0,$$

for all $w(t)$ with $w(t_1) = w(t_2) = 0$. This implies that $m\ddot{u} + ku = 0$ in $[t_1, t_2]$, since $w(t)$ can vary arbitrarily in the interval (t_1, t_2), and we obtain the desired equation.

213.3 A Pendulum with Fixed Support

We consider a pendulum in the form of a body of mass one attached to a weightless string of unit length fixed to the ceiling under the action of a vertical gravity force normalized to one. The action integral of the difference between kinetic and potential energy is given by

$$I(u) = \int_{t_1}^{t_2} \big(\frac{1}{2}\dot{u}^2(t) - (1-\cos(u(t))\big)\,dt,$$

where $u(t)$ represents the angle of the pendulum in radians at time t, measured from the vertical position, see Fig. 213.2.

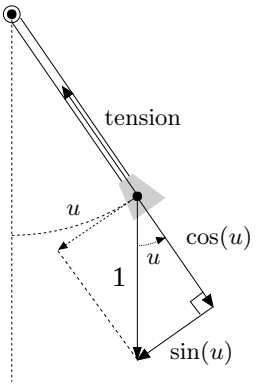


FIGURE 213.2. A pendulum.

The potential energy in this case is equal to the work of lifting the mass from the bottom position to the level $(1-\cos(v))$, which is exactly equal to $(1-\cos(v))$ if the gravitational constant is normalized to one. Stationarity of the action integral requires that for all perturbations $w(t)$ satisfying $w(t_1) = w(t_2) = 0$, we have

$$0 = \frac{d}{d\epsilon}\int_{t_1}^{t_2} \big(\frac{1}{2}(\dot{u}+\epsilon\dot{w})^2(t) - (1-\cos(u(t)+\epsilon w(t))\big)\,dt \quad \text{for } \epsilon = 0,$$

which gives as above

$$\int_{t_1}^{t_2} \big(\ddot{u} + \sin(u(t))\big)w\,dt = 0.$$

This yields the initial value problem

$$\begin{cases} \ddot{u} + \sin(u) = 0 & \text{for } t > 0 \\ u(0) = u_0,\ \dot{u}(0) = u_1, \end{cases} \tag{213.3}$$

where we added initial conditions for position and velocity.

The resulting differential equation $\ddot{u} = -\sin(u)$ is an expression of Newton's Law, since $\ddot{u}$ is the angular acceleration and $-\sin(u)$ is the restoring force in the angular direction. We conclude that the Principle of Least Action in the present case is a reformulation of Newton's Law.

If the angle of the pendulum stays small during the motion, then we can approximate $\sin(u)$ by u and obtain the linear equation $\ddot{u} + u = 0$, with solutions being linear combinations of $\sin(t)$ and $\cos(t)$.

213.4 A Pendulum with Moving Support

We now generalize to a pendulum with a support that is subject to a prescribed motion. Consider thus a body of mass m swinging in a weightless string of length l that is attached to a support moving according to a given function $r(t) = (r_1(t), r_2(t))$ in a coordinate system with the x_1-axis horizontal and the x_2-axis vertical upward. Let $u(t)$ be the angle of the string at time t measured from the vertical.

The potential energy is again equal to the height of the body, measured from some reference position, times mg with g the gravitational constant. Thus, we may choose

$$V(u(t)) = mg(r_2(t) - l\cos(u)).$$

To express the kinetic energy, we need to take into account the motion of the support. The velocity of the body relative to the support is $(l\dot{u}\cos u, l\dot{u}\sin u)$, and the total velocity is thus $(\dot{r}_1(t) + l\dot{u}\cos u, \dot{r}_2(t) + l\dot{u}\sin u)$. The kinetic energy is $m/2$ times the square of the *modulus of the velocity*, and thus

$$T = \frac{m}{2}\left[(\dot{r}_1 + l\dot{u}\cos u)^2 + (\dot{r}_1 + l\dot{u}\sin u)^2\right].$$

Using the Principle of Least Action, we obtain the following equation:

$$\ddot{u} + \frac{g}{l}\sin u + \frac{\ddot{r}_1}{l}\cos u + \frac{\ddot{r}_2}{l}\sin u = 0, \tag{213.4}$$

together with initial conditions for $u(0)$ and $\dot{u}(0)$.

If the support is fixed with $\ddot{r}_1 = \ddot{r}_2 = 0$, then we recover the equation (213.3) setting $l = m = g = 1$.

213.5 The Principle of Least Action

We now consider a mechanical system that is described by a vector function $u(t) = (u_1(t), u_2(t))$. We may think of a system consisting of two bodies

with positions given by the functions $u_1(t)$ and $u_2(t)$. The action integral is

$$I(u_1, u_2) = I(u) = \int_{t_1}^{t_2} L(u(t)\, dt,$$

where

$$L(u_1(t), u_2(t)) = L(u(t)) = T(\dot{u}(t)) - V(u(t))$$

is the difference of the kinetic energy $T(\dot{u}(t)) = T(\dot{u}_1(t), \dot{u}_2(t))$ and the potential energy $V(u(t)) = V(u_1(t), u_2(t))$. We refer to $L(u(t))$ as the *Lagrangean* of the state $u(t)$.

The Principle of Least Action states that the action integral is stationary at the true state $u(t)$ in the sense that for all perturbations $w_1(t)$ and $w_2(t)$ with $w_1(t_1) = w_2(t_1) = w_1(t_2) = w_2(t_2)$, we have for $\epsilon = 0$,

$$\frac{d}{d\epsilon} I(u_1 + \epsilon w_1, u_2) = 0$$
$$\frac{d}{d\epsilon} I(u_1, u_2 + \epsilon w_2) = 0.$$

Assuming that

$$T(\dot{u}_1(t), \dot{u}_2(t)) = \frac{m_1}{2}\dot{u}_1^2(t) + \frac{m_2}{2}\dot{u}_2^2(t),$$

we obtain performing the differentiation with respect to ϵ as above and setting $\epsilon = 0$,

$$\int_{t_1}^{t_2} \big(m\dot{u}_1(t)\dot{w}_1(t) - \frac{\partial V}{\partial u_1}(u_1(t), u_2(t))w_1(t)\big)\, dt = 0,$$
$$\int_{t_1}^{t_2} \big(m\dot{u}_2(t)\dot{w}_2(t) - \frac{\partial V}{\partial u_2}(u_1(t), u_2(t))w_2(t)\big)\, dt = 0.$$

Integrating by parts as above and letting w_1 and w_2 vary freely over (t_1, t_2), we obtain

$$\begin{aligned} m\ddot{u}_1(t) &= -\frac{\partial V}{\partial u_1}(u_1(t), u_2(t)), \\ m\ddot{u}_2(t) &= -\frac{\partial V}{\partial u_2}(u_1(t), u_2(t)). \end{aligned} \tag{213.5}$$

If we set

$$F_1 = -\frac{\partial V}{\partial u_1}, \quad F_2 = -\frac{\partial V}{\partial u_2},$$

then we can write the equations derived from the Principle of Least Action as

$$\begin{aligned} m\ddot{u}_1(t) &= F_1(u_1(t), u_2(t)), \\ m\ddot{u}_2(t) &= F_1(u_1(t), u_2(t)), \end{aligned} \tag{213.6}$$

which can be viewed as Newton's Law if F_1 and F_2 are interpreted as forces.

213.6 Conservation of the Total Energy

Defining the *total energy*

$$E(u(t)) = T(\dot{u}(t)) + V(u(t))$$

as the sum of the kinetic and potential energies and using (213.5), we get

$$\begin{aligned}\frac{d}{dt}E(u(t)) &= m_1\dot{u}_1\ddot{u}_1 + m_2\dot{u}_2\ddot{u}_2 + \frac{\partial V}{\partial u_1}\dot{u}_1 + \frac{\partial V}{\partial u_2}\dot{u}_2 \\ &= \dot{u}_1\big(m_1\ddot{u}_1 + \frac{\partial V}{\partial u_1}\big) + \dot{u}_2\big(m_2\ddot{u}_2 + \frac{\partial V}{\partial u_2}\big) = 0.\end{aligned}$$

We conclude that the total energy $E(u(t))$ is constant in time, that is the energy is *conserved*. Obviously, energy conservation is not a property of all systems, and thus the Principle of Least Action only applies to so called *conservative systems*, where the total energy is conserved. In particular, effects of *friction* are not present.

213.7 The Double Pendulum

We now consider a *double pendulum* consisting of two bodies of masses m_1 and m_2, where the first body of mass m_1 hangs on a weightless string of length l_1 attached to a fixed support and the second body of mass m_2 hangs on a weightless string of length l_2 attached to the first body. We shall now apply the Principle of Least Action to derive the equations of motion for this system.

To describe the state of the system, we use the angles $u_1(t)$ and $u_2(t)$ of the two bodies measured from vertical position.

We now seek expressions for the kinetic and potential energies of the system of the two bodies. The contributions from the second body is obtained from the expressions for a pendulum with moving support derived above if we set $(r_1(t), r_2(t)) = (l_1 \sin u_1, -l_1 \cos u_1)$.

The potential energy of the first pendulum is $-mgl_1 \cos u_1$ and the total potential energy is

$$V(u_1(t), u_2(t)) = -m_1 g l_1 \cos u_1(t) - m_2 g \left(l_1 \cos u_1(t) + l_2 \cos u_2(t)\right).$$

The total kinetic energy is obtained similarly adding the kinetic energies of the two bodies:

$$\begin{aligned}T(\dot{u}_1(t), \dot{u}_2(t)) = \frac{m_1}{2} l_1^2 \dot{u}_1^2 + \frac{m_2}{2}\big[&(l_1\dot{u}_1 \cos u_1 + l_2 \dot{u}_2 \cos u_2)^2 \\ &+ (l_1 \dot{u}_1 \sin u_1 + l_2 \dot{u}_2 \sin u_2)^2\big].\end{aligned}$$

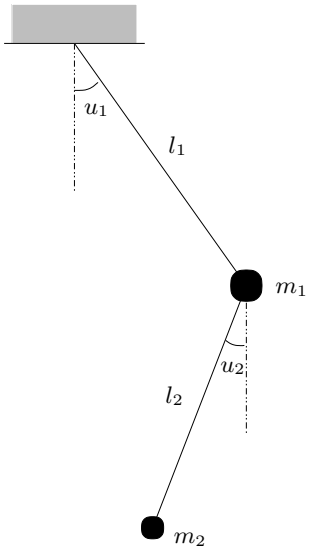


FIGURE 213.3. Double pendulum

Using the identities $\sin^2 u + \cos^2 u = 1$ and $\cos(u_1 - u_2) = \cos u_1 \cos u_2 + \sin u_1 \sin u_2$, we can rewrite this expression as

$$T = \frac{m_1}{2} l_1^2 \dot{u}_1^2 + \frac{m_2}{2} \left[l_1^2 \dot{u}_1^2 + l_2^2 \dot{u}_2^2 + 2 l_1 l_2 \dot{u}_1 \dot{u}_2 \cos(u_1 - u_2) \right].$$

Applying the Principle of Least Action, we obtain the following system of equations for a double pendulum:

$$\begin{aligned} \ddot{u}_1 + \frac{m_2}{m_1 + m_2} \frac{l_2}{l_1} [\ddot{u}_2 \cos(u_2 - u_1) - \dot{u}_2^2 \sin(u_2 - u_1)] + \frac{g}{l_1} \sin u_1 = 0, \\ \ddot{u}_2 + \frac{l_1}{l_2} [\ddot{u}_1 \cos(u_2 - u_1) + \dot{u}_1^2 \sin(u_2 - u_1)] + \frac{g}{l_2} \sin u_2 = 0. \end{aligned} \tag{213.7}$$

We note that if $m_2 = 0$, then the first equation is just the equation for a simple pendulum, and that if $\ddot{u}_1 = \dot{u}_1 = 0$, then the second equation is again the equation for a simple pendulum.

213.8 The Two-Body Problem

We consider the *two-body* problem for a small mass orbiting around a heavy mass, such as the Earth moving around the Sun neglecting the influence of the other planets. We assume that the motion takes place in a plane and use polar coordinates (r, θ) with the origin at the center of the heavy mass

to describe the position of the light body. Assuming that the heavy body is fixed, the action integral representing the difference between the kinetic and potential energy of the small mass is given by

$$\int_{t_1}^{t_2} \left(\frac{1}{2}\dot{r}^2 + \frac{1}{2}(\dot{\theta}r)^2 + \frac{1}{r}\right) dt \tag{213.8}$$

because the velocity is $(\dot{r}, r\dot{\theta})$ in the radial and angular directions respectively, and the gravity potential is $-r^{-1} = -\int_r^\infty s^{-2}\, ds$ corresponding to the work needed to move a particle of unit mass a distance r from the orbit center to infinity. The corresponding Euler-Lagrange equations are

$$\begin{cases} \ddot{r} - r\dot{\theta}^2 = -\frac{1}{r^2}, & t > 0, \\ \frac{d}{dt}(r^2\dot{\theta}) = 0, & t > 0, \end{cases} \tag{213.9}$$

which is a second order system to be complemented with initial values for position and velocity.

We construct the analytical solution of this system in a set of problems below, which may be viewed as a short course on Newton's *Principia Matematica*. We invite the reader to take this opportunity of getting on speaking terms with Newton himself.

213.9 Stability of the Motion of a Pendulum

The linearization of the equation for a pendulum at $\bar{u} \in \mathbb{R}$, $\ddot{u} + \sin(u) = 0$, is obtained by setting $u = \bar{u} + \varphi$ and noting that $\sin(u) \approx \sin(\bar{u}) + \cos(\bar{u})\varphi$. This leads to

$$0 = \ddot{u} + \sin(u) \approx \ddot{\varphi} + \sin(\bar{u}) + \cos(\bar{u})\varphi.$$

Assuming first that $\bar{u} = 0$, we obtain the following linearized equation for the perturbation φ,

$$\ddot{\varphi} + \varphi = 0, \tag{213.10}$$

with solution being a linear combination of $\sin(t)$ and $\cos(t)$. For example, if $\varphi(0) = \delta$ and $\dot{\varphi}(0) = 0$, then $\varphi(t) = \delta\cos(t)$, and we see that an initially small perturbation is kept small for all time: the pendulum stays close to the bottom position under small perturbations.

Setting next $\bar{u} = \pi$, we obtain

$$\ddot{\varphi} - \varphi = 0 \tag{213.11}$$

with the solution being a linear combination of $\exp(\pm t)$. Since $\exp(t)$ grows very quickly, the state $\bar{u} = \pi$ corresponding to the pendulum in the top position is *unstable*. A small perturbation will quickly develop into a large perturbation and the pendulum will move way from the top position.

We will return to the topic of this section in Chapter *Linearization and stability of initial value problems*

Chapter 213 Problems

213.1. Supply the missing details in the derivation of the equation for the pendulum. If the angle u stays small during the motion, then the simpler *linearized* model $\ddot{u} + u = 0$ may be used. Solve this equation analytically and compare with numerical results for the nonlinear pendulum equation to determine limits of validity of the linear model.

213.2. Carry out the details in the derivation of the equations for the pendulum with moving support and the double pendulum.

213.3. Study what happens for the double pendulum in the extreme cases, i.e. at zero and infinity, for the parameters m_1, m_2, l_1 and l_2.

213.4. Derive the second equation of motion for the double pendulum from the result for the pendulum with moving support by setting $(r_1(t), r_2(t)) = (l_1 \sin u_1, -l_1 \cos u_1)$.

213.5. Derive the equation of motion for a bead sliding on a frictionless plane vertical curve under the action of gravity.

213.6. In the foot-steps of Newton give an analysis and analytical solution of the two-body problem modeled by (213.9) through the following sequence of problems: (i) Prove that a stationary point of the action integral (213.8) satisfies (213.9). (ii) Prove that the total energy is constant in time. (iii) Introducing the change of variables $u = r^{-1}$, show that $\dot{\theta} = cu^2$ for c constant. Use this relation together with the fact that the chain rule implies that

$$\frac{dr}{dt} = \frac{dr}{du}\frac{du}{d\theta}\frac{d\theta}{dt} = -c\frac{du}{d\theta} \quad \text{and} \quad \ddot{r} = -c^2u^2\frac{d^2u}{d\theta^2}$$

to rewrite the system (213.9) as

$$\frac{d^2u}{d\theta^2} + u = c^{-2}. \tag{213.12}$$

Show that the general solution of (213.12) is

$$u = \frac{1}{r} = \gamma\cos(\theta - \alpha) + c^{-2},$$

where γ and α are constants. (iii) Prove that the solution is either an ellipse, parabola, or hyperbola. Hint: Use the fact that these curves can be described as the loci of points for which the ratio of the distance to a fixed point and to a fixed straight line, is constant. Polar coordinates are suitable for expressing this relation. (iv) Prove Kepler's three laws for planetary motion using the experience from the previous problem.

213.7. Study the linearizations of the double pendulum at $(u_1, u_2) = (0, 0)$ and $(u_1, u_2) = (\pi, \pi)$ and draw conclusions about stablity.

213.8. Attach an elastic string to a simple pendulum in some way and model the resulting system.

213.9. Compute solutions of the presented models numerically.

214
N-Body Systems*

> The reader will find no figures in this work. The methods which I set forth do not require either geometrical or mechanical reasonings, but only algebraic operations, subject to a regular and uniform rule of procedure. (Lagrange in *Méchanique Analytique*)

214.1 Introduction

We shall now model systems of N bodies interacting through mechanical forces that result from springs and dashpots, see Fig. 214.2, or from gravitational or electrostatic forces. We shall use two different modes of description. In the first formulation, we describe the system through the coordinates of (the centers of gravity of) the bodies. In the second, we use the *displacements* of the bodies measured from an initial reference configuration. In the latter case, we also *linearize* under an assumption of small displacements to obtain a linear system of equations. In the first formulation, the initial configuration is only used to initialize the system and is "forgotten" at a later time in the sense that the description of the system only contains the present position of the masses. In the second formulation, the reference configuration is retrievable through the evolution since the unknown is the displacement from the reference position. The different formulations have different advantages and ranges of application.

214.2 Masses and Springs

We consider the motion in $\mathbb{R}^3$ of a system of N bodies connected by a set of Hookean springs. For $i = 1, ..., N$, let the position at time t of body i be given by the vector function $u_i(t) = (u_{i1}(t), u_{i2}(t), u_{i3}(t))$, with $u_{ik}(t)$ denoting the x_k coordinate, $k = 1, 2, 3$, and suppose the mass of body i is m_i. Let body i be connected to body j with a Hookean spring of spring constant $k_{ij} \geq 0$ for $i, j = 1, ...N$. Some of the k_{ij} may be zero, which effectively means that there is no spring connection between body i and body j. In particular $k_{ii} = 0$. We assume to start with that the reference length of the spring corresponding to zero spring tension is equal to zero. This means that the spring forces are always attractive.

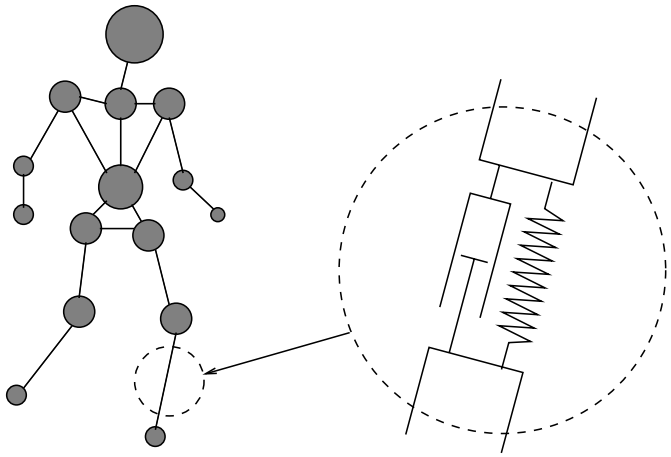


FIGURE 214.1. A typical system of masses, springs and dashpots in motion

We now derive the equations of motion for the mass-spring system using the Principle of Least Action. We assume first that the gravitational force is set to zero. The potential energy of the configuration $u(t)$ is given by

$$\begin{aligned} V(u(t)) &= \sum_{i,j=1}^{N} \frac{1}{2} k_{ij} |u_i - u_j|^2 \\ &= \sum_{i,j=1}^{N} \frac{1}{2} k_{ij} \big((u_{i1} - u_{j1})^2 + (u_{i2} - u_{j2})^2 + (u_{i3} - u_{j3})^2 \big), \end{aligned} \tag{214.1}$$

with the time dependence of the coordinates u_{ik} suppressed for readability. This is because the length of the spring connecting the body i and body j is equal to $|u_i - u_j|$, and the work to stretch the spring from zero length to length l is equal to $\frac{1}{2} k_{ij} l^2$.

The action integral is

$$I(u) = \int_{t_1}^{t_2} \Big(\sum_{i=1}^{N} \frac{1}{2} m_i \big(\dot{u}_{i1}^2 + \dot{u}_{i2}^2 + \dot{u}_{i3}^2\big) - V(u(t))\Big)\, dt,$$

and using the Principle of Least Action and the fact that

$$\frac{\partial V(u)}{\partial u_{ik}} = \sum_{j=1}^{N} k_{ij}(u_{ik} - u_{jk}),$$

we obtain the following system of equations of motion:

$$m_i \ddot{u}_{ik} = -\sum_{j=1}^{N} k_{ij}(u_{ik} - u_{jk}), \quad k = 1, 2, 3,\ i = 1, ..., N, \tag{214.2}$$

or in vector form

$$m_i \ddot{u}_i = -\sum_{j=1}^{N} k_{ij}(u_i - u_j),\ i = 1, ..., N, \tag{214.3}$$

together with initial conditions for $u_i(0)$ and $\dot{u}_i(0)$. We can view these equations as expressing Newton's Law

$$m_i \ddot{u}_i = F_i^s, \tag{214.4}$$

with the total spring force $F_i^s = (F_{i1}^s, F_{i2}^s, F_{i3}^s)$ acting on body i being equal to

$$F_i^s = -\sum_{j=1}^{N} k_{ij}(u_i - u_j). \tag{214.5}$$

Inclusion of gravity forces in the direction of the negative x_3 axis, adds a component $-m_i g$ to F_{i3}^s, where g is the gravitational constant.

The system (214.3) is linear in the unknowns $u_{ij}(t)$. If we assume that the reference length with zero spring force of the spring connecting body i and j is equal to $l_{ij} > 0$, then the potential changes to

$$V(u(t)) = \sum_{i,j=1}^{N} \frac{1}{2} k_{ij}(|u_i - u_j| - l_{ij})^2, \tag{214.6}$$

and the resulting equations of motion are no longer linear. Below, we shall consider a linearized form assuming $|u_i - u_j| - l_{ij}$ is small compared to l_{ij}.

214.3 The N-Body Problem

By tradition, a "N-body" problem refers to a system of N bodies in motion in $\mathbb{R}^3$ under the influence of mutual gravitational forces. An example is given by our solar system with 9 planets orbiting around the Sun, where we typically disregard moons, asteroids, and comets.

Let the position at time t of (the center of gravity of) body i be given by the vector function $u_i(t) = (u_{i1}(t), u_{i2}(t), u_{i3}(t))$, with $u_{ik}(t)$ denoting the x_k coordinate in $\mathbb{R}^3$, $k = 1, 2, 3$, and suppose the mass of body i is m_i. Newton's inverse square law of gravitation states that the gravitational force from the body j on the body i is given by

$$-\frac{\gamma m_i m_j}{|u_i(t) - u_j(t)|^2}\frac{u_i(t) - u_j(t)}{|u_i(t) - u_j(t)|} = -\gamma m_i m_j \frac{u_i(t) - u_j(t)}{|u_i(t) - u_j(t)|^3},$$

where γ is a gravitational constant. We thus obtain the following system of equations modeling the N-body problem:

$$m_i \ddot{u}_i = -\gamma m_i m_j \sum_{j \neq i} \frac{u_i - u_j}{|u_i(t) - u_j(t)|^3}, \tag{214.7}$$

together with initial conditions for $u_i(0)$ and $\dot{u}_i(0)$.

Alternatively, we may derive these equations using the Principle of Least Action using the gravity potential

$$V(u) = -\sum_{i,j=1,\, i\neq j}^{N} \frac{\gamma m_i m_j}{|u_i - u_j|},$$

and the fact that

$$\frac{\partial V}{\partial u_{ik}} = \sum_{j \neq i} \frac{\gamma m_i m_j}{|u_i - u_j|^3}(u_{ik} - u_{jk}). \tag{214.8}$$

The expression for the gravity potential is obtained by noticing that the work to bring body i from a distance r of body j to infinity is equal to

$$\int_r^\infty \frac{\gamma m_i m_j}{s^2}\, ds = \gamma m_i m_j [-\frac{1}{s}]_{s=r}^{s=\infty} = \frac{\gamma m_i m_j}{r}.$$

Notice the minus sign of the potential, arising from the fact that the body i loses potential energy as it approaches body j.

Analytical solutions are available only in the case of the 2-body problem. The numerical solution of for example the 10-body problem of our solar system is very computationally demanding in the case of long time simulation. As a result, the long time stability properties of our Solar system are unknown. For example, it does not seem to be known if eventually

the Earth will change orbit with Mercury, Pluto will spin away to another galaxy, or some other dramatic event will take place.

The general result of existence guarantees a solution, but the presence of the stability factor $\exp(tL_f)$ brings the accuracy in long-time simulation seriously in doubt.

214.4 Masses, Springs and Dashpots: Small Displacements

We now give a different description of the mass-spring system above. Let the initial position of body i, which is now chosen as reference position, be $a_i = (a_{i1}, a_{i2}, a_{i3})$, and let the actual position at time $t > 0$ be given by $a_i + u_i(t)$ where now $u_i(t) = (u_{i1}(t), u_{i2}(t), u_{i3}(t))$ is the *displacement* of body i from its reference position a_i.

The potential energy of the configuration $u(t)$ is given by

$$\begin{aligned} V(u(t)) &= \sum_{i,j=1}^{N} \frac{1}{2} k_{ij} \big(|a_i + u_i - (a_j + u_j)| - |a_i - a_j| \big)^2 \\ &= \frac{1}{2} k_{ij} \big(|a_i - a_j + (u_i - u_j)| - |a_i - a_j| \big)^2, \end{aligned}$$

assuming zero spring forces if the springs have the reference lengths $a_i - a_j$.

We now specialize to small displacements, assuming that $|u_i - u_j|$ is small compared to $|a_i - a_j|$. We then use that if $|b|$ is small compared to $|a|$, where $a, b \in \mathbb{R}^3$, then

$$\begin{aligned} |a+b| - |a| &= \frac{(|a+b| - |a|)(|a+b| + |a|)}{|a+b| + |a|} \\ &= \frac{|a+b|^2 - |a|^2}{|a+b| + |a|} = \frac{(a+b)\cdot(a+b) - a \cdot a}{|a+b| + |a|} \approx \frac{a \cdot b}{|a|}. \end{aligned}$$

Thus, if $|u_i - u_j|$ is small compared to $|a_i - a_j|$, then

$$|a_i - a_j + (u_i - u_j)| - |a_i - a_j| \approx \frac{(a_i - a_j)\cdot(u_i - u_j)}{|a_i - a_j|},$$

and we obtain the following approximation of the potential energy

$$\hat{V}(u(t) = \sum_{i,j=1}^{N} \frac{1}{2} k_{ij} \frac{\big((a_i - a_j)\cdot(u_i - u_j)\big)^2}{|a_i - a_j|^2}.$$

Using the Principle of Least Action we thus obtain the following linearized system of equations

$$m_i \ddot{u}_{ik} = -\sum_{j=1}^{N} \frac{k_{ij}(a_i - a_j)\cdot(u_i - u_j)}{|a_i - a_j|^2}(a_{ik} - a_{jk}), \quad k = 1, 2, 3,\ i = 1, ..., N,$$

or in vector form

$$m_i\ddot{u}_i = -\sum_{j=1}^{N} \frac{k_{ij}(a_i - a_j)\cdot(u_i - u_j)}{|a_i - a_j|^2}(a_i - a_j), \quad i = 1, ..., N. \tag{214.9}$$

together with initial conditions for $u_i(0)$ and $\dot{u}_i(0)$. We can view these equations as expressing Newton's Law

$$m_i\ddot{u}_i = F_i^s, \quad i + 1, ..., N, \tag{214.10}$$

with the spring force F_i^s acting on body i given by

$$F_i^s = -\sum_{j=1}^{N} b_{ij}e_{ij},$$

where

$$e_{ij} = \frac{a_i - a_j}{|a_i - a_j|}$$

is the normalized vector connecting a_j and a_i, and

$$b_{ij} = k_{ij}e_{ij} \cdot (u_i - u_j). \tag{214.11}$$

214.5 Adding Dashpots

A *dashpot* is a kind of shock absorber which may be thought of as consisting of a piston that moves inside a cylinder filled with oil or some other viscous fluid, see Fig. 214.2. As the piston moves, the flow of the fluid past

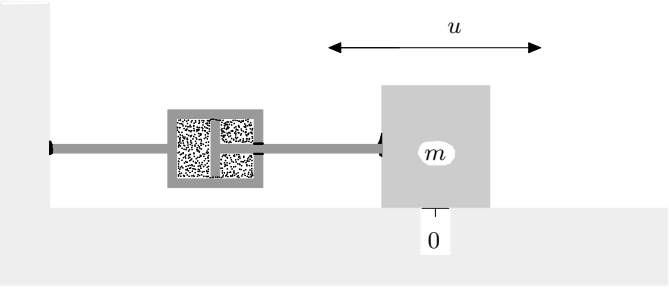


FIGURE 214.2. Cross section of a dashpot connected to a mass.

the piston creates a force opposite to the motion, which we assume is proportional to the velocity with the constant of proportionality representing the coefficient of *viscosity* of the dashpot.

We now expand the above mass-spring model to include springs and dashpots coupled in parallel. For each pair of nodes i and j, we let k_{ij} and μ_{ij} be the coefficients of a spring and a dashpot coupled in parallel, with $k_{ij} = 0$ and $\mu_{ij} = 0$ if the spring or dashpot is absent, and in particular $k_{ii} = \mu_{ii} = 0$. The dashpot force F_i^d acting on body i will then be given by

$$F_i^d = -\sum_{j=1}^{N} d_{ij} e_{ij},$$

where

$$d_{ij} = \mu_{ij} e_{ij} \cdot (\dot{u}_i - \dot{u}_j). \tag{214.12}$$

To get this result, we use the fact that

$$e_{ij} \cdot (\dot{u}_i - \dot{u}_j) e_{ij}$$

is the projection of $\dot{u}_i - \dot{u}_j$ onto e_{ij}. We thus assume that the dashpot reacts with a force that is proportional to the projection of $\dot{u}_i - \dot{u}_j$ onto the direction $a_i - a_j$.

This leads to the linearized mass-spring-dashpot model:

$$m_i \ddot{u}_i = F_i^s + F_i^d, \quad i = 1, ..., N, \tag{214.13}$$

together with initial conditions for $u_i(0)$ and $\dot{u}_i(0)$. We can write these equations as a system in the form

$$M\ddot{u} + D\dot{u} + Ku = 0, \tag{214.14}$$

with constant coefficient matrices matrices M, D and K, where u is a $3N$-vector listing all the components u_{ik}. The matrix M is diagonal with the masses m_i as entries, and D and K are symmetric positive semi-definite (see the problem section).

A system with dashpots is not conservative, since the dashpots consume energy, and therefor cannot be modeled using the Principle of Least Action.

The linear system (214.14) models a wide range of phenomena and can be solved numerically with appropriate solvers. We return to this issue below. We now consider the simplest example of one mass connected to the origin with a spring and a dashpot in parallel.

214.6 A Cow Falling Down Stairs

In Fig. **??** and Fig. **??** we show the result of computational simulation of a cow falling down a staircase. The computational model consists of a skeleton in the form of a mass-spring-dashpot-system together with a surface model built upon the skeleton. The skeleton deforms under the action of gravity forces and contact forces from the staircase and the surface model conforms to the deformation.

214.7 The Linear Oscillator

We now consider the simplest example consisting of one body of mass 1 connected to one end of a Hookean spring connected to the origin with the motion taking place along the x_1-axis. Assuming the spring has zero length at zero tension, the system is described by

$$\begin{cases} \ddot{u} + ku = 0 & \text{for } t > 0, \\ u(0) = u_0,\ \dot{u}(0) = \dot{u}_0. \end{cases} \tag{214.15}$$

with $u(t)$ denoting the x_1 coordinated of the body at time t, and u_0 and $\dot{u}_0$ given initial conditions. The solution is given by

$$u(t) = a\cos(\sqrt{k}t) + b\sin(\sqrt{k}t) = \alpha\cos(\sqrt{k}(t-\beta)), \tag{214.16}$$

where the constants a and b, or α and β, are determined by the initial conditions. We conclude that the motion of the mass is periodic with *frequency* $\sqrt{k}$ and *phase shift* β and *amplitude* α, depending on the initial data. This model is referred to as the *linear oscillator*. The solution is periodic with period $\frac{2\pi}{\sqrt{k}}$, and the *time scale* is similar.

214.8 The Damped Linear Oscillator

Adding a dashpot in parallel with the spring in the model above gives the model of a damped linear oscillator

$$\begin{cases} \ddot{u} + \mu\dot{u} + ku = 0, & \text{for } t > 0, \\ u(0) = u_0, \quad \dot{u}(0) = \dot{u}_0. \end{cases} \tag{214.17}$$

In the case $k = 0$, we obtain the model

$$\begin{cases} \ddot{u} + \mu\dot{u} = 0 & \text{for } t > 0, \\ u(0) = u_0, \quad \dot{u}(0) = \dot{u}_0, \end{cases} \tag{214.18}$$

with the solution

$$u(t) = -\frac{\dot{u}_0}{\mu}\exp(-\mu t) + u_0 + \frac{\dot{u}_0}{\mu}.$$

We see that the mass approaches the fixed position $u = u_0 + \frac{\dot{u}_0}{\mu}$ determined by the initial data as t increases to infinity. The time scale is of size $\frac{1}{\mu}$.

The characteristic polynomial equation for the full model $\ddot{u}+\mu\dot{u}+ku = 0$, is

$$r^2 + \mu r + kr = 0.$$

Completing the square we can write the characteristic equation in the form

$$(r+\frac{\mu}{2})^2=\frac{\mu^2}{4}-k=\frac{1}{4}(\mu^2-4k). \tag{214.19}$$

If $\mu^2-4k>0$, then there are two real roots $-\frac{1}{2}(\mu\pm\sqrt{\mu^2-4k})$, and the solution $u(t)$ has the form (see the Chapter The exponential function),

$$u(t)=ae^{-\frac{1}{2}(\mu+\sqrt{\mu^2-4k})t}+be^{-\frac{1}{2}(\mu-\sqrt{\mu^2-4k})t},$$

with the constants a and b determined by the initial conditions. In this case, the viscous damping of the dashpot dominates over the spring force, and the solution converges exponentially to a rest position, which is equal to $u=0$ if $k>0$. The fastest time scale is again of size $\frac{1}{\mu}$.

If $\mu^2-4k<0$, then we introduce the new variable $v(t)=e^{\frac{\mu t}{2}}u(t)$, with the objective of transforming the characteristic equation (214.19) into an equation of the form $s^2+(k-\frac{\mu^2}{4})=0$. Since $u(t)=e^{-\frac{\mu t}{2}}v(t)$, we have

$$\dot{u}(t)=\frac{d}{dt}\big(e^{-\frac{\mu t}{2}}v(t)\big)=(\dot{v}-\frac{\mu}{2}v)e^{-\frac{\mu t}{2}},$$
$$\ddot{u}(t)=(\ddot{v}-\mu\dot{v}+\frac{\mu^2}{4})e^{-\frac{\mu t}{2}},$$

and thus the differential equation $\ddot{u}+\mu\dot{u}+ku=0$ is transformed into

$$\ddot{v}+(k-\frac{\mu^2}{4})v=0,$$

with the solution $v(t)$ being a linear combination of $\cos(\frac{t}{2}\sqrt{4k-\mu^2})$ and $\sin(\frac{t}{2}\sqrt{4k-\mu^2})$. Transforming back to the variable $u(t)$ we get the solution formula

$$u(t)=ae^{-\frac{1}{2}\mu t}\cos(\frac{t}{2}\sqrt{4k-\mu^2})+be^{-\frac{1}{2}\mu t}\sin(\frac{t}{2}\sqrt{4k-\mu^2}).$$

The solution again converges to the zero rest position as time passes if $\mu>0$, but now it does so in an oscillatory fashion. Now two time scales appear: a time scale of size $\frac{1}{\mu}$ for the exponential decay and a time scale $1/\sqrt{k-\mu^2/4}$ of the oscillations.

Finally, in the limit case $\mu^2-4k=0$ the solution $v(t)$ of the corresponding equation $\ddot{v}=0$ is given by $v(t)=a+bt$, and thus

$$u(t)=(a+bt)e^{-\frac{1}{2}\mu t}.$$

This solution exhibits initial linear growth and eventually converges to a zero rest position as time tends to infinity. We illustrate the three possible behaviors in Fig. 214.3.

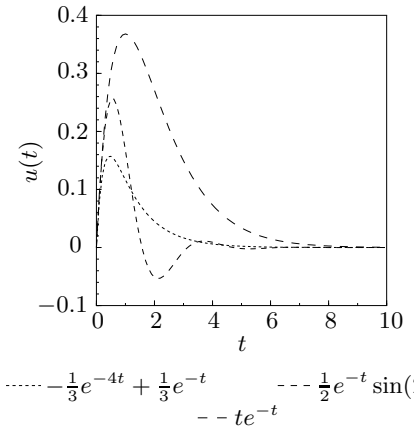


FIGURE 214.3. Three solutions of the mass-spring-dashpot model (214.17) satisfying the initial conditions $u(0) = 0$ and $\dot{u}(0) = 1$. The first solution corresponds to $\mu = 5$ and $k = 4$, the second to $\mu = 2$ and $k = 5$, and the third to $\mu = 2$ and $k = 1$.

214.9 Extensions

We have above studied systems of bodies interacting through Hookean springs, linear dashpots and gravitational forces. We can generalize to systems of non-linear springs, dashpots, and other mechanical devices like springs reacting to changes of angles between the bodies, or other forces like electrostatic forces. In this way, we can model very complex systems from macroscopic scales of galaxies to microscopic molecular scales. For example, electrostatic forces are related to potentials of the form

$$V^e(u) = \pm c \sum_{i,j=1}^{N} \frac{q_i q_j}{|u_i - u_j|}$$

where q_i is the charge of body i and c is a constant, and thus have a form similar to that of gravitational forces.

In particular, models for molecular dynamics take the form of N-body systems interacting through electrostatic forces and forces modeled by various springs reacting to bond lengths and bond angles between the atoms. In these applications, N may be of the order 10^4 and the smallest time scale of the dynamics may be of size 10^{-14} related to very stiff bond length springs. Needless to say, simulations with such models may be very computationally demanding and is often out of reach with present day computers. For more precise information, we refer to the survey article *Molecular modeling of proteins and mathematical prediction of protein structure*, SIAM REV. (39), No 3, 407-460, 1997, by A. Neumair.

Chapter 214 Problems

214.1. Verify the solution formulas for the three solutions shown in Fig. 214.3.

214.2. Write down the model (214.2) in a simple case of a system with a few bodies.

214.3. Derive the equations of motion with the potential (214.6).

214.4. Generalize the mass-spring-dashpot model to arbitrary displacements.

214.5. Generalize the mass-spring model to different non-linear springs.

214.6. Model the vertical motion of a floating buoy. Hint: use that by Archimedes' Principle, the upward force on a cylindrical vertical buoy from the water is proportional to the immersed depth of the buoy.

214.7. Prove that the matrices D and K in (214.14) are symmetric positive semi-definite.

215
Piecewise Linear Approximation

The beginners mind is empty, free of the habits of the expert, ready to accept, or doubt, and open to all the possibilities. It is a kind of mind which can see things as they are. (Shunryu Suzuki)

215.1 Introduction

Approximating a complicated function to arbitrary accuracy by "simpler" functions is a basic tool of applied mathematics. We have seen that piecewise polynomials are very useful for this purpose, and that is why approximation by piecewise polynomials plays a very important role in several areas of applied mathematics. For example, the *Finite Element Method* FEM is an extensively used tool for solving differential equations that is based on piecewise polynomial approximation, see the Chapters FEM for two-point boundary value problems and FEM for Poisson's equation.

In this chapter, we consider the problem of approximating a given real-valued function $f(x)$ on an interval $[a, b]$ by piecewise linear polynomials on a subdivision of $[a, b]$. We derive basic error estimates for interpolation with piecewise linear polynomials and we consider an application to least squares approximation.

215.2 Linear Interpolation on $[0, 1]$

Let $f : [0, 1] \to \mathbb{R}$ be a given Lipschitz continuous function. Consider the function $\pi f : [0, 1] \to \mathbb{R}$ defined by

$$\pi f(x) = f(0)(1 - x) + f(1)x = f(0) + (f(1) - f(0))x.$$

Clearly, $\pi f(x)$ is a *linear* function in x,

$$\pi f(x) = c_0 + c_1 x,$$

where $c_0 = f(0)$, $c_1 = f(1) - f(0)$, and $\pi f(x)$ *interpolates* $f(x)$ at the end-points 0 and 1 of the interval $[0, 1]$, by which we mean that πf takes the same values as f at the end-points, i.e.

$$\pi f(0) = f(0), \quad \pi f(1) = f(1).$$

We refer to $\pi f(x)$ as a linear *interpolant* of $f(x)$ that interpolates $f(x)$ at the end-points of the interval $[0, 1]$.

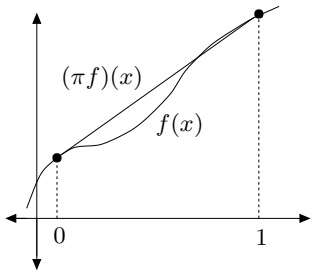


FIGURE 215.1. The linear interpolant πf of a function f.

We now study the *interpolation error* $f(x) - \pi f(x)$ for $x \in [0, 1]$. Before doing so we get some perspective on the space of linear functions on $[0, 1]$ to which the interpolant πf belongs.

The Space of Linear Functions

We let $\mathcal{P} = \mathcal{P}(0, 1)$ denote the set of first order (linear) polynomials

$$p(x) = c_0 + c_1 x,$$

defined for $x \in [0, 1]$, where the real numbers c_0 and c_1 are the coefficients of p. We recall that two polynomials $p(x)$ and $q(x)$ in $\mathcal{P}$ may be added to give a new polynomial $p + q$ in $\mathcal{P}$ defined by $(p + q)(x) = p(x) + q(x)$, and that a polynomial $p(x)$ in $\mathcal{P}$ may be multiplied by a scalar α to give a polynomial αp in $\mathcal{P}$ defined by $(\alpha p)(x) = \alpha p(x)$. Adding two polynomials

is carried out by adding their coefficients, and multiplying a polynomial by a real number is carried out by multiplying the coefficients by the real number.

We conclude that $\mathcal{P}$ is a vector space where each vector is a particular first order polynomial $p(x) = c_0 + c_1 x$ determined by the two real numbers c_0 and c_1. As a basis for $\mathcal{P}$ we may choose $\{1, x\}$. To see this, we note that each $p \in \mathcal{P}$ can be uniquely expressed as a linear combination of 1 and x: $p(x) = c_0 + c_1 x$, and we may refer to the pair (c_0, c_1) as the coordinates of the polynomial $p(x) = c_0 + c_1 x$ with respect to the basis $\{1, x\}$. For example, the coordinates of the polynomial $p(x) = x$ with respect to the basis $\{1, x\}$, are $(0, 1)$, right? Since there are two basis functions, we say that the dimension of the vector space $\mathcal{P}$ is equal to two.

We now consider an alternative basis $\{\lambda_0, \lambda_1\}$ for $\mathcal{P}$ consisting of the two functions λ_0 and λ_1 defined

$$\lambda_0(x) = 1 - x, \quad \lambda_1(x) = x.$$

Each of these functions takes the value 0 at one end-point and the value 1 at the other end-point, namely

$$\lambda_0(0) = 1,\ \lambda_0(1) = 0, \qquad \text{and } \lambda_1(0) = 0, \lambda_1(1) = 1.$$

See Fig. 215.2.

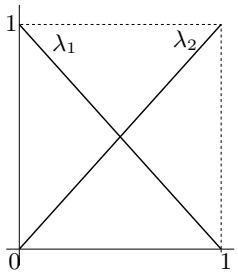


FIGURE 215.2. The basis functions λ_0 and λ_1.

Any polynomial $p(x) = c_0 + c_1 x$ in $\mathcal{P}$ can be expressed as a linear combination of the functions $\lambda_0(x)$ and $\lambda_1(x)$, i.e.

$$\begin{aligned} p(x) &= c_0 + c_1 x = c_0(1 - x) + (c_1 + c_0)x = c_0\lambda_0(x) + (c_1 + c_0)\lambda_1(x) \\ &= p(0)\lambda_0(x) + p(1)\lambda_1(x), \end{aligned}$$

A very nice feature of these functions is that the coefficients $p(0)$ and $p(1)$ are the values of $p(x)$ at $x = 0$ and $x = 1$. Moreover, λ_0 and λ_1 are linearly independent, since if

$$a_0\lambda_0(x) + a_1\lambda_1(x) = 0 \quad \text{for } x \in [0, 1],$$

then setting $x = 0$ and $x = 1$ shows that $a_1 = a_0 = 0$. We conclude that $\{\lambda_0, \lambda_1\}$ is a basis for $\mathcal{P}$.

In particular, we can express the interpolant $\pi f \in \mathcal{P}$ in the basis $\{\lambda_0, \lambda_1\}$ as follows:

$$\pi f(x) = f(0)\lambda_0(x) + f(1)\lambda_1(x), \tag{215.1}$$

where the end-point values $f(0)$ and $f(1)$ appear as coefficients.

The Interpolation Error

We want to estimate the interpolation error $f(x) - \pi f(x)$ for $x \in [0, 1]$. We prove that

$$|f(x) - \pi f(x)| \leq \frac{1}{2}x(1-x) \max_{y\in[0,1]} |f''(y)|, \quad x \in [0, 1]. \tag{215.2}$$

Since (convince yourself!)

$$0 \leq x(1-x) \leq \frac{1}{4} \quad \text{for } x \in [0, 1],$$

we can state the interpolation error estimate in the form

$$\max_{x\in[0,1]} |f(x) - \pi f(x)| \leq \frac{1}{8} \max_{y\in[0,1]} |f''(y)|. \tag{215.3}$$

This estimate states that the maximal value of the interpolation error $|f(x)-\pi f(x)|$ over $[0, 1]$ is bounded by a constant times the maximum value of the second derivative $|f''(y)|$ over $[0, 1]$, i.e. to the degree of concavity or convexity of f, or the amount that f curves away from being linear, see Fig. 215.3.

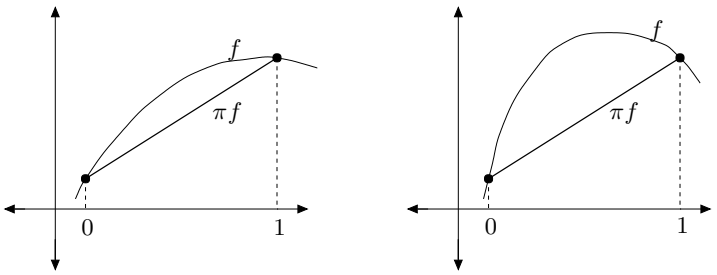


FIGURE 215.3. The error of a linear interpolant depends on the size of $|f''|$, which measures the degree that f curves away from being linear. Notice that the error of the linear interpolant of the function on the right is much larger than of the linear interpolant of the function on the left and the function on the right has a larger second derivative in magnitude.

To prove (215.2), we fix x in $(0,1)$ and use Taylor's theorem to express the values $f(0)$ and $f(1)$ in terms of $f(x)$, $f'(x)$, $f''(y_0)$ and $f''(y_1)$ where $y_0 \in (0,x)$ and $y_1 \in (x,1)$. This gives

$$\begin{aligned} f(0) &= f(x) + f'(x)(-x) + \frac{1}{2}f''(y_0)(-x)^2, \\ f(1) &= f(x) + f'(x)(1-x) + \frac{1}{2}f''(y_1)(1-x)^2. \end{aligned} \tag{215.4}$$

Substituting the Taylor expansions (215.4) into (215.1) and using the identities

$$\begin{aligned} \lambda_0(x) + \lambda_1(x) &= (1-x) + x \equiv 1, \\ (-x)\lambda_0(x) + (1-x)\lambda_1(x) &= (-x)(1-x) + (1-x)x \equiv 0, \end{aligned} \tag{215.5}$$

we obtain the *error representation*

$$f(x) - \pi f(x) = -\frac{1}{2}\big(f''(y_0)(-x)^2(1-x) + f''(y_1)(1-x)^2 x\big).$$

Using the identity $(-x)^2(1-x)+(1-x)^2x = x(1-x)(x+1-x) = x(1-x)$ gives (215.2),

$$|f(x) - \pi f(x)| \le \frac{1}{2}x(1-x)\max_{y\in[0,1]}|f''(y)| \le \frac{1}{8}\max_{y\in[0,1]}|f''(y)|. \tag{215.6}$$

Next, we prove the following estimate for the error in the first derivative,

$$|f'(x) - (\pi f)'(x)| \le \frac{x^2 + (1-x)^2}{2}\max_{y\in[0,1]}|f''(y)|, \quad x \in [0,1]. \tag{215.7}$$

Since $0 \le x^2 + (1-x)^2 \le 1$ for $x \in [0,1]$,

$$\max_{x\in[0,1]}|f'(x) - (\pi f)'(x)| \le \frac{1}{2}\max_{y\in[0,1]}|f''(y)|.$$

We illustrate in Fig. 215.4.

To prove (215.7), we differentiate (215.1) with respect to x (note that the x-dependence is carried by $\lambda_0(x)$ and $\lambda_1(x)$) and use (215.4) together with the obvious identities

$$\begin{aligned} \lambda_0'(x) + \lambda_1'(x) &= -1 + 1 \equiv 0, \\ (-x)\lambda_0'(x) + (1-x)\lambda_1'(x) &= (-x)(-1) + (1-x) \equiv 1. \end{aligned}$$

This gives the error representation:

$$f'(x) - (\pi f)'(x) = -\frac{1}{2}\big(f''(y_0)(-x)^2(-1) + f''(y_1)(1-x)^2\big),$$

where again $y_0 \in (0,x)$ and $y_1 \in (x,1)$. This proves the desired result.

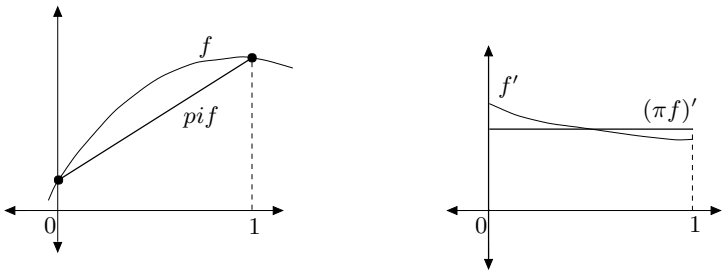


FIGURE 215.4. The derivative of a linear interpolant of f approximates the derivative of f. We show f and the linear interpolant πf on the left and their derivatives on the right.

Finally, we prove an estimate for $|f(x) - \pi f(x)|$ using only the first derivative f'. This is useful when the second derivative f'' does not exist. The Mean Value theorem implies

$$f(0) = f(x) + f'(y_0)(-x), \quad f(1) = f(x) + f'(y_1)(1-x), \tag{215.8}$$

where $y_0 \in [0, x]$ and $y_1 \in [x, 1]$. Substituting into (215.1), we get

$$|f(x) - \pi f(x)| = |f'(y_0)x(1-x) - f'(y_1)(1-x)x| \leq 2x(1-x) \max_{y\in[0,1]} |f'(y)|.$$

Since $2x(1-x) \leq \frac{1}{2}$ for $0 \leq x \leq 1$, we thus find that

$$\max_{x\in[0,1]} |f(x) - \pi f(x)| \leq \frac{1}{2} \max_{y\in[0,1]} |f'(y)|.$$

We summarize in the following theorem.

Theorem 215.1 *The linear polynomial $\pi f \in \mathcal{P}(0,1)$, which interpolates the given function $f(x)$ at $x = 0$ and $x = 1$, satisfies the following error bounds:*

$$\begin{aligned}
\max_{x\in[0,1]} |f(x) - \pi f(x)| &\leq \frac{1}{8} \max_{y\in[0,1]} |f''(y)|, \\
\max_{x\in[0,1]} |f(x) - \pi f(x)| &\leq \frac{1}{2} \max_{y\in[0,1]} |f'(y)|, \\
\max_{x\in[0,1]} |f'(x) - (\pi f)'(x)| &\leq \frac{1}{2} \max_{y\in[0,1]} |f''(y)|.
\end{aligned} \tag{215.9}$$

The corresponding estimates for an arbitrary interval $I = [a, b]$ of length $h = b - a$ takes the following form, where of course $\mathcal{P}(a, b)$ denotes the set of linear functions on $[a, b]$. Observe how the length $h = b-a$ of the interval enters, with the factor h^2 in the estimate for $f(x) - \pi f(x)$ with f'', and h in the estimate for $f'(x) - (\pi f)'(x)$.

Theorem 215.2 *The linear polynomial $\pi f \in \mathcal{P}(a,b)$, which interpolates the given function $f(x)$ at $x = a$ and $x = b$, satisfies the following error bounds:*

$$\begin{aligned}
\max_{x\in[a,b]} |f(x) - \pi f(x)| &\leq \frac{1}{8} \max_{y\in[a,b]} |h^2 f''(y)|,\\
\max_{x\in[a,b]} |f(x) - \pi f(x)| &\leq \frac{1}{2} \max_{y\in[a,b]} |hf'(y)|,\\
\max_{x\in[a,b]} |f'(x) - (\pi f)'(x)| &\leq \frac{1}{2} \max_{y\in[a,b]} |hf''(y)|,
\end{aligned} \tag{215.10}$$

where $h = b - a$.

If we define the *maximum norm* over $I = [a,b]$ by

$$\|v\|_{L_\infty(I)} = \max_{x\in[a,b]} |v(x)|,$$

then we can state (215.9) as follows

$$\begin{aligned}
\|f - \pi f\|_{L_\infty(I)} &\leq \frac{1}{8} \|h^2 f''\|_{L_\infty(I)},\\
\|f - \pi f\|_{L_\infty(I)} &\leq \frac{1}{2} \|hf'\|_{L_\infty(I)},\\
\|f' - (\pi f)'\|_{L_\infty(I)} &\leq \frac{1}{2} \|hf''\|_{L_\infty(I)}.
\end{aligned} \tag{215.11}$$

Below we shall use an analog of this estimate with the $L_\infty(I)$-norm replaced by the $L_2(I)$-norm.

215.3 The Space of Piecewise Linear Continuous Functions

For a given interval $I = [a,b]$, we let $a = x_0 < x_1 < x_2 < \cdots < x_N = b$ be a *partition* of I into N sub-intervals $I_i = (x_{i-1}, x_i)$ of length $h_i = x_i - x_{i-1}$, $i = 1, ..., N$. We denote by $h(x)$ the *mesh function* defined by $h(x) = h_i$ for $x \in I_i$ and we use $\mathcal{T}_h = \{I_i\}_{i=1}^N$ to denote the set of intervals or *mesh* or partition.

We introduce the vector space V_h of continuous piecewise linear functions on the mesh $\mathcal{T}_h$. A function $v \in V_h$ is linear on each subinterval I_i and is continuous on $[a,b]$. Adding two functions in V_h or multiplying a function in V_h by a real number gives a new function in V_h, and thus V_h is indeed a vector space. We show an example of such a function in Fig. 216.2.

We now present a particularly important basis for V_h that consists of the *hat functions* or *nodal basis functions* $\{\varphi_i\}_{i=0}^N$ illustrated in Fig. 215.5.

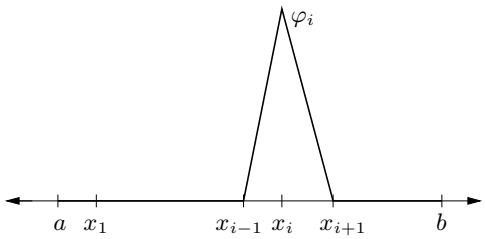


FIGURE 215.5. The hat function φ_i associated to node x_i.

The hat-function $\varphi_i(x)$ is a function in V_h satisfying

$$\varphi_i(x_j) = 1 \quad \text{if } j = i, \quad \varphi_i(x_j) = 0 \quad \text{if } j \neq i.$$

and is given by the formula:

$$\varphi_i(x) = \begin{cases} 0, & x \notin [x_{i-1}, x_{i+1}], \\ \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i], \\ \dfrac{x - x_{i+1}}{x_i - x_{i+1}}, & x \in [x_i, x_{i+1}]. \end{cases}$$

The basis functions φ_0 and φ_N associated to the boundary nodes x_0 and x_N look like "half hats". Observe that each hat function $\varphi_i(x)$ is defined on the whole interval $[a, b]$ and takes the value zero outside the interval $[x_{i-1}, x_{i+1}]$ (or $[a, x_1]$ if $i = 0$ and $[x_{N-1}, b]$ if $i = N$).

The set of hat-functions $\{\varphi_i\}_{i=0}^N$ is a basis for V_h because each $v \in V_h$ has the unique representation

$$v(x) = \sum_{i=0}^{N} v(x_i)\varphi_i(x),$$

where the nodal values $v(x_i)$ appear as coefficients. To see this, it is sufficient to realize that the functions on the left and right hand side are both continuous and piecewise linear and take the same values at the nodes, and thus coincide. Since the number of basis functions φ_i is equal to $N + 1$, the dimension of V_h is equal to $N + 1$.

The continuous piecewise linear interpolant $\pi_h f \in V_h$ of a given Lipschitz continuous function $f(x)$ on $[0, 1]$ is defined by

$$\pi_h f(x_i) = f(x_i) \quad \text{for } i = 0, 1, ..., N,$$

that is, $\pi_h f(x)$ interpolates $f(x)$ at the nodes x_i, see Fig. 215.6. We can express $\pi_h f$ in terms of the basis of hat functions $\{\varphi_i\}_{i=0}^N$ as follows:

$$\pi_h f = \sum_{i=0}^{N} f(x_i)\varphi_i \quad \text{or} \quad \pi_h f(x) = \sum_{i=0}^{N} f(x_i)\varphi_i(x) \quad \text{for } x \in [0, 1], \tag{215.12}$$

with the x-dependence indicated.

Since $\pi_h f(x)$ is linear on each subinterval I_i and interpolates $f(x)$ at the end-points of I_i, we can express $\pi f(x)$ analytically on I_i as follows:

$$\pi_h f(x) = f(x_{i-1})\frac{x - x_i}{x_{i-1} - x_i} + f(x_i)\frac{x - x_{i-1}}{x_i - x_{i-1}} \quad \text{for } x_{i-1} \le x \le x_i,$$

for $i = 1, ..., N$.

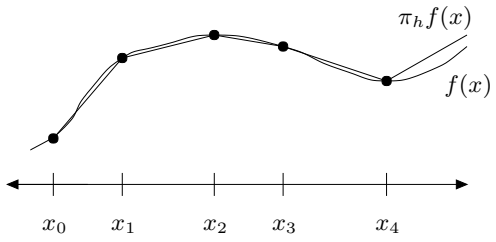


FIGURE 215.6. An example of a continuous piecewise linear interpolant.

Using Theorem 215.2, we obtain the following error estimate for piecewise linear interpolation:

Theorem 215.3 *The piecewise linear interpolant $\pi_h f(x)$ of a twice differentiable function $f(x)$ on a partition of $[a, b]$ with mesh function $h(x)$ satisfies*

$$\begin{aligned} \|f - \pi_h f\|_{L_\infty(a,b)} &\le \frac{1}{8}\|h^2 f''\|_{L_\infty(a,b)}, \\ \|f' - (\pi_h f)'\|_{L_\infty(a,b)} &\le \frac{1}{2}\|h f''\|_{L_\infty(a,b)}. \end{aligned} \tag{215.13}$$

If $f(x)$ is only once differentiable, then

$$\|f - \pi_h f\|_{L_\infty(a,b)} \le \frac{1}{2}\|h f'\|_{L_\infty(a,b)}. \tag{215.14}$$

Note that since the mesh function $h(x)$ may have jumps at the nodes, we interpret $\|h^2 f''\|_{L_\infty(a,b)}$ as

$$\max_{i=1,...,N} \max_{y \in [x_{i-1}, x_i]} |h^2(y) f''(y)|,$$

where $h(y) = x_i - x_{i-1}$ for $y \in [x_{i-1}, x_i]$.

215.4 The L_2 Projection into V_h

Let $f(x)$ be a given function on an interval $I = [a, b]$ and V_h denote the space of continuous piecewise linear functions V_h on a partition $a = x_0 < ... < x_N = b$ of I with mesh function $h(x)$.

The *orthogonal projection* $P_h f$ of the function f into V_h is the function $P_h f \in V_h$ such that

$$\int_I (f - P_h f)v\,dx = 0 \quad \text{for } v \in V_h. \tag{215.15}$$

Recalling the definition of the $L_2(I)$-scalar product

$$(v, w)_{L_2(I)} = \int_I v(x)w(x)\,dx,$$

with the corresponding $L_2(I)$-norm

$$\|v\|_{L_2(I)} = (\int_I v^2(x)\,dx)^{1/2},$$

we can write (215.15) in the form

$$(f - P_h f, v)_{L_2(I)} = 0 \quad \text{for } v \in V_h.$$

This says that $f - P_h f$ is orthogonal to V_h with respect to the $L_2(I)$ scalar product. We also call $P_h f$ the $L_2(I)$-*projection* of f onto V_h.

We first show that $P_h f$ is uniquely defined and then prove that $P_h f$ is the best V_h-approximation of f in the $L_2(I)$-norm.

To prove uniqueness and existence, we express $P_h f$ in the nodal basis $\{\varphi_i\}_{i=0}^N$:

$$P_h f(x) = \sum_{j=0}^N c_j \varphi_j(x),$$

where the $c_j = (P_h f)(x_j)$ are the nodal values of $P_h f$ that have to be determined. We insert this representation into (215.15) and choose $v = \varphi_i$ with $i = 0, ..., N$, to get for $i = 0, ..., N$,

$$\int_I \sum_{j=0}^N c_j \varphi_j(x)\varphi_i(x)\,dx = \sum_{j=0}^N c_j \int_I \varphi_j(x)\varphi_i(x)\,dx$$
$$= \int_I f\varphi_i\,dx \equiv b_i\,, \tag{215.16}$$

where we changed the order of integration and summation. This gives the following system of equations

$$\sum_{j=0}^N m_{ij}c_j = \int_I f\varphi_i\,dx \equiv b_i \quad i = 0, 1, ..., N, \tag{215.17}$$

where

$$m_{ij} = \int_I \varphi_j(x)\varphi_i(x)\,dx, \quad i, j = 0, ..., N.$$

We can write (215.17) in matrix form as

$$Mc = b$$

where $c = (c_0, ..., c_N)$ is a $N+1$-vector of the unknown coefficients c_j, and $b = (b_0, ..., b_N)$ is computable from $f(x)$, and $M = (m_{ij})$ is a $(N+1)\times(N+1)$-matrix that depends on the basis functions φ_i, but not on the function $f(x)$. We refer to the matrix M as the *mass matrix*.

We can now easily prove the uniqueness of $P_h f$. Since the difference $P_h f - \bar{P}_h f$ of two functions $P_h f \in V_h$ and $\bar{P}_h f \in V_h$ satisfying the relation (215.15), also satisfy

$$\int_I (P_h f - \bar{P}_h f) v \, dx = 0 \quad \text{for } v \in V_h,$$

by choosing $v = P_h f - \bar{P}_h f$, we get

$$\int_I (P_h f - \bar{P}_h f)^2 \, dx = 0,$$

and thus $P_h f(x) = \bar{P}_h f(x)$ for $x \in I$. Solutions of the system $Mc = b$ are therefore unique, and since M is a square matrix, existence follows from the Fundamental Theorem of Linear Algebra. We sum up:

Theorem 215.4 *The $L_2(I)$-projection $P_h f$ of a given function f onto the set of piecewise linear functions V_h on I is uniquely defined by (215.15) or the equivalent system of equations $Mc = b$, where $c_j = P_h f(x_j)$ are the nodal values of $P_h f$, M is the mass matrix with coefficients $m_{ij} = (\varphi_j, \varphi_i)_{L_2(I)} = (\varphi_i, \varphi_j)_{L_2(I)}$ and the coefficients of the right hand side b are given by $b_i = (f, \varphi_i)$.*

EXAMPLE 215.1. We compute the mass matrix M in the case of a uniform subdivision with $h(x) = h = (b-a)/N$ for $x \in I$. We get by a direct computation

$$m_{ii} = \int_{x_{i-1}}^{x_{i+1}} \varphi_i^2(x) \, dx = \frac{2h}{3} \quad i = 1, ...N-1, \qquad m_{00} = m_{NN} = \frac{h}{3},$$

$$m_{i,i+1} = \int_{x_{i-1}}^{x_{i+1}} \varphi_i(x)\varphi_{i+1}(x) \, dx = \frac{h}{6} \quad i = 1, ...N-1.$$

The corresponding "lumped" mass matrix $\hat{M} = (\hat{m}_{ij})$, which is a diagonal matrix with the diagonal element in each row being the sum of the elements in the corresponding row of M, takes the form

$$\hat{m}_{ii} = h \quad i = 1, ..., N-1, \quad \hat{m}_{00} = \hat{m}_{NN} = h/2.$$

We see that $\hat{M}$ may be viewed as a h-scaled variant of the identity matrix and M can be viewed as an h-scaled approximation of the identity matrix.

We now prove that the $L_2(I)$-projection $P_h f$ of a function f satisfies

$$\|f - P_h f\|_{L_2(I)} \leq \|f - v\|_{L_2(I)}, \quad \text{for all } v \in V_h. \tag{215.18}$$

This implies that $P_h f$ is the element in V_h with smallest deviation from f in the $L_2(I)$-norm. Applying Cauchy's inequality to (215.15) with $v \in V_h$ gives

$$\begin{aligned}
&\int_I (f - P_h f)^2\,dx \\
&= \int_I (f - P_h f)(f - P_h f)\,dx + \int_I (f - P_h f)(P_h f - v)\,dx \\
&= \int_I (f - P_h f)(f - v)\,dx \leq (\int_I (f - P_h f)^2\,dx)^{1/2} (\int_I (f - v)^2\,dx)^{1/2},
\end{aligned}$$

which proves the desired result. We summarize:

Theorem 215.5 *The $L_2(I)$-projection P_h into V_h defined by (215.15), is the unique element in V_h which minimizes $\|f - v\|_{L_2(I)}$ with v varying over V_h.*

In particular, choosing $v = \pi_h f$ in (215.18), we obtain

$$\|f - P_h f\|_{L_2(I)} \leq \|f - \pi_h f\|_{L_2(I)},$$

where $\pi_h f$ is the nodal interpolant of f introduced above. One can prove the following analog of (215.13)

$$\|f - \pi_h f\|_{L_2(I)} \leq \frac{1}{\pi^2} \|h^2\, f''\|_{L_2(I)},$$

where the interpolation constant happens to be π^{-2}. We thus conclude the following basic result:

Theorem 215.6 *The $L_2(I)$-projection P_h into the space of piecewise linear functions V_h on I with mesh function $h(x)$, satisfies the following error estimate:*

$$\|f - P_h f\|_{L_2(I)} \leq \frac{1}{\pi^2} \|h^2\, f''\|_{L_2(I)}. \tag{215.19}$$

Chapter 215 Problems

215.1. Give a different proof of the first estimate of Theorem Theorem 215.1 by considering for a given $x \in (0, 1)$, the function

$$g(y) = f(y) - \pi f(y) - \gamma(x) y(1 - y), \quad y \in [0, 1],$$

where $\gamma(x)$ is chosen so that $g(x) = 0$. Hint: the function $g(y)$ vanishes at 0, x and 1. Show by repeated use of the Mean Value theorem that g'' vanishes at some point ξ, from which it follows that $\gamma(x) = -f''(\xi)/2$.

215.2. Prove Theorem 215.2 from Theorem 215.1 by using the change of variables $x = a + (b-a)z$ transforming the interval $[0,1]$ onto $[a,b]$, setting $F(z) = f(a+(b-a)z)$ and using that by the Chain Rule, $F' = \frac{dF}{dz} = (b-a)f' = (b-a)\frac{df}{dx}$.

215.3. Develop approximation/interpolation with piecewise constant (discontinuous) functions on a partition of an interval. Consider interpolation at left-hand endpoint, right-hand endpoint, midpoint and mean value for each subinterval. Prove error estimates of the form $\|u - \pi_h u\|_{L_\infty(I)} \leq C\|hu'\|_{L_\infty(I)}$, with $C = 1$ or $C = \frac{1}{2}$.

216

FEM for Two-Point Boundary Value Problems

> The results, however, of the labour and invention of this century are not to be found in a network of railways, in superb bridges, in enormous guns, or in instantaneous communication. We must compare the social state of the inhabitants of the country with what it was. The change is apparent enough. The population is double what it was a century back; the people are better fed and better housed, and comforts and even luxuries that were only within the reach of the wealthy can now be obtained by all classes alike.....But with these advantages there are some drawbacks. These have in many cases assumed national importance, and it has become the province of the engineer to provide a remedy. (Reynolds, 1868)

216.1 Introduction

We begin by deriving a model that is based on a *conservation principle* which states:

> The rate at which a specified quantity changes in a region is equal to the rate that the quantity leaves and enters the region plus the rate at which the quantity is created and destroyed inside the region.

Such a conservation principle holds for a wide variety of quantities, including animals, automobiles, bacteria, chemicals, fluids, heat and energy, etc. So the model we derive has a wide application.

In this chapter, we assume that the quantity to be modeled exists in a very small diameter "tube" with constant cross section and that the quantity varies in the direction along the tube but not at all within a fixed cross section, see Fig. 216.1. We use x to denote the position along the length of the tube and let t denote time. We assume that the quantity in the tube is sufficiently abundant that it makes sense to talk about a *density* $u(x,t)$, measured in amount of the quantity per unit volume, that varies continuously with the position x and time t. This is certainly valid for quantities such as heat and energy, and may be more or less valid for quantities such as bacteria and chemicals provided there is a sufficient number of creatures or molecules respectively.

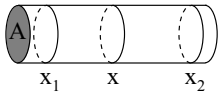


FIGURE 216.1. Variation in a very narrow "tube".

We next express the conservation principle mathematically. We consider a small region of the tube of width dx and cross-sectional area A. The amount of quantity in this region is $u(x,t)A\,dx$. We let $q(x,t)$ denote the *flux* at position x and time t, or the amount of the quantity crossing the section at x at time t measured in amount per unit area per unit time. We choose the orientation so that q is positive when the flow is to the right. The amount of quantity crossing the section at position x at time t is therefore $Aq(x,t)$. Lastly, we let $f(x,t)$ denote the rate that the quantity is created or destroyed within the section at x at time t measured in amount per unit volume per unit time. So, $f(x,t)Adx$ is the amount of the quantity created or destroyed in the small region of width dx per unit time.

The conservation principle for a fixed length of pipe between $x = x_1$ and $x = x_2$ implies that the rate of change of the quantity in this section must equal the rate at which it flows in at $x = x_1$ minus the rate at which it flows out at $x = x_2$ plus the rate at which it is created in $x_1 \le x \le x_2$. In mathematical terms,

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} u(x,t)A\,dx = Aq(x_1,t) - Aq(x_2,t) + \int_{x_1}^{x_2} f(x,t)A\,dx.$$

or

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} u(x,t)\,dx = q(x_1,t) - q(x_2,t) + \int_{x_1}^{x_2} f(x,t)\,dx. \tag{216.1}$$

Equation (216.1) is called the *integral formulation* of the conservation principle.

We can reformulate (216.1) as a partial differential equation provided $u(x,t)$ and $q(x,t)$ are sufficiently smooth. For we can write,

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} u(x,t)\,dx = \int_{x_1}^{x_2} \frac{\partial}{\partial t}u(x,t)\,dx,$$
$$q(x_1,t) - q(x_2,t) = \int_{x_1}^{x_2} \frac{\partial}{\partial x}q(x,t)\,dx,$$

and therefore collecting terms,

$$\int_{x_1}^{x_2} \big(\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}q(x,t) - f(x,t)\big)\,dx = 0.$$

Since x_1 and x_2 are arbitrary, the integrand must be zero at each point, or

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}q(x,t) = f(x,t). \tag{216.2}$$

Equation (216.2) is the *pointwise* or *differential formulation* of the conservation principle.

So far we have one equation for two unknowns. To complete the model, we use a *constitutive relation* that describes the relation between the flux and the density. This relation is specific to the physical properties of the quantity being modeled, yet it is often unclear exactly how to model these properties. A constitutive relation used in practice is often only an approximation to the true unknown relation.

Many quantities have the property that the quantity flows from regions of high concentration to regions of low concentration, and the rate of flow increases as the differences in concentration increases. As a first approximation, we assume a simple linear relation

$$q(x,t) = -a(x,t)\frac{\partial}{\partial x}u(x,t), \tag{216.3}$$

where $a(x,t) > 0$ is the *diffusion coefficient*. In case u represents heat, (216.3) is known as *Newton's Heat Law*. In general, equation (216.3) is known as *Fick's Law*. Note that the choice of sign of a guarantees for example that flow is to the right if $u_x < 0$, i.e. if u decreases across the section at x. Substituting (216.3) into (216.2), we obtain the general time-dependent reaction-diffusion equation,

$$\frac{\partial}{\partial t}u(x,t) - \frac{\partial}{\partial x}\big(a(x,t)\frac{\partial}{\partial x}u(x,t)\big) = f(x,t).$$

To simplify the notation, we use $\dot{u}$ to denote $\partial u/\partial t$ and u' to denote $\partial u/\partial x$. This yields

$$\dot{u}(x,t) - (a(x,t)u'(x,t))' = f(x,t). \tag{216.4}$$

Convection or transport is another important process to take into account in this model.

EXAMPLE 216.1. When modeling populations of animals, diffusion reflects the natural tendency of most creatures to spread out over a region due to randomly occurring interactions between pairs of creatures, while convection models phenomena such as migration.

Convection is modeled by assuming a constitutive relation in which the flux is proportional to the density, i.e.

$$\varphi(x,t) = b(x,t)u(x,t),$$

which results in a convection term in the differential equation of the form $(bu)'$. The convection coefficient $b(x,t)$ determines the rate and direction of transport of the quantity being modeled.

In general, many quantities are modeled by a constitutive relation of the form

$$\varphi(x,t) = -a(x,t)u'(x,t) + b(x,t)u(x,t)$$

which combines diffusion and convection. Arguing as above, we obtain the general reaction-diffusion-convection equation

$$\dot{u}(x,t) - (a(x,t)u'(x,t))' + (b(x,t)u(x,t))' = f(x,t). \tag{216.5}$$

216.2 Initial Boundary-Value Problems

We have to add suitable data to (216.4) or (216.5) in order to specify a unique solution. We model the amount of substance in a fixed length of tube located between $x = 0$ and $x = 1$, as in Fig. 216.1, and specify some information about u called *boundary conditions* at $x = 0$ and $x = 1$. We also need to give some initial data at some initial time, which we take to be $t = 0$. The *evolutionary* or *time-dependent initial two point boundary value problem* reads: find $u(x,t)$ such that

$$\begin{cases} \dot{u} - (au')' + (bu)' = f & \text{in } (0,1)\times(0,T), \\ u(0,t) = u(1,t) = 0 & \text{for } t \in (0,T) \\ u(x,0) = u_0(x) & \text{for } x \in (0,1), \end{cases} \tag{216.6}$$

where a, b, c are given coefficients and f and g are given data. The boundary values $u(0,t) = u(1,t) = 0$ are known as *homogeneous Dirichlet boundary conditions*.

EXAMPLE 216.2. In the case that we use (216.4) to model the heat u in a long thin wire, the coefficient a represents the heat conductivity of the metal in the wire, f is a given heat source, and the homogeneous Dirichlet boundary conditions at the end-points means that the temperature of the wire is held fixed at 0 there. Such conditions are realistic for example if the wire is attached to very large masses at the ends.

Other boundary conditions found in practice include: *nonhomogeneous Dirichlet boundary conditions* $u(0) = u_0$, $u(1) = u_1$ with constants u_0, u_1; one homogeneous Dirichlet $u(0) = 0$ and one *nonhomogeneous Neumann boundary condition* $a(1)u'(1) = g_1$ with constant g_1; and more general *Robin boundary conditions*

$$-a(0)u'(0) = \gamma(0)(u_0 - u(0)), \quad a(1)u'(1) = \gamma(1)(u_1 - u(1))$$

with constants $\gamma(0)$, u_0, $\gamma(1)$, u_1.

216.3 Stationary Boundary Value Problems

In many situations, u is independent of time and the model reduces to the *stationary* reaction-diffusion equation

$$-(a(x)u'(x))' = f(x) \tag{216.7}$$

in the case of pure diffusion and

$$-(a(x)u'(x))' + (b(x)u(x))' = f(x) \tag{216.8}$$

in case there is convection as well. For these problems, we only need to specify boundary conditions. For example, we consider the *two-point boundary value problem*: find the function $u(x)$ satisfying

$$\begin{cases} -(au')' = f & \text{in } (0,1), \\ u(0) = u(1) = 0 \end{cases} \tag{216.9}$$

and when there is convection: find $u(x)$ such that

$$\begin{cases} -(au')' + (bu)' = f & \text{in } (0,1), \\ u(0,t) = u(1,t) = 0. \end{cases} \tag{216.10}$$

216.4 The Finite Element Method

We begin the discussion of discretization by studying the simplest model above, namely the two-point boundary value problem for the stationary reaction-diffusion model (216.9).

We can express the solution $u(x)$ of (216.9) analytically in terms of data by integrating twice (setting $w = au'$)

$$u(x) = \int_0^x \frac{w(y)}{a(y)}\,dy + \alpha_1, \quad w(y) = -\int_0^y f(z)\,dz + \alpha_2,$$

where the constants α_1 and α_2 are chosen so that $u(0) = u(1) = 0$. We can use this solution formula to compute the value of the solution $u(x)$ for any given $x \in (0,1)$ by evaluating the integrals analytically or numerically using quadrature. However, this is very time consuming if we want the solution at many points in $[0,1]$. This motivates consideration of an alternative way of computing the solution $u(x)$ using the *Finite Element Method* (FEM), which is a general method for solving differential equations numerically. FEM is based on rewriting the differential equation in *variational form* and seeking an approximate solution as a piecewise polynomial.

Note that we do not use the solution by integration outlined above, one important consequence of that procedure is that u is "twice as differentiable" as the data f, since we integrate twice to get from f to u.

We present FEM for (216.9) based on continuous piecewise linear approximation. We let $\mathcal{T}_h : 0 = x_0 < x_1 < ... < x_{M+1} = 1$, be a *partition* (or *triangulation*) of $I = (0,1)$ into sub-intervals $I_j = (x_{j-1}, x_j)$ of length $h_j = x_j - x_{j-1}$. We look for an approximate solution in the set V_h of continuous piecewise linear functions $v(x)$ on $\mathcal{T}_h$ such that $v(0) = 0$ and $v(1) = 0$. We show an example of such a function in Fig. 216.2. In Chapter 215, we

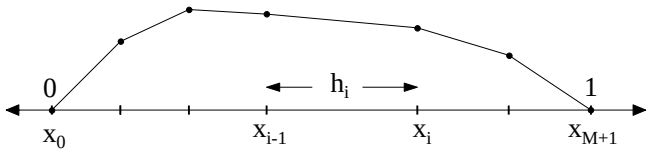


FIGURE 216.2. A continuous piecewise linear function in V_h.

saw that V_h is a finite dimensional vector space of dimension M with a basis consisting of the hat functions $\{\varphi_j\}_{j=1}^M$ illustrated in Fig. 215.5, associated with the *interior nodes* $x_1, \cdots, x_M$. The coordinates of a function v in V_h in this basis are the values $v(x_j)$ at the interior nodes since a function $v \in V_h$ can be written

$$v(x) = \sum_{j=1}^{M} v(x_j)\varphi_j(x).$$

Note that because $v \in V_h$ is zero at 0 and 1, we do not include φ_0 and φ_{M+1} in the set of basis functions for V_h.

The finite element method is based on restating the differential equation $-(au')' = f$ in an average or *variational* form

$$-\int_0^1 (au')'v\,dx = \int_0^1 fv\,dx, \tag{216.11}$$

where the function v varies over an appropriate set of *test functions*. The variational form results from multiplying the differential equation $-(au')' =$

f by the test function $v(x)$ and integrating over the interval $(0,1)$. The variational formulation says that the residual $-(au')'-f$ of the true solution is orthogonal to all test functions v with respect to the $L_2(0,1)$ scalar product.

The basic idea of FEM is to compute an approximate solution $U \in V_h$ that satisfies (216.11) for a restricted set of test functions. This approach to computing an approximate solution is known as the *Galerkin method* in memory of the Russian engineer and scientist Galerkin (1871-1945), see Fig. 216.3. He invented his method while imprisoned for anti-Tsarist activities during 1906-7. We call the set V_h, where we seek the FEM-solution U, the *trial space* and we call the space of test functions the *test space*. In the present case of homogeneous Dirichlet boundary conditions, we usually choose the test space to be equal to V_h. Consequently, the dimensions of the trial and test spaces are equal, which is necessary for the existence and uniqueness of the approximate solution U.

FIGURE 216.3. Boris Galerkin, inventor of the Finite Element Method: "It is really quite simple; just multiply by $v(x)$ and then integrate".

However since the functions in V_h do not have second derivatives, we can not simply plug a potential approximate solution U in V_h directly into (216.11). To get around this difficulty, we use integration by parts to move one derivative from $(au')'$ onto v, noting that functions in V_h are piecewise differentiable. Assuming v is differentiable and $v(0)=v(1)=0$:

$$-\int_0^1 (au')'v\,dx = -a(1)u'(1)v(1)+a(0)u'(0)v(0)+\int_0^1 au'v'\,dx$$
$$=\int_0^1 au'v'\,dx.$$

This leads to the *continuous Galerkin finite element method of order* 1 (*cG(1)-method*) for (216.9): compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V_h. \tag{216.12}$$

We note that the derivatives U' and v' of the functions U and $v \in V_h$ are piecewise constant functions of the form depicted in Fig. 216.4 and are not

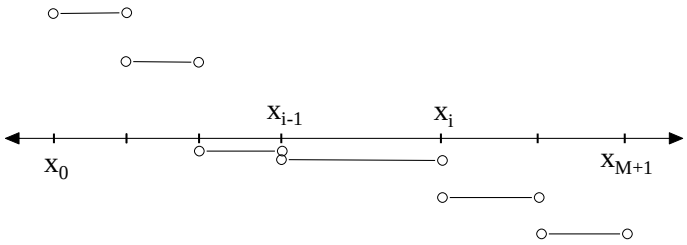


FIGURE 216.4. The derivative of the continuous piecewise linear function in Fig. 216.2.

defined at the nodes x_i. However, the value of an integral is independent of the value of the integrand at isolated points. Therefore, the integral (216.12) with integrand $aU'v'$ is uniquely defined as the sum of the integrals over the sub-intervals I_j.

Discretization of the Stationary Reaction-Diffusion-Convection Problem

To solve (216.10) numerically let $0 = x_0 < x_1 < ... < x_{L+1} = 1$ be a partition of $(0,1)$, and let V_h be the corresponding space of continuous piecewise linear functions $v(x)$ such that $v(0) = v(1) = 0$. The cG(1) FEM for (216.10) takes the form: compute $U \in V_h$ such that

$$\int_0^1 (aU')v' + (bU)'v)\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V_h.$$

216.5 The Discrete System of Equations

We have not yet proved that the set of equations (216.12) has a unique solution nor discussed what is involved in computing the solution U. This is an important issue considering we constructed the FEM precisely because the original problem is likely impossible to solve analytically.

We prove that the cG(1)-method (216.12) corresponds to a square linear system of equations for the unknown nodal values $\xi_j = U(x_j)$, $j = 1, ..., M$. We write U using the basis of hat functions as

$$U(x) = \sum_{j=1}^{M} \xi_j \varphi_j(x) = \sum_{j=1}^{M} U(x_j)\varphi_j(x).$$

Substituting into (216.12), we change the order of summation and integration to obtain

$$\sum_{j=1}^{M} \xi_j \int_0^1 a\varphi_j' v' \, dx = \int_0^1 fv \, dx, \tag{216.13}$$

for all $v \in V_h$. Now, it suffices to check (216.13) with v varying over the set of basis functions $\{\varphi_i\}_{i=1}^M$, since any function in V_h can be expressed as a linear combination of the basis functions. We are thus led to the $M \times M$ linear system of equations

$$\sum_{j=1}^{M} \xi_j \int_0^1 a\varphi_j' \varphi_i' \, dx = \int_0^1 f\varphi_i \, dx, \quad i = 1, ..., M, \tag{216.14}$$

for the unknown coefficients $\xi_1, ..., \xi_M$. We let $\xi = (\xi_1, ..., \xi_M)^\top$ denote the M-vector of unknown coefficients and define the $M \times M$ *stiffness matrix* $A = (a_{ij})$ with elements

$$a_{ij} = \int_0^1 a\varphi_j' \varphi_i' \, dx, \qquad i, j = 1, ..., M,$$

and the *load vector* $b = (b_i)$ with

$$b_i = \int_0^1 f\varphi_i \, dx, \qquad i = 1, ..., M.$$

These names originate from early applications of the finite element method in *structural mechanics* describing deformable structures like the body and wing of an aircraft or buildings. Using this notation, (216.14) is equivalent to the system of linear equations

$$A\xi = b. \tag{216.15}$$

In order to solve for the unknown vector ξ of nodal values of U, we first have to compute the stiffness matrix A and the load vector b. In the first instance, we assume that $a(x) = 1$ for $x \in [0, 1]$. We note that a_{ij} is zero unless $i = j - 1$, $i = j$, or $i = j + 1$ because otherwise either $\varphi_i(x)$ or $\varphi_j(x)$ is zero on each sub-interval occurring in the integration. We illustrate this in Fig. 216.5. We compute a_{ii} first. Using the definition of the hat function

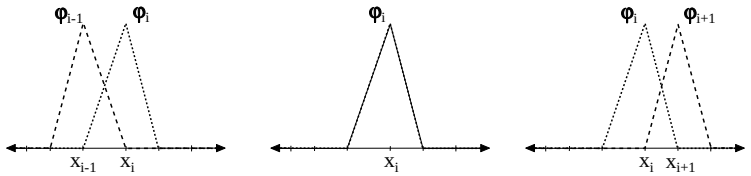


FIGURE 216.5. Three possibilities to obtain a non-zero element in the stiffness matrix.

φ_i,

$$\varphi_i(x) = \begin{cases} (x - x_{i-1})/h_i, & x_{i-1} \le x \le x_i, \\ (x_{i+1} - x)/h_{i+1}, & x_i \le x \le x_{i+1}, \\ 0, & \text{elsewhere,} \end{cases}$$

the integration breaks down into two integrals:

$$a_{ii} = \int_{x_{i-1}}^{x_i} \big(\frac{1}{h_i}\big)^2 \, dx + \int_{x_i}^{x_{i+1}} \big(\frac{-1}{h_{i+1}}\big)^2 \, dx = \frac{1}{h_i} + \frac{1}{h_{i+1}} \text{ for } i = 1, 2, ..., M,$$

since $\varphi_i' = 1/h_i$ on (x_{i-1}, x_i) and $\varphi_i' = -1/h_{i+1}$ on (x_i, x_{i+1}) and φ_i is zero on the other sub-intervals. Similarly,

$$a_{i\,i+1} = \int_{x_i}^{x_{i+1}} \frac{-1}{(h_{i+1})^2} \, dx = -\frac{1}{h_{i+1}} \text{ for } i = 1, 2, ..., M,$$

while $a_{i\,i-1} = -1/h_i$ for $i = 2, 3, ..., M$.

We compute the elements of the load vector of b in the same way to get

$$b_i = \int_{x_{i-1}}^{x_i} f(x) \frac{x - x_{i-1}}{h_i} dx + \int_{x_i}^{x_{i+1}} f(x) \frac{x_{i+1} - x}{h_{i+1}} dx, \quad i = 1, ..., M.$$

The matrix A is a *sparse* matrix in the sense that most of its entries are zero. In particular, A is a *banded* matrix with non-zero entries occurring only in the diagonal, super-diagonal and sub-diagonal positions. A is also called a tri-diagonal matrix. Moreover, A is a *symmetric* matrix since $\int_0^1 \varphi_i' \varphi_j' \, dx = \int_0^1 \varphi_j' \varphi_i' \, dx$. Finally, A is *positive-definite* in the sense that

$$\eta^\top A \eta = \sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j > 0,$$

unless $\eta_i = 0$ for $i = 1, ..., M$. This follows by noting that if $v(x) = \sum_{j=1}^{M} \eta_j \varphi_j(x)$ then by reordering the summation (check!)

$$\sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j = \sum_{i,j=1}^{M} \eta_i \int_0^1 a\varphi_j' \varphi_i' \, dx \, \eta_j$$
$$= \int_0^1 a \sum_{j=1}^{M} \eta_j \varphi_j' \sum_{i=1}^{M} \eta_i \varphi_i' \, dx = \int_0^1 a v'(x) v'(x) \, dx > 0$$

unless $v'(x) = 0$ for all $x \in [0, 1]$, that is $v(x) = 0$ for $x \in [0, 1]$, since $v(0) = 0$, that is $\eta_i = 0$ for $i = 1, ..., M$. This implies that A is invertible, so that (216.15) has a unique solution for all data b.

We sum up: the stiffness matrix A is sparse, symmetric and positive definite, and thus in particular the system $A\xi = b$ has a unique solution for all b.

We expect the accuracy of the approximate solution to increase as M increases since the work involved in solving for U increases. Systems of dimension $10^2 - 10^3$ in one space dimension and up to 10^6 in two or three space dimensions are common. An important issue is the efficient numerical solution of the system $A\xi = b$.

216.6 Handling Different Boundary Conditions

We consider briefly the discretization of the two-point boundary value problem $-(au')' = f$ in $(0, 1)$ with the different boundary conditions.

Non-Homogeneous Dirichlet Boundary Conditions

We begin with the boundary conditions $u(0) = u_0$ and $u(1) = u_1$, where u_0 and u_1 are given boundary values, where the conditions are non-homogeneous if $u_0 u_1 \neq 0$. In this situation, we compute an approximate solution in the trial space V_h of continuous piecewise linear functions $v(x)$ on a partition $\mathcal{T}_h : 0 = x_0 < x_1 < ... < x_{M+1} = 1$, satisfying the boundary conditions $v(0) = u_0$, $v(1) = u_1$, and we let the test functions vary over the space V_h^0 of continuous piecewise linear functions $v(x)$ satisfying the homogeneous boundary conditions $v(0) = v(1) = 0$. The trial and test spaces are different in this case, but we note that they have equal dimension (equal to the number M of internal nodes). Multiplying by a test function and integrating by parts, we are led to the following method: compute $U \in V_h$ such that

$$\int_0^1 aU'v' \, dx = \int_0^1 fv \, dx \quad \text{for all } v \in V_h^0. \tag{216.16}$$

As above this leads to a symmetric positive definite system of equations in the internal unknown nodal values $U(x_1), ..., U(x_M)$.

Neumann Boundary Conditions

We now consider the problem

$$\begin{cases} -(au')' = f, & \text{in } (0,1), \\ u(0) = 0,\ a(1)u'(1) = g_1, \end{cases} \tag{216.17}$$

with a non-homogeneous Neumann boundary condition at $x = 1$, which in the case of modeling heat in a wire, corresponds to prescribing the heat flux $a(1)u'(1)$ at $x = 1$ to be g_1.

To derive a variational formulation of this problem, we multiply the differential equation $-(au')' = f$ by a test function v and integrate by parts to get

$$\int_0^1 fv\,dx = -\int_0^1 (au')'v\,dx = \int_0^1 au'v'\,dx - a(1)u'(1)v(1) + a(0)u'(0)v(0).$$

Now $a(1)u'(1) = g_1$ is specified but $a(0)u'(0)$ is unknown. So it is convenient to assume that v satisfies the homogeneous Dirichlet condition $v(0) = 0$. Correspondingly, we define V_h to be the space of continuous functions v that are piecewise linear on a partition $\mathcal{T}_h$ of $(0,1)$ satisfying $v(0) = 0$. Replacing $a(1)u'(1)$ by g_1, we are led to the following FEM for (216.17): compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx = \int_0^1 fv\,dx + g_1v(1) \quad \text{for all } v \in V_h. \tag{216.18}$$

We substitute $U(x) = \sum_{i=1}^{M+1} \xi_i\varphi_i(x)$, noting that the value $\xi_{M+1} = U(x_{M+1})$ at the node x_{M+1} is now undetermined, into (216.18) and choose $v = \varphi_1, \cdots, \varphi_{M+1}$ to get a $(M+1) \times (M+1)$ system of equations for ξ. We show the form of the resulting stiffness matrix with $a = 1$ and load vector in Fig. 216.6. Note that the last equation

$$\frac{U(x_{M+1}) - U(x_M)}{h_{M+1}} = b_{M+1} + g_1$$

is a discrete analog of the boundary condition $u'(1) = g_1$ since $b_{M+1} \approx \frac{h_{M+1}}{2}f(1)$.

To conclude, a Neumann boundary condition, unlike a Dirichlet condition, is not explicitly enforced in the trial space. Instead, the Neumann condition is automatically satisfied as a consequence of the variational formulation by letting the test functions vary freely at the corresponding boundary point. In the case of Neumann boundary conditions, we thus simply can

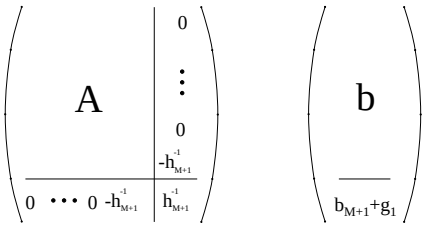


FIGURE 216.6. The stiffness matrix and load vector computed from (216.18) in the case that $a \equiv 1$. A and b are the stiffness matrix and load vector previously obtained in the problem with homogeneous Dirichlet boundary conditions and $b_{M+1} = \int_0^1 f\varphi_{M+1}\,dx$.

"forget" the boundary conditions in the definition of the trial space V_h and let the test space coincide with V_h. A Dirichlet boundary condition is called an *essential* boundary condition and a Neumann condition is called a *natural* boundary condition. An essential boundary condition is imposed explicitly in the definition of the trial space, i.e. it is a *strongly imposed* boundary condition, and the test space satisfy the corresponding homogeneous boundary condition. A natural boundary condition is not imposed in the trial space and becomes automatically satisfied through the variational formulation by letting the test functions vary freely at the corresponding boundary point.

Robin Boundary Conditions

A natural generalization of Neumann conditions for the problem $-(au')' = f$ in $(0,1)$ are called *Robin* boundary conditions. These take the form

$$-a(0)u'(0) = \gamma(0)(u_0 - u(0)), \quad a(1)u'(1) = \gamma(1)(u_1 - u(1)). \quad (216.19)$$

In the case of modeling heat in a wire, $\gamma(0)$ and $\gamma(1)$ are given (non-negative) boundary heat conductivities and u_0 and u_1 are given "outside temperatures". The Robin boundary condition at $x = 0$ states that the heat flux $-a(0)u'(0)$ is proportional to the temperature difference $u_0 - u(0)$ between the outside and inside temperature. If $u_0 > u(0)$ then heat will flow from outside to inside and if $u_0 < u(0)$ then heat will flow from inside out.

EXAMPLE 216.3. We may experience this kind of boundary condition with $\gamma(0)$ quite large in a poorly insulated house on a cold winter day. The size of the boundary heat conductivity γ is an important issue in the real estate business in the north of Sweden.

When $\gamma = 0$, (216.19) reduces to a homogeneous Neumann boundary condition. Conversely, letting γ tend to infinity, the Robin boundary condition

$-a(0)u'(0) = \gamma(0)(u_0 - u(0))$ approaches the Dirichlet boundary condition $u(0) = u_0$.

Robin boundary conditions are natural boundary conditions like Neumann conditions, Therefore, we let V_h be the space of continuous piecewise linear functions on a partition of $(0, 1)$ without any boundary conditions imposed. Multiplying the equation $-(au')' = f$ by a function $v \in V_h$ and integrating by parts, we get

$$\int_0^1 fv\,dx = -\int_0^1 (au')'v\,dx = \int_0^1 au'v'\,dx - a(1)u'(1)v(1) + a(0)u'(0)v(0).$$

Replacing $a(0)u'(0)$ and $a(1)u'(1)$ using the Robin boundary conditions, we get

$$\int_0^1 fv\,dx = \int_0^1 au'v'\,dx + \gamma(1)(u(1) - u_1)v(1) + \gamma(0)(u(0) - u_0)v(0).$$

Collecting data on the right hand side, we are led to the following cG(1) method: compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx + \gamma(0)u(0)v(0) + \gamma(1)u(1)v(1) = \int_0^1 fv\,dx + \gamma(0)u_0v(0) + \gamma(1)u_1v(1)$$

for all $v \in V_h$.

An even more general Robin boundary condition has the form $-a(0)u'(0) = \gamma(0)(u_0 - u(0)) + g_0$, where g_0 is a given heat flux. This Robin boundary condition thus includes Neumann boundary conditions ($\gamma = 0$) and Dirichlet boundary conditions (letting $\gamma \to \infty$). The implementation of a Robin boundary conditions is facilitated by the fact that the trial and test space are the same.

216.7 Error Estimates and Adaptive Error Control

When conducting scientific experiments in a laboratory or building a suspension bridge, for example, there is always a lot of worry about the errors in the process. In fact, if we were to summarize the philosophy behind the scientific revolution, a main component would be the modern emphasis on the quantitative analysis of error in measurements during experiments and the reporting of the errors along with the results. The same issue comes up in computational mathematical modeling: whenever we make a computation on a practical problem, we must be concerned with the accuracy of the results and the related issue of how to compute efficiently. These

issues naturally fit into a wider framework which also addresses how well the differential equation models the underlying physical situation and what effect errors in data and the model have on the conclusions we can draw from the results.

We address these issues by deriving two kinds of error estimates for the error $u - U$ of the finite element approximation. First we prove an *a priori* error estimate which shows that the Galerkin finite element method for (216.9) produces the best possible approximation in V_h of the solution u in a certain sense. If u has continuous second derivatives, then we know that V_h contains good approximations of u, for example the piecewise linear interpolant. So the a priori estimate implies that the error of the finite element approximation can be made arbitrarily small by refining the mesh provided that the solution u is sufficiently smooth to allow the interpolation error to go to zero as the mesh is refined. This kind of result is called an *a priori* error estimate because the error bound does not depend on the approximate solution to be computed. One the other hand, it does requires knowledge about the derivatives of the (unknown) exact solution.

After that, we prove an *a posteriori* error bound that bounds the error of the finite element approximation in terms of its residual error. This error bound can be evaluated once the finite element solution has been computed and used to estimate the error. Through the a posteriori error estimate, it is possible to estimate and adaptively control the finite element error to a desired tolerance level by suitably refining the mesh.

To measure the size of the error $e = u - U$, we shall use the *weighted* L_2 *norm*

$$\|w\|_a = \left(\int_0^1 a\, w^2 \, dx \right)^{1/2},$$

with *weight* a. More precisely we shall estimate the quantity

$$\|(u - U)'\|_a$$

which we refer to as the *energy norm* of the error $u - U$.

We will use the following variations of Cauchy's inequality with the weight a present:

$$\left| \int_0^1 a v' w' \, dx \right| \leq \|v'\|_a \|w'\|_a \quad \text{and} \quad \left| \int_0^1 vw \, dx \right| \leq \|v\|_a \|w\|_{a^{-1}}. \tag{216.20}$$

An A Priori Error Estimate

We shall prove that the finite element approximation $U \in V_h$ is the best approximation of u in V_h with respect to the energy norm. This is a consequence of the *Galerkin orthogonality* built into the finite element method expressed by

$$\int_0^1 a(u - U)' v' \, dx = 0 \quad \text{for all } v \in V_h \tag{216.21}$$

which results from subtracting (216.12) from (216.11) (integrated by parts) with $v \in V_h$. This is analogous to the best approximation property of the L_2 projection studied in the Chapter Piecewise linear approximation.

We have for any $v \in V_h$,

$$\begin{aligned}\|(u-U)'\|_a^2 &= \int_0^1 a(u-U)'(u-U)'\,dx \\ &= \int_0^1 a(u-U)'(u-v)'\,dx + \int_0^1 a(u-U)'(v-U)'\,dx \\ &= \int_0^1 a(u-U)'(u-v)'\,dx,\end{aligned}$$

where the last line follows because $v - U \in V_h$. Estimating using Cauchy's inequality, we get

$$\|(u-U)'\|_a^2 \le \|(u-U)'\|_a \|(u-v)'\|_a,$$

so that

$$\|(u-U)'\|_a \le \|(u-v)'\|_a \quad \text{for all } v \in V_h.$$

This is the best approximation property of U. We now choose in particular $v = \pi_h u$, where $\pi_h u \in V_h$ is the nodal interpolant of u, and use the following weighted analog of (215.11)

$$\|(u - \pi_h u)'\|_a \le C_i \|hu''\|_a,$$

where C_i is an interpolation constant that depends only on (the variation of) a. We then obtain the following error estimate.

Theorem 216.1 *The finite element approximation U satisfies $\|(u-U)'\|_a \le \|(u-v)'\|_a$ for all $v \in V_h$. In particular, there is a constant C_i depending only on a such that*

$$\|u' - U'\|_a \le C_i \|hu''\|_a.$$

This energy norm estimate says that the derivative of the error of the finite element approximation converges to zero at a first order rate in the mesh size h. By integration it follows that the error itself, say pointwise or in the L_2 norm, also tends to zero. One can also prove a more precise bound for the error $u - U$ itself that is second order in the mesh size h.

An A Posteriori Error Estimate

We shall now estimate the energy norm error $\|u' - U'\|_a$ in terms of the residual $R(U) = (aU')' + f$ of the finite element solution U on each subinterval. The residual measures how well U solves the differential equation and it is completely computable once U has been computed.

We start by using the variational form of (216.11) with $v = e = u - U$ to find an expression for $\|u - U\|_a^2$:

$$\begin{aligned}\|e'\|_a^2 &= \int_0^1 ae'e'\,dx = \int_0^1 au'e'\,dx - \int_0^1 aU'e'\,dx \\ &= \int_0^1 fe\,dx - \int_0^1 aU'e'\,dx.\end{aligned}$$

We then use (216.12), with $v = \pi_h e$ denoting the nodal interpolant of e in V_h, to obtain

$$\begin{aligned}\|e'\|_a^2 &= \int_0^1 f\,(e - \pi_h e)\,dx - \int_0^1 aU'(e - \pi_h e)'\,dx \\ &= \int_0^1 f\,(e - \pi_h e)\,dx - \sum_{j=1}^{M+1} \int_{I_j} aU'(e - \pi_h e)'\,dx.\end{aligned}$$

Now, we integrate by parts over each sub-interval I_j in the last term and use the fact that all the boundary terms disappear because $(e - \pi_h e)(x_j) = 0$ to get the *error representation formula*

$$\|e'\|_a^2 = \int_0^1 R(U)(e - \pi_h e)\,dx, \tag{216.22}$$

where the residual error $R(U)$ is the discontinuous function defined on $(0, 1)$ by

$$R(U) = f + (aU')' \quad \text{on each sub-interval } I_j.$$

From the weighted Cauchy inequality (216.20) (inserting factors h and h^{-1}), we get

$$\|e'\|_a^2 \le \|hR(U)\|_{a^{-1}} \|h^{-1}(e - \pi_h e)\|_a.$$

One can prove the following analog of the second estimate of (215.11)

$$\|h^{-1}(e - \pi_h e)\|_a \le C_i \|e'\|_a,$$

where C_i is an interpolation constant depending on a, and we notice the appearance of the factor h^{-1} on the left hand side. This proves the basic a posteriori error estimate:

Theorem 216.2 *There is an interpolation constant C_i depending only on a such that the finite element approximation U satisfies*

$$\|u' - U'\|_a \le C_i \|hR(U)\|_{a^{-1}}. \tag{216.23}$$

Adaptive Error Control

Since the a posteriori error estimate (216.23) indicates the size of the error of an approximation on a given mesh in terms of computable information, it is natural to try to use this information to compute an accurate approximation. This is the basis of *adaptive error control*.

The computational problem that arises once a two-point boundary value problem is specified is to find a mesh such that the finite element approximation achieves a given level of accuracy, or in other words, such that the error of the approximation is bounded by an *error tolerance* TOL. In practice, we are also concerned with efficiency, which means that we want to determine a mesh with the fewest number of elements that yields an approximation with the desired accuracy. We try to reach this optimal mesh by starting with a coarse mesh and successively refining based on the size of the a posteriori error estimate. By starting with a coarse mesh, we try to keep the number of elements as small as possible.

More precisely, we choose an initial mesh $\mathcal{T}_h$, compute the corresponding cG(1) approximation U, and then check whether or not

$$C_i\|hR(U)\|_{a^{-1}} \leq \text{TOL}.$$

This is the *stopping criterion*, which guarantees that $\|u' - U'\|_a \leq \text{TOL}$ by (216.23). Therefore when the stopping criterion is satisfied, U is sufficiently accurate. If the stopping criterion is not satisfied, we try to construct a new mesh $\mathcal{T}_{\tilde{h}}$ of mesh size $\tilde{h}$ with as few elements as possible such that

$$C_i\|\tilde{h}R(U)\|_{a^{-1}} = \text{TOL}.$$

This is the *mesh modification criterion* from which the new mesh size $\tilde{h}$ is computed based on the size of the residual error $R(U)$ of the approximation on the old mesh. In order to minimize the number of mesh points, it turns out that the mesh size should be chosen to *equidistribute* the residual error in the sense that the contribution from each element to the integral giving the total residual error is roughly the same. In practice, this means that elements with large residual errors are refined, while elements in intervals where the residual error is small are combined together to form bigger elements.

We repeat the adaptive cycle of mesh modification followed by solution on the new mesh until the stopping criterion is satisfied. By the a priori error estimate, we know that if u'' is bounded then the error tends to zero as the mesh is refined. Hence, the stopping criterion will be satisfied eventually. In practice, the adaptive error control rarely requires more than a few iterations.

216.8 Discretization of Time-Dependent Reaction-Diffusion-Convection Problems

We now return to original time dependent problem (216.6).

To solve (216.6) numerically, we apply the cG(1) method for time discretization and the cG(1) FEM for discretization in space. More precisely, let $0 = x_0 < x_1 < ... < x_{L+1} = 1$ be a partition of $(0,1)$, and let V_h be the corresponding space of continuous piecewise linear functions $v(x)$ such that $v(0) = v(1) = 0$. Let $0 = t_0 < t_1 < t_2 < ... < t_N = T$ be a sequence of discrete time levels with corresponding time intervals $I_n = (t_{n-1}, t_n)$ and time steps $k_n = t_n - t_{n-1}$, for $n = 1, ..., N$. We look for a numerical solution $U(x,t)$ that is linear in t on each time interval I_n. For $n = 1, ..., N$, we compute $U^n \in V_h$ such that for all $v \in V_h$,

$$\begin{aligned}&\int_{I_n}\int_0^1 \dot{U}v\,dx\,dt + \int_{I_n}\int_0^1 (aU')v' + (bU)'v)\,dx\,dt\\ &= \int_{I_n}\int_0^1 fv\,dx\,dt + \int_{I_n}(g(0,t)v(0) + g(1,t)v(1))\,dt,\end{aligned} \tag{216.24}$$

where $U(t_n, x) = U^n(x)$ denotes the *time nodal value* for $n = 1, 2, ..., N$ and $U^0 = u_0$, assuming that $u_0 \in V_h$. Since U is linear on each time interval, it is determined completely once we have computed its nodal values.

Arguing as above using the expansion in terms of the basis functions for V_h leads to a sequence of systems of equations for $n = 1, ..., N$,

$$MU^n + k_n A_n U^n = MU^{n-1} + k_n b^n, \tag{216.25}$$

where M is the mass matrix corresponding to V_h and A_n is a stiffness matrix related to time interval I_n. Solving this system successively for $n = 1, 2, ..., N$, we obtain an approximate solution U of (216.10).

216.9 Non-Linear Reaction-Diffusion-Convection Problems

In many situations, the coefficients or data depend on the solution u, which leads to a nonlinear problem. For example if f depends on u, we get a problem of the form

$$\begin{cases}\dot{u} - (au')' + (bu)' = f(u) & \text{in } (0,1)\times(0,T),\\ u(0,t) = u(1,t) = 0, & \text{for } t \in (0,T),\\ u(x,0) = u_0(x) & \text{for } x \in (0,1).\end{cases} \tag{216.26}$$

Discretization as above eventually yields a discrete system of the form

$$MU^n + k_n A_n U^n = MU^{n-1} + k_n b^n(U^n), \tag{216.27}$$

where b^n depends on U^n. This nonlinear system may be solved by fixed point iteration or Newton's method.

We conclude this section by presenting some examples of systems of nonlinear reaction-diffusion-convection problems arising in physics, chemistry and biology. These systems may be solved numerically by a direct extension of the cG(1) method in space and time presented above. In all examples, a and the α_i are a positive constants.

EXAMPLE 216.4. *The bistable equation for ferro-magnetism*

$$\dot{u} - au'' = u - u^3, \tag{216.28}$$

with a small.

EXAMPLE 216.5. *Model of a superconductivity of a fluid*

$$\begin{aligned} \dot{u}_1 - au_1'' &= (1 - |u|^2)u_1, \\ \dot{u}_2 - au_2'' &= (1 - |u|^2)u_2. \end{aligned} \tag{216.29}$$

EXAMPLE 216.6. *Model of flame propagation*

$$\begin{aligned} \dot{u}_1 - au_1'' &= -u_1 e^{-\alpha_1/u_2}, \\ \dot{u}_2 - au_2'' &= \alpha_2 u_1 e^{-\alpha_1/u_2}. \end{aligned} \tag{216.30}$$

EXAMPLE 216.7. *Field-Noyes equations for chemical reactions*

$$\begin{aligned} \dot{u}_1 - au_1'' &= \alpha_1(u_2 - u_1u_3 + u_1 - \alpha_2 u_1^2), \\ \dot{u}_2 - au_2'' &= \alpha^{-1}(\alpha_3 u_3 - u_2 - u_1u_2), \\ \dot{u}_2 - au_2'' &= \alpha_4(u_1 - u_3). \end{aligned} \tag{216.31}$$

EXAMPLE 216.8. *Spread of rabies in foxes*

$$\begin{aligned} \dot{u}_1 - au_1'' &= \alpha_1(1 - u_1 - u_2 - u_3) - u_3u_1, \\ \dot{u}_2 - au_2'' &= u_3u_1 - (\alpha_2 + \alpha_3 + \alpha_1u_1 + \alpha_1u_1 + \alpha_1u_3)u_2, \\ \dot{u}_2 - au_2'' &= \alpha_2u_2 - (\alpha_4 + \alpha_1u_1 + \alpha_1u_1 + \alpha_1u_3)u_3, \end{aligned} \tag{216.32}$$

where $\alpha_4 < (1 + (\alpha_3 + \alpha_1)/\alpha_2)^{-1} - \alpha_1$.

EXAMPLE 216.9. *Interaction of two species*

$$\begin{aligned} \dot{u}_1 - au_1'' &= u_1M(u_1, u_2), \\ \dot{u}_2 - au_2'' &= u_2N(u_1, u_2), \end{aligned} \tag{216.33}$$

where $M(u_1, u_2)$ and $N(u_1, u_2)$ are given functions describing various situations such as (i) predator-prey ($M_{u_2} < 0$, $N_{u_1} > 0$) (ii) competing species ($M_{u_2} < 0$, $N_{u_1} < 0$) and (iii) symbiosis ($M_{u_2} > 0$, $N_{u_1} > 0$).

EXAMPLE 216.10. *Morphogenesis of patterns (zebra or tiger)*

$$\begin{aligned} \dot{u}_1 - a u_1'' &= -u_1 u_2^2 + \alpha_1(1 - u_1) \\ \dot{u}_2 - a u_2'' &= u_1 u_2^2 - (\alpha_1 + \alpha_2) u_2. \end{aligned} \tag{216.34}$$

EXAMPLE 216.11. *Fitz-Hugh-Nagumo model for transmission of axons*

$$\begin{aligned} \dot{u}_1 - a u_1'' &= -u_1(u_1 - \alpha_1)(u_1 - 1) - u_2 \\ \dot{u}_2 - a u_2'' &= \alpha_2 u_1 - \alpha_3 u_2, \end{aligned} \tag{216.35}$$

$0 < \alpha_1 < 1$.

Chapter 216 Problems

216.1. Compute the stiffness matrix and load vector for the cG(1) method on a uniform partition for (216.9) with $a(x) = 1+x$ and $f(x) = \sin(x)$. Use quadrature if exact integration is inconvenient.

216.2. Formulate the cG(1) method for the problem $-(au')' + cu = f$ in $(0, 1)$, $u(0) = u(1) = 0$, where $a(x)$ and $c(x)$ are positive coefficients. Compute the corresponding stiffness matrix when $a = c = 1$, assuming a uniform partition. Is the stiffness matrix still symmetric, positive-definite, and tridiagonal?

216.3. Determine the resulting system of equations corresponding to the cG(1) method (216.16) with non-homogeneous Dirichlet boundary conditions.

216.4. Prove a priori and a posteriori error estimates for cG(1) for $-(au')' = f$ in $(0, 1)$ with Robin boundary conditions (a positive).

216.5. Prove a priori and a posteriori error estimates for cG(1) for $-(au')'+cu = f$ in $(0, 1)$ with Robin boundary conditions (a and c positive).

> The "classical" phase of my career was summed up in the book *The Large Scale Structure of Spacetime* which Ellis and I wrote in 1973. I would not advise readers of this book to consult that work for further information: it is highly technical, and quite unreadable. I hope that since then I have learned how to write in a manner that is easier to understand. (Stephen Hawking in *A Brief History of Time*)

Part XIII

MultiD Calculus

(Partly from Applied Mathematics Body and Soul, Vol 3, Springer 2003, coauthored with Kenneth Eriksson and Don Estep).

$$\Delta u = \nabla \cdot \nabla u$$
$$\frac{\partial u}{\partial t} - \Delta u = f$$
$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f$$
$$i\frac{\partial u}{\partial t} = -\frac{1}{2}\Delta u + Vu$$
$$\frac{\partial u}{\partial t} + u \cdot \nabla u + \nabla p - \nu \Delta u = f,$$
$$\nabla \cdot u = 0$$
$$\nabla \times H = J, \quad \nabla \cdot B = 0, \quad B = \mu H.$$

217

Vector-Valued Functions of Several Real Variables

Auch die Chemiker müssen sich allmänlich an den Gedanken gewöhnen, dass ihnen die theoretische Chemie ohne die Beherrschung der Elemente der höheren Analysis ein Buch mit sieben Siegeln blieben wirt. Ein Differential- oder Integralzeichen muss aufhören, für den Chemiker eine unvertständliche Hieroglyphe zu sein,... wenn er sich nicht der Gefahr aussetzen will, für die Entwicklung der theoretischen Chemie jedes Verständnis zu verlieren. (H. Jahn, Grundriss der Elektrochemie, 1895)

217.1 Introduction

We now turn to the extension of the basic concepts of real-valued functions of one real variable, such as Lipschitz continuity and differentiability, to vector-valued functions of several variables. We have carefully prepared the material so that this extension will be as natural and smooth as possible. We shall see that the proofs of the basic theorems like the Chain rule, the Mean Value theorem, Taylor's theorem, the Contraction Mapping theorem and the Inverse Function theorem, extend almost word by word to the more complicated situation of vector valued functions of several real variables.

We consider functions $f : \mathbb{R}^n \to \mathbb{R}^m$ that are vector valued in the sense that the value $f(x) = (f_1(x), ..., f_m(x))$ is a vector in $\mathbb{R}^m$ with components $f_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, ..., m$, where with $f_i(x) = f_i(x_1, ..., x_n)$ and $x = (x_1, ..., x_n) \in \mathbb{R}^n$. As usual, we view $x = (x_1, ..., x_n)$ as a n-column vector and $f(x) = (f_1(x), ..., f_m(x))$ as a m-column vector.

As particular examples of vector-valued functions, we first consider *curves*, which are functions $g : \mathbb{R} \to \mathbb{R}^n$, and *surfaces*, which are functions $g : \mathbb{R}^2 \to \mathbb{R}^n$. We then discuss composite functions $f \circ g : \mathbb{R} \to \mathbb{R}^m$, where $g : \mathbb{R} \to \mathbb{R}^n$ is a curve and $f : \mathbb{R}^n \to \mathbb{R}^m$, with $f \circ g$ again being a curve. We recall that $f \circ g(t) = f(g(t))$.

The inputs to the functions reside in the n dimensional vector space $\mathbb{R}^n$ and it is worthwhile to consider the properties of $\mathbb{R}^n$. Of particular importance is the notion of Cauchy sequence and convergence for sequences $\{x^{(j)}\}_{j=1}^{\infty}$ of vectors $x^{(j)} = (x_1^{(j)}, ..., x_n^{(j)}) \in \mathbb{R}^n$ with coordinates $x_k^{(j)}$, $k = 1, ..., n$. We say that the sequence $\{x^{(j)}\}_{j=1}^{\infty}$ is a *Cauchy sequence* if for all $\epsilon > 0$, there is a natural number N so that

$$\|x^{(i)} - x^{(j)}\| \leq \epsilon \quad \text{for } i, j > N.$$

Here $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^n$, that is, $\|x\| = (\sum_{i=1}^n x_i^2)^{1/2}$. Sometimes, it is convenient to work with the norms $\|x\|_1 = \sum_{i=1}^n |x_i|$ or $\|x\|_\infty = \max_{i=1,...,n} |x_i|$. We say that the sequence $\{x^{(j)}\}_{j=1}^{\infty}$ of vectors in $\mathbb{R}^n$ *converges* to $x \in \mathbb{R}^n$ if for all $\epsilon > 0$, there is a natural number N so that

$$\|x - x^{(i)}\| \leq \epsilon \quad \text{for } i > N.$$

It is easy to show that a convergent sequence is a Cauchy sequence and conversely that a Cauchy sequence converges. We obtain these results applying the corresponding results for sequences in $\mathbb{R}$ to each of the coordinates of the vectors in $\mathbb{R}^n$.

EXAMPLE 217.1. The sequence $\{x^{(i)}\}_{i=1}^{\infty}$ in $\mathbb{R}^2$, $x^{(i)} = (1-i^{-2}, \exp(-i))$, converges to $(1, 0)$.

217.2 Curves in $\mathbb{R}^n$

A function $g : I \to \mathbb{R}^n$, where $I = [a, b]$ is an interval of real numbers, is a *curve* in $\mathbb{R}^n$, see Fig. 217.1. If we use t as the independent variable ranging over I, then we say that the curve $g(t)$ is *parametrized* by the variable t. We also refer to the set of points $\Gamma = \{g(t) \in \mathbb{R}^n : t \in I\}$ as the curve Γ parameterized by the function $g : I \to \mathbb{R}^n$.

EXAMPLE 217.2. The simplest example of a curve is a straight line. The function $g : \mathbb{R} \to \mathbb{R}^2$ given by

$$g(t) = \bar{x} + tz,$$

where $z \in \mathbb{R}^2$ and $\bar{x} \in \mathbb{R}^2$, is a straight line in $\mathbb{R}^2$ through the point $\bar{x}$ with direction z, see Fig. 217.2.

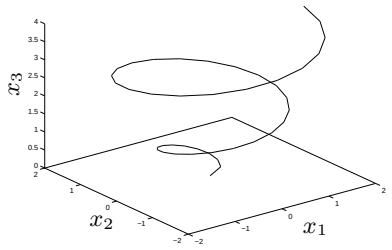


FIGURE 217.1. The curve $g : [0, 4] \to \mathbb{R}^3$ with $g(t) = \big(t^{1/2}\cos(\pi t), t^{1/2}\sin(\pi t), t\big)$.

EXAMPLE 217.3. Let $f : [a, b] \to \mathbb{R}$ be given, and define $g : [a, b] \to \mathbb{R}^2$ by $g(t) = (g_1(t), g_2(t)) = (t, f(t))$. This curve is simply the graph of the function $f : [a, b] \to \mathbb{R}$, see Fig. 217.2.

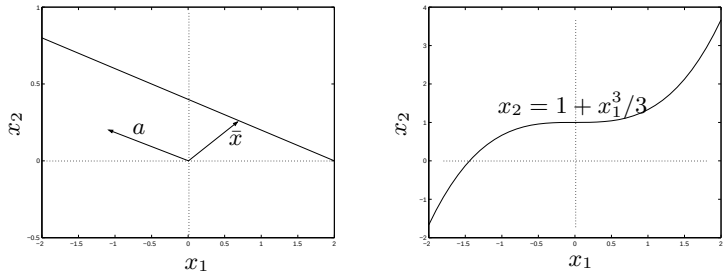


FIGURE 217.2. On the left: the curve $g(t) = \bar{x} + ta$. On the right: a curve $g(t) = (t, f(t))$.

217.3 Different Parameterizations of a Curve

It is possible to use different parametrizations for the set of points forming a curve. If $h : [c, d] \to [a, b]$ is a one-to-one mapping, then the composite function $f = g \circ h : [c, d] \to \mathbb{R}^2$ is a *reparameterization* of the curve $\{g(t) : t \in [a, b]\}$ given by $g : [a, b] \to \mathbb{R}^2$.

EXAMPLE 217.4. The function $f : [0, \infty) \to \mathbb{R}^3$ given by

$$f(\tau) = (\tau\cos(\pi\tau^2), \tau\sin(\pi\tau^2), \tau^2),$$

is a reparameterization of the curve $g : [0, \infty) \to \mathbb{R}^3$ given by

$$g(t) = (\sqrt{t}\cos(\pi t), \sqrt{t}\sin(\pi t), t),$$

obtained setting $t = h(\tau) = \tau^2$. We have $f = g \circ h$.

217.4 Surfaces in $\mathbb{R}^n$, $n \geq 3$

A function $g : Q \to \mathbb{R}^n$, where $n \geq 3$ and Q is a subdomain of $\mathbb{R}^2$, may be viewed to be a *surface* S in $\mathbb{R}^n$, see Fig. 217.3. We write $g = g(y)$ with $y = (y_1, y_2) \in Q$ and say that S is parameterized by $y \in Q$. We may also identify the surface S with the set of points $S = \{g(y) \in \mathbb{R}^n : y \in Q\}$, and reparameterize S by $f = g \circ h : \tilde{Q} \to \mathbb{R}^n$ if $h : \tilde{Q} \to Q$ is a one-to-one mapping of a domain $\tilde{Q}$ in $\mathbb{R}^2$ onto Q.

EXAMPLE 217.5. The simplest example of a surface $g : \mathbb{R}^2 \to \mathbb{R}^3$ is a plane in $\mathbb{R}^3$ given by

$$g(y) = g(y_1, y_2) = \bar{x} + y_1 b_1 + y_2 b_2, \quad y \in \mathbb{R}^2,$$

where $\bar{x}$, b_1, $b_2 \in \mathbb{R}^3$.

EXAMPLE 217.6. Let $f : [0,1] \times [0,1] \to \mathbb{R}$ be given, and define $g : [0,1] \times [0,1] \to \mathbb{R}^3$ by $g(y_1, y_2) = (y_1, y_2, f(y_1, y_2))$. This is a surface, which is the graph of $f : [0,1] \times [0,1] \to \mathbb{R}$. We also refer to this surface briefly as the surface given by the function $x_3 = f(x_1, x_2)$ with $(x_1, x_2) \in [0,1] \times [0,1]$.

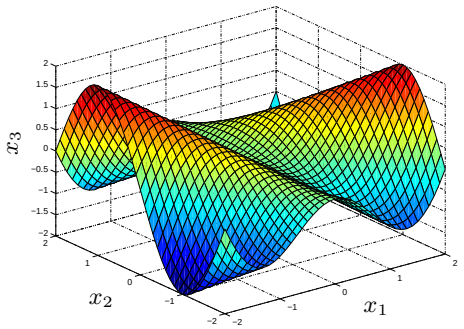


FIGURE 217.3. The surface $s(y_1, y_2) = \big(y_1, y_2, y_1 \sin\big((y_1 + y_2)\pi/2\big)\big)$ with $-1 \leq y_1, y_2 \leq 1$, or briefly the surface $x_3 = x_1 \sin\big((x_1 + x_2)\pi/2\big)$ with $-1 \leq x_1, x_2 \leq 1$.

217.5 Lipschitz Continuity

We say that $f : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous on $\mathbb{R}^n$ if there is a constant L such that

$$\|f(x) - f(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n. \tag{217.1}$$

This definition extends easily to functions $f : A \to \mathbb{R}^m$ with the domain $D(f) = A$ being a subset of $\mathbb{R}^n$. For example, A may be the unit n-cube $[0,1]^n = \{x \in \mathbb{R}^n : 0 \leq x_i \leq 1, i = 1, ..., n\}$ or the unit n-disc $\{x \in \mathbb{R}^n : \|x\| \leq 1\}$.

To check if a function $f : A \to \mathbb{R}^m$ is Lipschitz continuous on some subset A of $\mathbb{R}^n$, it suffices to check that the component functions $f_i : A \to \mathbb{R}$ are Lipschitz continuous. This is because

$$|f_i(x) - f_i(y)| \leq L_i\|x - y\| \quad \text{for } i = 1, ..., m,$$

implies

$$\|f(x) - f(y)\|^2 = \sum_{i=1}^m |f_i(x) - f_i(y)|^2 \leq \sum_{i=1}^m L_i^2\|x - y\|^2,$$

which shows that $\|f(x) - f(y)\| \leq L\|x - y\|$ with $L = (\sum_i L_i^2)^{\frac{1}{2}}$.

EXAMPLE 217.7. The function $f : [0,1] \times [0,1] \to \mathbb{R}^2$ defined by $f(x_1, x_2) = (x_1 + x_2, x_1x_2)$, is Lipschitz continuous with Lipschitz constant $L = 2$. To show this, we note that $f_1(x_1, x_2) = x_1 + x_2$ is Lipschitz continuous on $[0,1] \times [0,1]$ with Lipschitz constant $L_1 = \sqrt{2}$ because $|f_1(x_1, x_2) - f_1(y_1, y_2)| \leq |x_1 - y_1| + |x_2 - y_2| \leq \sqrt{2}\|x - y\|$ by Cauchy's inequality. Similarly, $f_2(x_1, x_2) = x_1x_2$ is Lipschitz continuous on $[0,1] \times [0,1]$ with Lipschitz constant $L_2 = \sqrt{2}$ since $|x_1x_2 - y_1y_2| = |x_1x_2 - y_1x_2 + y_1x_2 - y_1y_2| \leq |x_1 - y_1| + |x_2 - y_2| \leq \sqrt{2}\|x - y\|$.

EXAMPLE 217.8. The function $f : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$f(x_1, ..., x_n) = (x_n, x_{n-1}, ..., x_1),$$

is Lipschitz continuous with Lipschitz constant $L = 1$.

EXAMPLE 217.9. A linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ given by an $m \times n$ matrix $A = (a_{ij})$, with $f(x) = Ax$ and x a n-column vector, is Lipschitz continuous with Lipschitz constant $L = \|A\|$. We made this observation in Chapter *Analytic geometry in* $\mathbb{R}^n$. We repeat the argument:

$$L = \max_{x \neq y} \frac{\|f(x) - f(y)\|}{\|x - y\|} = \max_{x \neq y} \frac{\|Ax - Ay\|}{\|x - y\|}$$
$$= \max_{x \neq y} \frac{\|A(x - y)\|}{\|x - y\|} = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \|A\|.$$

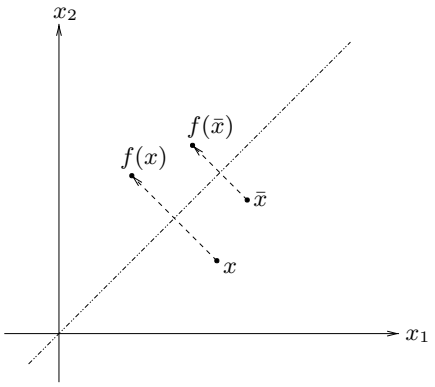


FIGURE 217.4. Illustration of the mapping $f(x_1, x_2) = (x_2, x_1)$, which is clearly Lipschitz continuous with $L = 1$.

Concerning the definition of the matrix norm $\|A\|$, we note that the function $F(x) = \|Ax\|/\|x\|$ is homogeneous of degree zero, that is, $F(\lambda x) = F(x)$ for all non-zero real numbers λ, and thus $\|A\|$ is the maximum value of $F(x)$ on the closed and bounded set $\{x \in \mathbb{R}^n : \|x\| = 1\}$, which is a finite real number.

We recall that if A is a diagonal $n \times n$ matrix with diagonal elements λ_i, then $\|A\| = \max_i |\lambda_i|$.

217.6 Differentiability: Jacobian, Gradient and Tangent

We say that $f : \mathbb{R}^n \to \mathbb{R}^m$ is *differentiable at* $\bar{x} \in \mathbb{R}^n$ if there is a $m \times n$ matrix $M(\bar{x}) = (m_{ij}(\bar{x}))$, called the *Jacobian* of the function $f(x)$ at $\bar{x}$, and a constant $K_f(\bar{x})$ such that for all x close to $\bar{x}$,

$$f(x) = f(\bar{x}) + M(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}), \tag{217.2}$$

where $E_f(x, \bar{x}) = (E_f(x, \bar{x})_i)$ is an m-vector satisfying $\|E_f(x, \bar{x})\|$ $\leq K_f(\bar{x})\|x - \bar{x}\|^2$. We also denote the Jacobian by $Df(\bar{x})$ or $f'(\bar{x})$ so that $M(\bar{x}) = Df(\bar{x}) = f'(\bar{x})$. Since $f(x)$ is a m-column vector, or $m \times 1$ matrix, and x is a n-column vector, or $n \times 1$ matrix, $M(\bar{x})(x - \bar{x})$ is the product of the $m \times n$ matrix $M(\bar{x})$ and the $n \times 1$ matrix $x - \bar{x}$ yielding a $m \times 1$ matrix or a m-column vector.

We say that $f : A \to \mathbb{R}^m$, where A is a subset of $\mathbb{R}^n$, is *differentiable on* A if $f(x)$ is differentiable at $\bar{x}$ for all $\bar{x} \in A$. We say that $f : A \to \mathbb{R}^m$ is *uniformly differentiable on* A if the constant $K_f(\bar{x}) = K_f$ can be chosen independently of $\bar{x} \in A$.

FIGURE 217.5. Carl Jacobi (1804-51);: "It is often more convenient to possess the ashes of great men than to possess the men themselves during their lifetime" (on the return of Descarte's remains to France).

We now show how to determine a specific element $m_{ij}(\bar{x})$ of the Jacobian using the relation (217.2). We consider the coordinate function $f_i(x_1, ..., x_n)$ and setting $x = \bar{x} + se_j$, where e_j is the j^{th} standard basis vector and s is a small real number, we focus on the variation of $f_i(x_1, ..., x_n)$ as the variable x_j varies in a neighborhood of $\bar{x}_j$. The relation (217.2) states that for small non-zero real numbers s,

$$f_i(\bar{x} + se_j) = f_i(\bar{x}) + m_{ij}(\bar{x})s + E_f(\bar{x} + se_j, \bar{x})_i, \tag{217.3}$$

where $\|x - \bar{x}\|^2 = \|se_j\|^2 = s^2$ implies

$$|E_f(\bar{x} + se_j, \bar{x})_i| \leq K_f(\bar{x})s^2.$$

Note that by assumption $\|E_f(x, \bar{x})\| \leq K_f(\bar{x})\|x - \bar{x}\|^2$, and so each coordinate function $E_f(\bar{x} + se_j, \bar{x})_i$ satisfies $|E_f(x, \bar{x})_i| \leq K_f(\bar{x})\|x - \bar{x}\|^2$.

Now, dividing by s in (217.3) and letting s tend to zero, we find that

$$m_{ij}(\bar{x}) = \lim_{s \to 0} \frac{f_i(\bar{x} + se_j) - f_i(\bar{x})}{s}, \tag{217.4}$$

which we can also write as

$$m_{ij}(\bar{x}) = \lim_{x_j \to \bar{x}_j} \frac{f_i(\bar{x}_1, ..., \bar{x}_{j-1}, x_j, \bar{x}_{j+1}, ..., \bar{x}_n) - f_i(\bar{x}_1, ..., \bar{x}_{j-1}, \bar{x}_j, \bar{x}_{j+1}, ..., \bar{x}_n)}{x_j - \bar{x}_j}. \tag{217.5}$$

We refer to $m_{ij}(\bar{x})$ as the *partial derivative* of f_i with respect to x_j at $\bar{x}$, and we use the alternative notation $m_{ij}(\bar{x}) = \frac{\partial f_i}{\partial x_j}(\bar{x})$. To compute $\frac{\partial f_i}{\partial x_j}(\bar{x})$ we freeze all coordinates at $\bar{x}$ but the coordinate x_j and then let x_j vary in a neighborhood of $\bar{x}_j$. The formula

$$\frac{\partial f_i}{\partial x_j}(\bar{x}) = \lim_{x_j \to \bar{x}_j} \frac{f_i(\bar{x}_1, ..., \bar{x}_{j-1}, x_j, \bar{x}_{j+1}, ..., \bar{x}_n) - f_i(\bar{x}_1, ..., \bar{x}_{j-1}, \bar{x}_j, \bar{x}_{j+1}, ..., \bar{x}_n)}{x_j - \bar{x}_j}, \tag{217.6}$$

states that we compute the partial derivative with respect to the variable x_j by keeping all the other variables x_1,...,x_{j-1}, x_{j+1},...,x_n constant. Thus, computing partial derivatives should be a pleasure using our previous expertise of computing derivatives of functions of one real variable!

We may express the computation alternatively as follows:

$$\frac{\partial f_i}{\partial x_j}(\bar{x}) = m_{ij}(\bar{x}) = g'_{ij}(0) = \frac{dg_{ij}}{ds}(0), \tag{217.7}$$

where $g_{ij}(s) = f_i(\bar{x} + se_j)$.

EXAMPLE 217.10. Let $f : \mathbb{R}^3 \to \mathbb{R}$ be given by $f(x_1, x_2, x_3) = x_1 e^{x_2} \sin(x_3)$. We compute

$$\frac{\partial f}{\partial x_1}(\bar{x}) = e^{\bar{x}_2} \sin(\bar{x}_3), \quad \frac{\partial f}{\partial x_2}(\bar{x}) = \bar{x}_1 e^{\bar{x}_2} \sin(\bar{x}_3), \quad \frac{\partial f}{\partial x_3}(\bar{x}) = \bar{x}_1 e^{\bar{x}_2} \cos(\bar{x}_3),$$

and thus

$$f'(\bar{x}) = (e^{\bar{x}_2} \sin(\bar{x}_3), \bar{x}_1 e^{\bar{x}_2} \sin(\bar{x}_3), \bar{x}_1 e^{\bar{x}_2} \cos(\bar{x}_3))$$

EXAMPLE 217.11. If $f : \mathbb{R}^3 \to \mathbb{R}^2$ is given by $f(x) = \begin{pmatrix} \exp(x_1^2 + x_2^2) \\ \sin(x_2 + 2x_3) \end{pmatrix}$, then

$$f'(x) = \begin{pmatrix} 2x_1 \exp(x_1^2 + x_2^2) & 2x_2 \exp(x_1^2 + x_2^2) & 0 \\ 0 & \cos(x_2 + 2x_3) & 2\cos(x_2 + 2x_3) \end{pmatrix}.$$

We have now shown how to compute the elements of a Jacobian using the usual rules for differentiation with respect to one real variable. This opens a whole new world of applications to explore. The setting is thus a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$ satisfying for suitable $x, \bar{x} \in \mathbb{R}^n$:

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}), \tag{217.8}$$

with $\|E_f(x,\bar{x})\| \leq K_f(\bar{x})\|x-\bar{x}\|^2$, where $f'(\bar{x}) = Df(\bar{x})$ is the Jacobian $m \times n$ matrix with elements $\frac{\partial f_i}{\partial x_j}$:

$$f'(\bar{x}) = Df(\bar{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\bar{x}) & \frac{\partial f_1}{\partial x_2}(\bar{x}) & ... & \frac{\partial f_1}{\partial x_n}(\bar{x}) \\ \frac{\partial f_2}{\partial x_1}(\bar{x}) & \frac{\partial f_2}{\partial x_2}(\bar{x}) & ... & \frac{\partial f_2}{\partial x_n}(\bar{x}) \\ ... & ... & ... & \\ \frac{\partial f_m}{\partial x_1}(\bar{x}) & \frac{\partial f_m}{\partial x_2}(\bar{x}) & ... & \frac{\partial f_m}{\partial x_n}(\bar{x}) \end{pmatrix}.$$

Sometimes we use the following notation for the Jacobian $f'(x)$ of a function $y = f(x)$ with $f : \mathbb{R}^n \to \mathbb{R}^m$:

$$f'(x) = \frac{dy_1, ..., dy_m}{dx_1, ..., dx_n}(x) \tag{217.9}$$

The function $x \to \hat{f}(x) = f(\bar{x}) + f'(\bar{x})(x-\bar{x})$ is called the *linearization* of the function $x \to f(x)$ at $x = \bar{x}$. We have

$$\hat{f}(x) = f'(\bar{x})x + f(\bar{x}) - f'(\bar{x})\bar{x} = Ax + b,$$

with $A = f'(\bar{x})$ a $m \times n$ matrix and $b = f(\bar{x}) - f'(\bar{x})\bar{x}$ a m-column vector. We say that $\hat{f}(x)$ is an *affine transformation*, which is a transformation of the form $x \to Ax + b$, where x is a n-column vector, A is a $m \times n$ matrix and b is a m-column vector. The Jacobian $\hat{f}'(x)$ of the linearization $\hat{f}(x) = Ax + b$ is a constant matrix equal to the matrix A, because the partial derivatives of Ax with respect to x are simply the elements of the matrix A.

If $f : \mathbb{R}^n \to \mathbb{R}$, that is $m = 1$, then we also denote the Jacobian f' by ∇f, that is,

$$f'(\bar{x}) = \nabla f(\bar{x}) = (\frac{\partial f}{\partial x_1}(\bar{x}), ..., \frac{\partial f}{\partial x_n}(\bar{x})).$$

In words, $\nabla f(\bar{x})$ is the n-row vector or $1 \times n$ matrix of partial derivatives of $f(x)$ with respect to x_1, x_2,...,x_n at $\bar{x}$. We refer to $\nabla f(\bar{x})$ as the *gradient* of $f(x)$ at $\bar{x}$. If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\bar{x}$, we thus have

$$f(x) = f(\bar{x}) + \nabla f(\bar{x})(x-\bar{x}) + E_f(x,\bar{x}), \tag{217.10}$$

with $|E_f(x,\bar{x})| \leq K_f(\bar{x})\|x-\bar{x}\|^2$, and $\hat{f}(x) = f(\bar{x}) + \nabla f(\bar{x})(x-\bar{x})$ is the linearization of $f(x)$ at $x = \bar{x}$. We may alternatively express the product $\nabla f(\bar{x})(x-\bar{x})$ of the n-row vector ($1 \times n$ matrix) $\nabla f(\bar{x})$ with the n-column vector ($n \times 1$ matrix) $(x-\bar{x})$ as the scalar product $\nabla f(\bar{x}) \cdot (x-\bar{x})$ of the n-vector $\nabla f(\bar{x})$ with the n-vector $(x-\bar{x})$. We thus often write (217.10) in the form

$$f(x) = f(\bar{x}) + \nabla f(\bar{x}) \cdot (x-\bar{x}) + E_f(x,\bar{x}). \tag{217.11}$$

EXAMPLE 217.12. If $f : \mathbb{R}^3 \to \mathbb{R}$ is given by $f(x) = x_1^2 + 2x_2^3 + 3x_3^4$, then

$$\nabla f(x) = (2x_1, 6x_2^2, 12x_3^3).$$

EXAMPLE 217.13. The equation $x_3 = f(x)$ with $f : \mathbb{R}^2 \to \mathbb{R}$ and $x = (x_1, x_2)$ represents a surface in $\mathbb{R}^3$ (the graph of the function f). The linearization

$$x_3 = f(\bar{x}) + \nabla f(\bar{x}) \cdot (x - \bar{x}) = f(\bar{x}) + \frac{\partial f}{\partial x_1}(\bar{x})(x_1 - \bar{x}_1) + \frac{\partial f}{\partial x_2}(\bar{x})(x_2 - \bar{x}_2)$$

with $\bar{x} = (\bar{x}_1, \bar{x}_2)$, represents the *tangent plane* at $x = \bar{x}$, see Fig. 217.13.

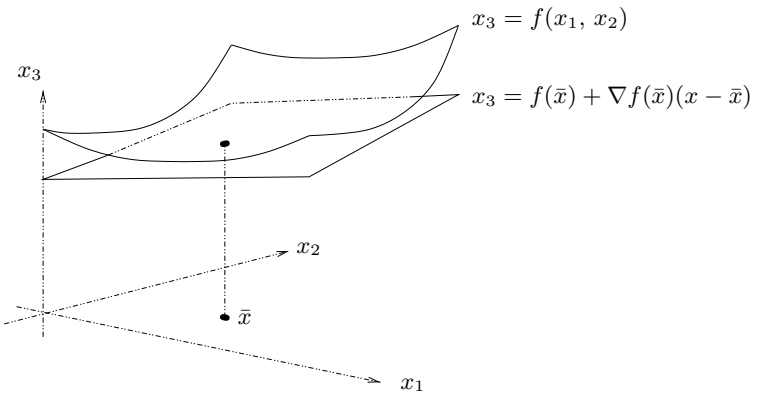


FIGURE 217.6. The surface $x_3 = f(x_1, x_2)$ and its tangent plane.

EXAMPLE 217.14. Consider now a curve $f : \mathbb{R} \to \mathbb{R}^m$, that is, $f(t) = (f_1(t), ..., f_m(t))$ with $t \in \mathbb{R}$ and we have a situation with $n = 1$. The linearization $t \to \hat{f}(t) = f(\bar{t}) + f'(\bar{t})(t - \bar{t})$ at $\bar{t}$ represents a straight line in $\mathbb{R}^m$ through the point $f(\bar{t})$ and the Jacobian $f'(\bar{t}) = (f_1'(\bar{t}), ..., f_m'(\bar{t}))$ gives the direction of the *tangent* to the curve $f : \mathbb{R} \to \mathbb{R}^m$ at $f(\bar{t})$, see Fig. 217.7.

217.7 The Chain Rule

Let $g : \mathbb{R}^n \to \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}^p$ and consider the composite function $f \circ g : \mathbb{R}^n \to \mathbb{R}^p$ defined by $f \circ g(x) = f(g(x))$. Under suitable assumptions of differentiability and Lipschitz continuity, we shall prove a *Chain rule* generalizing the Chain rule of Chapter *Differentiation rules* in the case $n = m = p = 1$. Using linearizations of f and g, we have

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

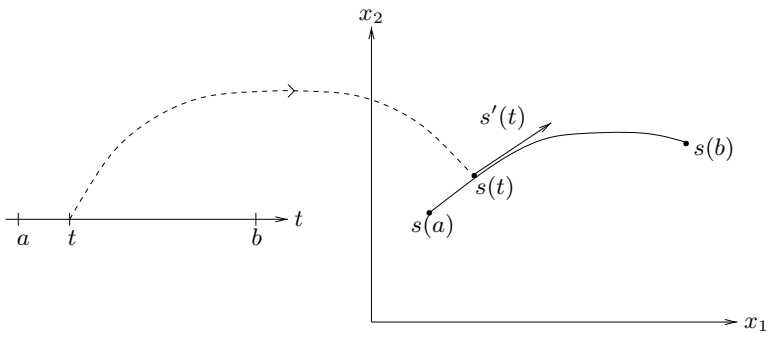


FIGURE 217.7. The tangent $s'(t)$ to a curve given by $s(t)$.

where we may naturally assume that

$$\|E_f(g(x), g(\bar{x}))\| \leq K_f\|g(x) - g(\bar{x})\|^2 \leq K_f L_g^2\|x - \bar{x}\|^2,$$

and $\|f'(g(\bar{x}))E_g(x, \bar{x})\| \leq \|f'(g(\bar{x}))\|K_g\|x - \bar{x}\|^2$, with suitable constants of differentiability K_f and K_g and Lipschitz constant L_g. We have now proved:

Theorem 217.1 (The Chain rule) *If $g : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\bar{x} \in \mathbb{R}^n$, and $f : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $g(\bar{x}) \in \mathbb{R}^m$ and further $g : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous, then the composite function $f \circ g : \mathbb{R}^n \to \mathbb{R}^p$ is differentiable at $\bar{x} \in \mathbb{R}^n$ with Jacobian*

$$(f \circ g)'(\bar{x}) = f'(g(\bar{x}))g'(\bar{x}).$$

The Chain rule has a wealth of applications and we now turn to harvest a couple of the most basic examples.

217.8 The Mean Value Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on $\mathbb{R}^n$ with a Lipschitz continuous gradient, and for given $x, \bar{x} \in \mathbb{R}^n$ consider the function $h : \mathbb{R} \to \mathbb{R}$ defined by

$$h(t) = f(\bar{x} + t(x - \bar{x})) = f \circ g(t),$$

with $g(t) = \bar{x} + t(x - \bar{x})$ representing the straight line through $\bar{x}$ and x. We have

$$f(x) - f(\bar{x}) = h(1) - h(0) = h'(\bar{t}),$$

for some $\bar{t} \in [0, 1]$, where we applied the usual Mean Value theorem to the function $h(t)$. By the Chain rule we have

$$h'(t) = \nabla f(g(t)) \cdot g'(t) = \nabla f(g(t)) \cdot (x - \bar{x}),$$

and we have now proved:

Theorem 217.2 (Mean Value theorem) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on $\mathbb{R}^n$ with a Lipschitz continuous gradient ∇f. Then for given x and $\bar{x}$ in $\mathbb{R}^n$, there is $y = x + \bar{t}(x - \bar{x})$ with $\bar{t} \in [0, 1]$, such that*

$$f(x) - f(\bar{x}) = \nabla f(y) \cdot (x - \bar{x}).$$

With the help of the Mean Value theorem we express the difference $f(x) - f(\bar{x})$ as the scalar product of the gradient $\nabla f(y)$ with the difference $x - \bar{x}$, where y is a point somewhere on the straight line between x and $\bar{x}$.

We may extend the Mean Value theorem to a function $f : \mathbb{R}^n \to \mathbb{R}^m$ to take the form

$$f(x) - f(\bar{x}) = f'(y)(x - \bar{x}),$$

where y is a point on the straight line between x and $\bar{x}$, which may be different for different rows of $f'(y)$. We may then estimate:

$$\|f(x) - f(\bar{x})\| = \|f'(y) \cdot (x - \bar{x})\| \leq \|f'(y)\| \|x - \bar{x}\|,$$

and we may thus estimate the Lipschitz constant of f by $\max_y \|f'(y)\|$ with $\|f'(y)\|$ the (Euclidean) matrix norm of $f'(y)$.

EXAMPLE 217.15. Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(x) = \sin(\sum_{j=1}^n x_j)$. We have

$$\frac{\partial f}{\partial x_i}(\bar{x}) = \cos(\sum_{j=1}^n \bar{x}_j) \quad \text{for } i = 1, ..., n,$$

and thus $|\frac{\partial f}{\partial x_i}(\bar{x})| \leq 1$ for $i = 1, ..., n$, and therefore

$$\|\nabla f(\bar{x})\| \leq \sqrt{n}.$$

We conclude that $f(x) = \sin(\sum_{j=1}^n x_j)$ is Lipschitz continuous with Lipschitz constant $\sqrt{n}$.

217.9 Direction of Steepest Descent and the Gradient

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a given function and suppose we want to study the variation of $f(x)$ in a neighborhood of a given point $\bar{x} \in \mathbb{R}^n$. More precisely, let x vary on the line through $\bar{x}$ in a given direction $z \in \mathbb{R}^n$, that is assume that $x = \bar{x} + tz$ where t is a real variable varying in a neighborhood of 0. Assuming f to be differentiable, the linearization formula (217.8) implies

$$f(x) = f(\bar{x}) + t\nabla f(\bar{x}) \cdot z + E_f(x, \bar{x}), \tag{217.12}$$

where $|E_f(x, \bar{x})| \leq t^2 K_f \|z\|^2$ and $\nabla f(\bar{x}) \cdot z$ is the scalar product of the gradient $\nabla f(\bar{x}) \in \mathbb{R}^n$ and the vector $z \in \mathbb{R}^n$. If $\nabla f(\bar{x}) \cdot z \neq 0$, then the

linear term $t\nabla f(\bar{x}) \cdot z$ will dominate the quadratic term $E_f(x, \bar{x})$ for small t. So the linearization

$$\hat{f}(x) = f(\bar{x}) + t\nabla f(\bar{x}) \cdot z$$

will be a good approximation of $f(x)$ for $x = \bar{x} + tz$ close to $\bar{x}$. Thus if $\nabla f(\bar{x}) \cdot z \neq 0$, then we get good information on the variation of $f(x)$ along the line $x = \bar{x} + tz$ by studying the linear function $t \to f(\bar{x}) + t\nabla f(\bar{x}) \cdot z$ with slope $\nabla f(\bar{x}) \cdot z$. In particular, if $\nabla f(\bar{x}) \cdot z > 0$ and $x = \bar{x} + tz$ then $\hat{f}(x)$ increases as we increase t and decreases as we decrease t. Similarly, if $\nabla f(\bar{x}) \cdot z < 0$ and $x = \bar{x} + tz$ then $\hat{f}(x)$ decreases as we increase t and increases as we decrease t.

Alternatively, we may consider the composite function $F_z : \mathbb{R} \to \mathbb{R}$ defined by $F_z(t) = f(g_z(t))$ with $g_z : \mathbb{R} \to \mathbb{R}^n$ given by $g_z(t) = \bar{x} + tz$. Obviously, $F_z(t)$ describes the variation of $f(x)$ on the straight line through $\bar{x}$ with direction z, with $F_z(0) = f(\bar{x})$. Of course, the derivative $F_z'(0)$ gives important information on this variation close to $\bar{x}$. By the Chain rule we have

$$F_z'(0) = \nabla f(\bar{x})z = \nabla f(\bar{x}) \cdot z,$$

and we retrieve $\nabla f(\bar{x}) \cdot z$ as a quantity of interest. In particular, the sign of $\nabla f(\bar{x}) \cdot z$ determines if $F_z(t)$ is increasing or decreasing at $t = 0$.

We may now ask how to choose the direction z to get maximal increase or decrease. We assume $\nabla f(\bar{x}) \neq 0$ to avoid the trivial case with $F_z'(0) = 0$ for all z. It is then natural to normalize z so $\|z\| = 1$ and we study the quantity $F_z'(0) = \nabla f(\bar{x}) \cdot z$ as we vary $z \in \mathbb{R}^n$ with $\|z\| = 1$. We conclude that the scalar product $\nabla f(\bar{x}) \cdot z$ is maximized if we choose z in the direction of the gradient $\nabla f(\bar{x})$,

$$z = \frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|},$$

which is called the direction of *steepest ascent*. For this gives

$$\max_{\|z\|=1} F_z'(0) = \nabla f(\bar{x}) \cdot \frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|} = \|\nabla f(\bar{x})\|.$$

Similarly, the scalar product is minimized if we choose z in the opposite direction of the gradient $\nabla f(\bar{x})$,

$$z = -\frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|},$$

which is called the direction of *steepest descent*, see Fig. 217.8. For then

$$\min_{\|z\|=1} F_z'(0) = -\nabla f(\bar{x}) \cdot \frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|} = -\|\nabla f(\bar{x})\|.$$

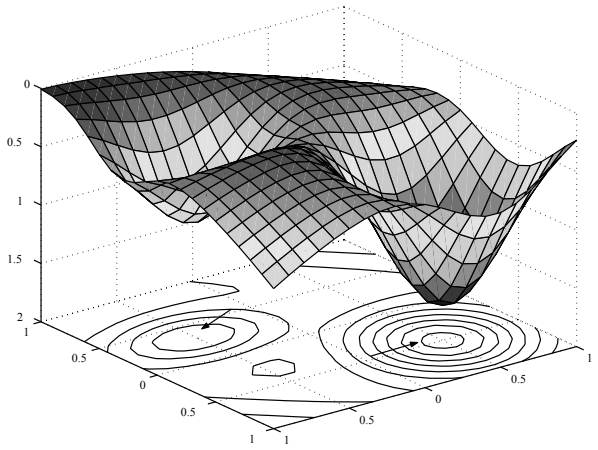


FIGURE 217.8. Directions of steepest descent on a "hiking map".

If $\nabla f(\bar{x}) = 0$, then $\bar{x}$ is said to be a *stationary point*. If $\bar{x}$ is a stationary point, then evidently $\nabla f(\bar{x}) \cdot z = 0$ for any direction z and

$$f(x) = f(\bar{x}) + E_f(x, \bar{x}).$$

The difference $f(x) - f(\bar{x})$ is then quadratically small in the distance $\|x - \bar{x}\|$, that is $|f(x) - f(\bar{x})| \leq K_f \|x - \bar{x}\|^2$, and $f(x)$ is very close to the constant value $f(\bar{x})$ for x close to $\bar{x}$.

217.10 A Minimum Point Is a Stationary Point

Suppose $\bar{x} \in \mathbb{R}^n$ is a *minimum point* for the function $f : \mathbb{R}^n \to \mathbb{R}$, that is

$$f(x) \geq f(\bar{x}) \quad \text{for } x \in \mathbb{R}^n. \tag{217.13}$$

We shall show that if $f(x)$ is differentiable at a minimum point $\bar{x}$, then

$$\nabla f(\bar{x}) = 0. \tag{217.14}$$

For if $\nabla f(\bar{x}) \neq 0$, then we could move in the direction of steepest descent from $\bar{x}$ to a point x close to $\bar{x}$ with $f(x) < f(\bar{x})$, contradicting (217.13). Consequently, in order to find minimum points of a function $f : \mathbb{R}^n \to \mathbb{R}$, we are led to try to solve the equation $g(x) = 0$, where $g = \nabla f : \mathbb{R}^n \to \mathbb{R}^n$. Here, we interpret $\nabla f(x)$ as a n-column vector.

A whole world of applications in mechanics, physics and other areas may be formulated as solving equations of the form $\nabla f(x) = 0$, that is as finding stationary points. We shall meet many applications below.

217.11 The Method of Steepest Descent

Let $f : \mathbb{R}^n \to \mathbb{R}$ be given and consider the problem of finding a minimum point $\bar{x}$. To do so it is natural to try a *method of Steepest Descent*: Given an approximation $\bar{y}$ of $\bar{x}$ with $\nabla f(\bar{y}) \neq 0$, we move from $\bar{y}$ to a new point y in the direction of steepest descent:

$$y = \bar{y} - \alpha \frac{\nabla f(\bar{y})}{\|\nabla f(\bar{y})\|},$$

where $\alpha > 0$ is a step length to be chosen. We know that $f(y)$ decreases as α increases from 0 and the question is just to find a reasonable value of α. This can be done by increasing α in small steps until $f(y)$ doesn't decrease anymore. The procedure is then repeated with $\bar{y}$ replaced by y. Evidently, the method of Steepest Descent is closely connected to Fixed Point Iteration for solving the equation $\nabla f(x) = 0$ in the form

$$x = x - \alpha \nabla f(x)$$

where we let $\alpha > 0$ include the normalizing factor $1/\|\nabla f(\bar{y})\|$.

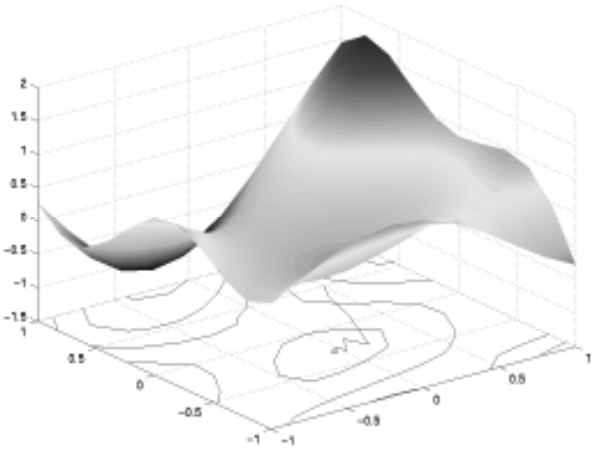


FIGURE 217.9. The method of Steepest Descent for $f(x_1, x_2) = x_1 \sin(x_1 + x_2) + x_2 \cos(2x_1 - 3x_2)$ starting at $(.5, .5)$ with $\alpha = .3$.

217.12 Directional Derivatives

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$, let $g_z(t) = \bar{x} + tz$ with $z \in \mathbb{R}^n$ a given vector normalized to $\|z\| = 1$, and consider the composite function $F_z(t) =$

$f(\bar{x}+tz)$. The Chain rule implies

$$F_z'(0) = \nabla f(\bar{x}) \cdot z,$$

and

$$\nabla f(\bar{x}) \cdot z$$

is called the *derivative of* $f(x)$ *in the direction* z at $\bar{x}$, see Fig. 217.10.

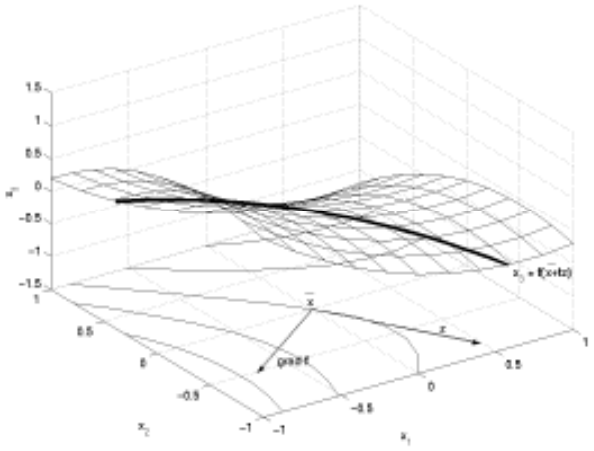


FIGURE 217.10. Illustration of directional derivative.

217.13 Higher Order Partial Derivatives

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on $\mathbb{R}^n$. Each partial derivative $\frac{\partial f}{\partial x_i}(\bar{x})$ is a function of $\bar{x} \in \mathbb{R}^n$ may be itself be differentiable. We denote its partial derivatives by

$$\frac{\partial}{\partial x_j}\frac{\partial f}{\partial x_i}(\bar{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\bar{x}), \quad i,j = 1,...,n,\ \bar{x} \in \mathbb{R}^n,$$

which are called the *partial derivatives of* f *of second order at* $\bar{x}$. It turns out that under appropriate continuity assumptions, the order of differentiation does not matter. In other words, we shall prove that

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(\bar{x}) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\bar{x}).$$

We carry out the proof in the case $n = 2$ with $i = 1$ and $j = 2$. We rewrite the expression

$$A = f(x_1,x_2) - f(\bar{x}_1,x_2) - f(x_1,\bar{x}_2) + f(\bar{x}_1,\bar{x}_2), \tag{217.15}$$

as

$$A = f(x_1, x_2) - f(x_1, \bar{x}_2) - f(\bar{x}_1, x_2) + f(\bar{x}_1, \bar{x}_2), \tag{217.16}$$

by shifting the order of the two mid terms. First, we set $F(x_1, x_2) = f(x_1, x_2) - f(\bar{x}_1, x_2)$ and use (217.15) to write

$$A = F(x_1, x_2) - F(x_1, \bar{x}_2).$$

The Mean Value theorem implies

$$A = \frac{\partial F}{\partial x_2}(x_1, y_2)(x_2 - \bar{x}_2) = \big(\frac{\partial f}{\partial x_2}(x_1, y_2) - \frac{\partial f}{\partial x_2}(\bar{x}_1, y_2)\big)(x_2 - \bar{x}_2)$$

for some $y_2 \in [\bar{x}_2, x_2]$. We use the Mean value theorem once again to get

$$A = \frac{\partial^2 f}{\partial x_1 \partial x_2}(y_1, y_2)(x_1 - \bar{x}_1)(x_2 - \bar{x}_2),$$

with $y_1 \in [\bar{x}_1, x_1]$. We next rewrite A using (217.16) in the form

$$A = G(x_1, x_2) - G(\bar{x}_1, x_2),$$

where $G(x_1, x_2) = f(x_1, x_2) - f(x_1, \bar{x}_2)$. Using the Mean Value theorem twice as above, we obtain

$$A = \frac{\partial^2 f}{\partial x_2 \partial x_1}(z_1, z_2)(x_1 - \bar{x}_1)(x_2 - \bar{x}_2),$$

where $z_i \in [\bar{x}_i, x_i]$, $i = 1, 2$. Assuming the second partial derivatives are Lipschitz continuous at $\bar{x}$ and letting x_i tend to $\bar{x}_i$ for $i = 1, 2$ gives

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(\bar{x}) = \frac{\partial^2 f}{\partial x_2 \partial x_1}(\bar{x}).$$

We have proved the following fundamental result:

Theorem 217.3 *If the partial derivatives of second order of a function $f : \mathbb{R}^n \to \mathbb{R}$ are all Lipschitz continuous, then the order of application of the derivatives of second order is irrelevant.*

The result directly generalizes to higher order partial derivatives: if the derivatives are Lipschitz continuous, then the order of application doesn't matter. What a relief!

217.14 Taylor's Theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz continuous partial derivatives of order 2. For given $x, \bar{x} \in \mathbb{R}^n$, consider the function $h : \mathbb{R} \to \mathbb{R}$ defined by

$$h(t) = f(\bar{x} + t(x - \bar{x})) = f \circ g(t),$$

where $g(t) = \bar{x} + t(x - \bar{x})$ is the straight line through $\bar{x}$ and x. Clearly $h(1) = f(x)$ and $h(0) = f(\bar{x})$, so the Taylor's theorem applied to $h(t)$ gives

$$h(1) = h(0) + h'(0) + \frac{1}{2}h''(\bar{t}),$$

for some $\bar{t} \in [0, 1]$. We compute using the Chain rule:

$$h'(t) = \nabla f(g(t)) \cdot (x - \bar{x}) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(g(t))(x_i - \bar{x}_i),$$

and similarly by a further differentiation with respect to t:

$$h''(t) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(g(t))(x_i - \bar{x}_i)(x_j - \bar{x}_j).$$

We thus obtain

$$f(x) = f(\bar{x}) + \nabla f(\bar{x}) \cdot (x - \bar{x}) + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(y)(x_i - \bar{x}_i)(x_j - \bar{x}_j), \quad (217.17)$$

for some $y = \bar{x} + \bar{t}(x - \bar{x})$ with $t \in [0, 1]$. The $n \times n$ matrix $H(\bar{x}) = (h_{ij}(\bar{x}))$ with elements $h_{ij}(\bar{x}) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\bar{x})$ is called the *Hessian* of $f(x)$ at $x = \bar{x}$. The Hessian is the matrix of all second partial derivatives of $f : \mathbb{R}^n \to \mathbb{R}$. With matrix vector notation with x a n-column vector, we can write

$$\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(y)(x_i - \bar{x}_i)(x_j - \bar{x}_j) = (x - \bar{x})^\top H(y)(x - \bar{x}).$$

We summarize:

Theorem 217.4 (Taylor's theorem) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable with Lipschitz continuous Hessian $H = (h_{ij})$ with elements $h_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$. Then, for given x and $\bar{x} \in \mathbb{R}^n$, there is $y = x + \bar{t}(x - \bar{x})$ with $\bar{t} \in [0, 1]$, such that*

$$\begin{aligned} f(x) &= f(\bar{x}) + \nabla f(\bar{x}) \cdot (x - \bar{x}) + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(y)(x_i - \bar{x}_i)(x_j - \bar{x}_j) \\ &= f(\bar{x}) + \nabla f(\bar{x}) \cdot (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^\top H(y)(x - \bar{x}). \end{aligned}$$

217.15 The Contraction Mapping Theorem

We shall now prove the following generalization of the Contraction Mapping theorem.

Theorem 217.5 *If $g : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with Lipschitz constant $L < 1$, then the equation $x = g(x)$ has a unique solution $\bar{x} = \lim_{i\to\infty} x^{(i)}$, where $\{x^{(i)}\}_{i=1}^{\infty}$ is a sequence in $\mathbb{R}^n$ generated by Fixed Point Iteration: $x^{(i)} = g(x^{(i-1)})$, $i = 1, 2, ...$, starting with any initial value $x^{(0)}$.*

The proof is word by word the same as in the case $g : \mathbb{R} \to \mathbb{R}$ considered in Chapter *Fixed Points and Contraction Mappings*. We repeat the proof for the convenience of the reader. Subtracting the equation $x^{(k)} = g(x^{(k-1)})$ from $x^{(k+1)} = g(x^{(k)})$, we get

$$x^{(k+1)} - x^{(k)} = g(x^{(k)}) - g(x^{(k-1)}),$$

and using the Lipschitz continuity of g, we thus have

$$\|x^{(k+1)} - x^{(k)}\| \le L\|x^{(k)} - x^{(k-1)}\|.$$

Repeating this estimate, we find that

$$\|x^{(k+1)} - x^{(k)}\| \le L^k\|x^{(1)} - x^{(0)}\|,$$

and thus for $j > i$

$$\begin{aligned}\|x^{(i)} - x^{(j)}\| &\le \sum_{k=i}^{j-1} \|x^{(k)} - x^{(k+1)}\| \\ &\le \|x^{(1)} - x^{(0)}\| \sum_{k=i}^{j-1} L^k = \|x^{(1)} - x^{(0)}\| L^i \, \frac{1 - L^{j-i}}{1 - L}.\end{aligned}$$

Since $L < 1$, $\{x^{(i)}\}_{i=1}^{\infty}$ is a Cauchy sequence in $\mathbb{R}^n$, and therefore converges to a limit $\bar{x} = \lim_{i\to\infty} x^{(i)}$. Passing to the limit in the equation $x^{(i)} = g(x^{(i-1)})$ shows that $\bar{x} = g(\bar{x})$ and thus $\bar{x}$ is a fixed point of $g : \mathbb{R}^n \to \mathbb{R}^n$. Uniqueness follows from the fact that if $\bar{y} = g(\bar{y})$, then $\|\bar{x} - \bar{y}\| = \|g(\bar{x}) - g(\bar{y})\| \le L\|\bar{x} - \bar{y}\|$ which is impossible unless $\bar{y} = \bar{x}$, because $L < 1$.

Example 217.16. Consider the function $g : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $g(x) = (g_1(x), g_2(x))$ with

$$g_1(x) = \frac{1}{4 + |x_1| + |x_2|}, \quad g_2(x) = \frac{1}{4 + |\sin(x_1)| + |\cos(x_2)|}.$$

We have

$$|\frac{\partial g_i}{\partial x_j}| \le \frac{1}{16},$$

and thus by simple estimates

$$\|g(x) - g(y)\| \leq \frac{1}{4}\|x - y\|,$$

which shows that $g : \mathbb{R}^2 \to \mathbb{R}^2$ is Lipschitz continuous with Lipschitz constant $L_g \leq \frac{1}{4}$. The equation $x = g(x)$ thus has a unique solution.

217.16 Solving $f(x) = 0$ with $f : \mathbb{R}^n \to \mathbb{R}^n$

The Contraction Mapping theorem can be applied as follows. Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ is given and we want to solve the equation $f(x) = 0$. Introduce

$$g(x) = x - Af(x),$$

where A is some non-singular $n \times n$ matrix with constant coefficients to be chosen. The equation $x = g(x)$ is then equivalent to the equation $f(x) = 0$. If $g : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with Lipschitz constant $L < 1$, then $g(x)$ has a unique fixed point $\bar{x}$ and thus $f(\bar{x}) = 0$. We have

$$g'(x) = I - Af'(x),$$

and thus we are led to choose the matrix A so that

$$\|I - Af'(x)\| \leq 1$$

for x close to the root $\bar{x}$. The ideal choice seems to be:

$$A = f'(\bar{x})^{-1},$$

assuming that $f'(\bar{x})$ is non-singular, since then $g'(\bar{x}) = 0$. In applications, we may seek to choose A close to $f'(\bar{x})^{-1}$ with the hope that the corresponding $g'(x) = I - Af'(x)$ will have $\|g'x)\|$ small for x close to the root $\bar{x}$, leading to a quick convergence. In Newton's method we choose $A = f'(x)^{-1}$, see below.

EXAMPLE 217.17. Consider the initial value problem $\dot{u}(t) = f(u(t))$ for $t > 0$, $u(0) = u_0$, where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a given Lipschitz continuous function with Lipschitz constant L_f, and as usual $\dot{u} = \frac{du}{dt}$. Consider the backward Euler method

$$U(t_i) = U(t_{i-1}) + k_i f(U(t_i)), \tag{217.18}$$

where $0 = t_0 < t_1 < t_2 ....$ is a sequence of increasing discrete time levels with time steps $k_i = t_i - t_{i-1}$. To determine $U(t_i) \in \mathbb{R}^n$ satisfying (217.18) having already determined $U(t_{i-1})$, we have to solve the nonlinear system of equations

$$V = U(t_{i-1}) + k_i f(V) \tag{217.19}$$

in the unknown $V \in \mathbb{R}^n$. This equation is of the form $V = g(V)$ with $g(V) = U(t_{i-1}) + k_i f(V)$ and $g : \mathbb{R}^n \to \mathbb{R}^n$.

Therefore, we use the Fixed Point Iteration

$$V^{(m)} = U(t_{i-1}) + k_i f(V^{(m-1)}), \quad m = 1, 2, ...,$$

choosing say $V^{(0)} = U(t_{i-1})$ to try to solve for the new value. If L_f denotes the Lipschitz constant of $f : \mathbb{R}^n \to \mathbb{R}^n$, then

$$\|g(V) - g(W)\| = \|k_i(f(V) - f(W))\| \leq k_i L_f \|V - W\|, \quad V, W \in \mathbb{R}^n,$$

and thus $g : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with Lipschitz constant $L_g = k_i L_f$. Now $L_g < 1$ if the time step k_i satisfies $k_i < 1/L_f$ and thus the Fixed Point Iteration to determine $U(t_i)$ in (217.18) converges if $k_i < 1/L_f$. This gives a method for numerical solution of a very large class of initial value problems of the form $\dot{u}(t) = f(u(t))$ for $t > 0$, $u(0) = u_0$. The only restriction is to choose sufficiently small time steps, which however can be a severe restriction if the Lipschitz constant L_f is very large in the sense of requiring massive computational work (very small time steps). Thus, caution for large Lipschitz constants L_f!!

217.17 The Inverse Function Theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ is a given function and let $\bar{y} = f(\bar{x})$, where $\bar{x} \in \mathbb{R}^n$ is given. We shall prove that if $f'(\bar{x})$ is non-singular, then for $y \in \mathbb{R}^n$ close to $\bar{y}$, the equation

$$f(x) = y \tag{217.20}$$

has a unique solution x. Thus, we can define x as a function of y for y close to $\bar{y}$, which is called the inverse function $x = f^{-1}(y)$ of $y = f(x)$. To show that (217.20) has a unique solution x for any given y close to $\bar{y}$, we consider the Fixed Point iteration for $x = g(x)$ with $g(x) = x - (f'(\bar{x}))^{-1}(f(x) - y)$, which has the fixed point x satisfying $f(x) = y$ as desired. The iteration is

$$x^{(j)} = x^{(j-1)} - (f'(\bar{x}))^{-1}(f(x^{(j-1)}) - y), \quad j = 1, 2, ...,$$

with $x^{(0)} = \bar{x}$. To analyze the convergence, we subtract

$$x^{(j-1)} = x^{(j-2)} - (f'(\bar{x}))^{-1}(f(x^{(j-2)}) - y)$$

and write $e^j = x^{(j)} - x^{(j-1)}$ to get

$$e^j = e^{j-1} - (f'(\bar{x}))^{-1}(f(x^{(j-1)}) - f(x^{(j-2)}) \quad \text{for } j = 1, 2, ...$$

The Mean Value theorem implies

$$f_i(x^{(j-1)}) - f_i(x^{(j-2)}) = f'(z)e^{j-1},$$

where z lies on the straight line between $x^{(j-1)}$ and $x^{(j-2)}$. Note there might be possibly different z for different rows of $f'(z)$. We conclude that

$$e^j = \left(I - (f'(\bar{x}))^{-1} f'(z)\right) e^{j-1}.$$

Assuming now that

$$\|I - (f'(\bar{x}))^{-1} f'(z)\| \le \theta, \tag{217.21}$$

where $\theta < 1$ is a positive constant, we have

$$\|e^j\| \le \theta \|e^{j-1}\|.$$

As in the proof of the Contraction Mapping theorem, this shows that the sequence $\{x^{(j)}\}_{j=1}^{\infty}$ is a Cauchy sequence and thus converges to a vector $x \in \mathbb{R}^n$ satisfying $f(x) = y$.

The condition for convergence is obviously (217.21). This condition is satisfied if the coefficients of the Jacobian $f'(x)$ are Lipschitz continuous close to $\bar{x}$ and $f'(\bar{x})$ is non-singular so that $(f'(\bar{x}))^{-1}$ exists, and we restrict y to be sufficiently close to $\bar{y}$.

We summarize in the following (very famous):

Theorem 217.6 (Inverse Function theorem) *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ and assume the coefficients of $f'(x)$ are Lipschitz continuous close to $\bar{x}$ and $f'(\bar{x})$ is non-singular. Then for y sufficiently close to $\bar{y} = f(\bar{x})$, the equation $f(x) = y$ has a unique solution x. This defines x as a function $x = f^{-1}(y)$ of y.*

Carl Jacobi (1804-51), German mathematician, was the first to study the role of the determinant of the Jacobian in the inverse function theorem, and also gave important contributions to many areas of mathematics including the budding theory of first order partial differential equations .

217.18 The Implicit Function Theorem

There is an important generalization of the Inverse Function theorem. Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be a given function with value $f(x, y) \in \mathbb{R}^n$ for $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Assume that $f(\bar{x}, \bar{y}) = 0$ and consider the equation in $x \in \mathbb{R}^n$,

$$f(x, y) = 0,$$

for $y \in \mathbb{R}^m$ close to $\bar{y}$. In the case of the Inverse Function theorem we considered a special case of this situation with $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ defined by $f(x, y) = g(x) - y$ with $g : \mathbb{R}^n \to \mathbb{R}^n$.

We define the Jacobian $f'_x(x, y)$ of $f(x, y)$ with respect to x at (x, y) to be the $n \times n$ matrix with elements

$$\frac{\partial f_i}{\partial x_j}(x, y).$$

Assuming now that $f'_x(\bar{x},\bar{y})$ is non-singular, we consider the Fixed Point iteration:

$$x^{(j)} = x^{(j-1)} - (f'_x(\bar{x},\bar{y}))^{-1} f(x^{(j-1)},y),$$

connected to solving the equation $f(x,y) = 0$. Arguing as above, we can show this iteration generates a sequence $\{x^{(j)}\}_{j=1}\infty$ that converges to $x \in \mathbb{R}^n$ satisfying $f(x,y) = 0$ assuming $f'_x(x,y)$ is Lipschitz continuous for x close to $\bar{x}$ and y close to $\bar{y}$. This defines x as a function $g(y)$ of y for y close to $\bar{y}$. We have now proved the (also very famous):

Theorem 217.7 (Implicit Function theorem) *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ with $f(x,y) \in \mathbb{R}^n$ and $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, and assume that $f(\bar{x},\bar{y}) = 0$. Assume that the Jacobian $f'_x(x,y)$ with respect to x is Lipschitz continuous for x close to $\bar{x}$ and y close to $\bar{y}$, and that $f'_x(\bar{x},\bar{y})$ is non-singular. Then for y close to $\bar{y}$, the equation $f(x,y) = 0$ has a unique solution $x = g(y)$. This defines x as a function $g(y)$ of y.*

217.19 Newton's Method

We next turn to *Newton's method* for solving an equation $f(x) = 0$ with $f : \mathbb{R}^n \to \mathbb{R}^n$, which reads:

$$x^{(i+1)} = x^{(i)} - f'(x^{(i)})^{-1} f(x^{(i)}), \quad \text{for } i = 0,1,2,..., \tag{217.22}$$

where $x^{(0)}$ is an initial approximation. Newton's method corresponds to Fixed Point iteration for $x = g(x)$ with $g(x) = x - f'(x)^{-1}f(x)$. We shall prove that Newton's method converges quadratically close to a root $\bar{x}$ when $f'(\bar{x})$ is non-singular. The argument is the same is as in the case $n = 1$ considered above. Setting $e^i = \bar{x} - x^{(i)}$, and using $\bar{x} = \bar{x} - f'(x^{(i)})^{-1}f(\bar{x})$ if $f(\bar{x}) = 0$, we have

$$\begin{aligned}\bar{x} - x^{(i+1)} &= \bar{x} - x^{(i)} - f'(x^{(i)})^{-1}(f(\bar{x}) - f(x^{(i)}))\\ &= \bar{x} - x^{(i)} - f'(x^{(i)})^{-1}(f'(x^{(i)}) + E_f(x^{(i)},\bar{x})) = f'(x^{(i)})^{-1}E_f(x^{(i)},\bar{x}).\end{aligned}$$

We conclude that

$$\|\bar{x} - x^{(i+1)}\| \le C\|\bar{x} - x^{(i)}\|^2$$

provided

$$\|f'(x^{(i)})^{-1}\| \le C,$$

where C is some positive constant. We have proved the following fundamental result:

Theorem 217.8 (Newton's method) *If $\bar{x}$ is a root of $f : \mathbb{R}^n \to \mathbb{R}^n$ such that $f(x)$ is uniformly differentiable with a Lipschitz continuous derivative close to $\bar{x}$ and $f'(\bar{x})$ is non-singular, then Newton's method for solving $f(x) = 0$ converges quadratically if started sufficiently close to $\bar{x}$.*

In concrete implementations of Newton's method we may rewrite (217.22) as

$$f'(x^{(i)})z = -f(x^{(i)}),$$
$$x^{(i+1)} = x^{(i)} + z,$$

where $f'(x^{(i)})z = -f(x^{(i)})$ is a system of equations in z that is solved by Gaussian elimination or by some iterative method.

EXAMPLE 217.18. We return to the equation (217.19), that is,

$$h(V) = V - k_i f(V) - U(t_{i-1}) = 0.$$

To apply Newton's method to solve the equation $h(V) = 0$, we compute

$$h'(v) = I - k_i f'(v),$$

and conclude that $h'(v)$ will be non-singular at v, if $k_i < \|f'(v)\|^{-1}$. We conclude that Newton's method converges if k_i is sufficiently small and we start close to the root. Again the restriction on the time step is connected to the Lipschitz constant L_f of f, since L_f reflects the size of $\|f'(v)\|$.

217.20 Differentiation Under the Integral Sign

Finally, we show that if the limits of integration of an integral are independent of a variable x_1, then the operation of taking the partial derivative with respect x_1 can be moved past the integral sign. Let then $f : \mathbb{R}^2 \to \mathbb{R}$ be a function of two real variables x_1 and x_2 and consider the integral

$$\int_0^1 f(x_1, x_2)\, dx_2 = g(x_1),$$

which is a function $g(x_1)$ of x_1. We shall now prove that

$$\frac{dg}{dx_1}(\bar{x}_1) = \int_0^1 \frac{\partial f}{\partial x_1}(\bar{x}_1, x_2)\, dx_2, \tag{217.23}$$

which is referred to as "differentiation under the integral sign". The proof starts by writing

$$f(x_1, x_2) = f(\bar{x}_1, x_2) + \frac{\partial f}{\partial x_1}(\bar{x}_1, x_2)(x_1 - \bar{x}_1) + E_f(x_1, \bar{x}_1, x_2),$$

where we assume that

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

Taylor's theorem implies this is true provided the second partial derivatives of f are bounded. Integration with respect to x_2 yields

$$\int_0^1 f(x_1,x_2)\,dx_2 = \int_0^1 f(\bar{x}_1,x_2)\,dx_2 + (x_1-\bar{x}_1)\int_0^1 \frac{\partial f}{\partial x_1}(\bar{x}_1,x_2)\,dx_2 + \int_0^1 E_f(x_1,\bar{x}_1,x_2)\,dx_2.$$

Since

$$\left|\int_0^1 E_f(x_1,\bar{x}_1,x_2)\,dx_2\right| \le K_f(\bar{x}_1-x_1)^2$$

(217.23) follows after dividing by $(x_1-\bar{x}_1)$ and taking the limit as x_1 tends to $\bar{x}_1$. We summarize:

Theorem 217.9 (Differentiation under the integral sign) *If the second partial derivatives of $f(x_1,x_2)$ are bounded, then for $x_1\in\mathbb{R}$,*

$$\frac{d}{dx_1}\int_0^1 f(x_1,x_2)\,dx_2 = \int_0^1 \frac{\partial f}{\partial x_1}(x_1,x_2)\,dx_2 \tag{217.24}$$

Example 217.19.

$$\frac{d}{dx}\int_0^1 (1+xy^2)^{-1}\,dy = \int_0^1 \frac{\partial}{\partial x}(1+xy^2)^{-1}\,dy = -\int_0^1 \frac{y^2}{(1+xy^2)^2}\,dy.$$

Chapter 217 Problems

217.1. Sketch the following surfaces in $\mathbb{R}^3$: (a) $\Gamma=\{x : x_3=x_1^2+x_2^2\}$, (b) $\Gamma=\{x : x_3=x_1^2-x_2^2\}$, (c) $\Gamma=\{x : x_3=x_1+x_2^2\}$, (d) $\Gamma=\{x : x_3=x_1^4+x_2^6\}$. Determine the tangent planes to the surfaces at different points.

217.2. Determine whether the following functions are Lipschitz continuous or not on $\{x : |x|<1\}$ and determine Lipschitz constants:

- (a) $f:\mathbb{R}^3\to\mathbb{R}^3$ where $f(x)=x|x|^2$,
- (b) $f:\mathbb{R}^3\to\mathbb{R}$ where $f(x)=\sin|x|^2$,
- (c) $f:\mathbb{R}^2\to\mathbb{R}^3$ where $f(x)=(x_1,x_2,\sin|x|^2)$,
- (d) $f:\mathbb{R}^3\to\mathbb{R}$ where $f(x)=1/|x|$,
- (e) $f:\mathbb{R}^3\to\mathbb{R}^3$ where $f(x)=x\sin(|x|)$, (optional)
- (f) $f:\mathbb{R}^3\to\mathbb{R}$ where $f(x)=\sin(|x|)/|x|$. (optional)

217.3. For the functions in the previous exercise, determine which are contractions in $\{x : |x|<1\}$ and find their fixed points (optional).

217.4. Linearize the following functions on $\mathbb{R}^3$ at $x = (1, 2, 3)$:

- (a) $f(x) = |x|^2$,
- (b) $f(x) = \sin(|x|^2)$,
- (c) $f(x) = (|x|^2, \sin(x_2))$,
- (d) $f(x) = (|x|^2, \sin(x_2), x_1 x_2 \cos(x_3))$.

217.5. Compute the determinant of the Jacobian of the following functions: (a) $f(x) = (x_1^3 - 3x_1x_2^2, 3x_1x_2^2 - x_2^3)$, (b) $f(x) = (x_1e^{x_2}\cos(x_3), x_1e^{x_2}\sin(x_3), x_1e^{x_2})$.

217.6. Compute the second order Taylor polynomials at $(0, 0, 0)$ of the following functions $f : \mathbb{R}^3 \to \mathbb{R}$: (a) $f(x) = \sqrt{1 + x_1 + x_2 + x_3}$, (b) $f(x) = (x_1 - 1)x_2x_3$, (c) $f(x) = \sin(\cos(x_1x_2x_3))$, (d) $\exp(-x_1^2 - x_2^2 - x_3^2)$, (e) try to estimate the errors in the approximations in (a)-(d).

217.7. Linearize $f \circ s$, where $f(x) = x_1x_2x_3$ at $t = 1$ with (a) $s(t) = (t, t^2, t^3)$, (b) $s(t) = (\cos(t), \sin(t), t)$, (c) $s(t) = (t, 1, t^{-1})$.

217.8. Evaluate $\int_0^\infty y^n e^{-xy}\, dy$ for $x > 0$ by repeated differentiation with respect to x of $\int_0^\infty e^{-xy}\, dy$.

217.9. Try to minimize the function $u(x) = x_1^2 + x_2^2 + 2x_3^2$ by starting at $x = (1, 1, 1)$ using the method of steepest descent. Seek the largest step length for which the iteration converges.

217.10. Compute the roots of the equation $(x_1^2 - x_2^2 - 3x_1 + x_2 + 4, 2x_1x_2 - 3x_2 - x_1 + 3) = (0, 0)$ using Newton's method.

217.11. Generalize Taylor's theorem for a function $f : \mathbb{R}^n \to \mathbb{R}$ to third order.

217.12. Is the function $f(x_1, x_2) = \frac{x_1^2 - x_2^2}{x_1^2 + x_2^2}$ Lipschitz continuous close to $(0, 0)$?

> Jacobi and Euler were kindred spirits in the way they created their mathematics. Both were prolific writers and even more prolific calculators; both drew a great deal of insight from immense algorithmical work; both laboured in many fields of mathematics (Euler, in this respect, greatly surpassed Jacobi) ; and both at any moment could draw from the vast armoury of mathematical methods just those weapons which would promise the best results in the attack of a given problem. (Sciba)

218

Level Curves/Surfaces and the Gradient

> It would make no sense to overload the student with all kinds of little things that might be of occasional use. Instead, it is important that students become familiar with ways to think mathematically, recognize the need for applying mathematical methods to engineering problems, realize that mathematics is a systematic science built on relatively few principles and get a firm grasp for the interrelation between theory, computing and experiment. (E. Kreyszig, in Preface to Advanced Engineering Mathematics, 1993)

218.1 Level Curves

A *level curve* of a function $u : \mathbb{R}^2 \to \mathbb{R}$ is a curve $g : [a, b] \to \mathbb{R}^2$ such that

$$u(g(t)) = c \quad \text{for } t \in [a, b], \tag{218.1}$$

where c is a constant. A level curve is also called an *isoline*. The points x on a level curve $x = g(t)$ satisfying (218.1), all have the same function value $u(x) = u(g(t)) = c$. By plotting the level curves or isolines for a collection of different constants c, we get a *level curve plot* or *contour plot* of the function $u(x)$. The level curves are the projections onto $\mathbb{R}^2$ of the intersections of the planes $x_3 = c$ in $\mathbb{R}^3$ with the graph $\{x \in \mathbb{R}^2 : x_3 = u(x_1, x_2),\ (x_1, x_2) \in \mathbb{R}^2\}$. We illustrate in Fig. 218.2.

EXAMPLE 218.1. The level curves of the function $u(x) = x_1^2 + x_2^2$ are the circles $x_1^2 + x_2^2 = c$ with $c \geq 0$ a constant. The level curves of the

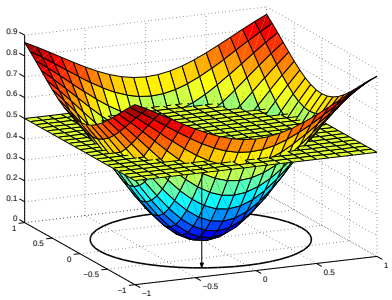


FIGURE 218.1. Projection onto $\mathbb{R}^2$ of the intersection of $x_3 = c$ and $x_3 = u(x_1, x_2$ (with $u(x_1, x_2) = 1 - \exp(-x_1^2 - x_2^2)$ and $c = .5$) gives a level curve.

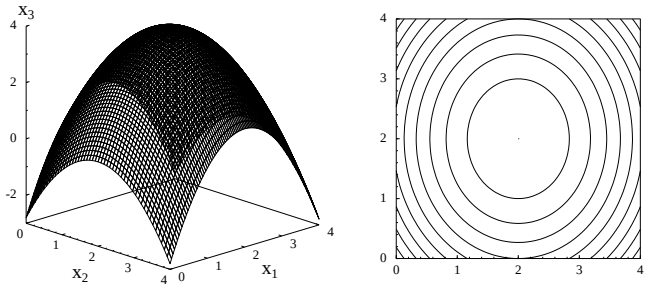


FIGURE 218.2. A plot of a surface and the corresponding contour plot with contour curves shown every .7 units starting at the maximum height of 4.

function $u(x) = 2x_1^2 + x_2^2$ are the ellipses $2x_1^2 + x_2^2 = c$ with $c \geq 0$. The level curves of the function $u(x) = x_1^2 - x_2$ are the parabolas $x_2 = x_1^2 - c$ with c a constant.

EXAMPLE 218.2. A hiking map indicates the level curves of the function $u : \mathbb{R}^2 \to \mathbb{R}$ that gives the height of a point $x \in \mathbb{R}^2$ above a reference level, like the see level. The difference in height between two nearby level curves is typically 10 meters. The change in height between two points can be obtained by counting the number of contour lines intersected by a line joining the two points. This is useful when planning a hiking trip. Recall Fig. 217.8

A level curve $u(g(t)) = c$ may be thought of as the shore-lines with the sea level equal to c above the reference level.

218.2 Local Existence of Level Curves

The local existence of level curves follows from the following special case of the Implicit Function theorem, where the level curve is given by $t \to (t, g(t))$ or $t \to (g(t), t)$, where $g : \mathbb{R} \to \mathbb{R}$.

Theorem 218.1 *Assume $u : \mathbb{R}^2 \to \mathbb{R}$ has continuous partial derivatives and $u(\bar{x}_1, \bar{x}_2) = c$. If $\frac{\partial u}{\partial x_2}(\bar{x}_1, \bar{x}_2) \neq 0$, then there is a $\delta > 0$ such that $u(x_1, x_2) = c$ has a unique solution $x_2 = g(x_1)$ for $|x_1 - \bar{x}_1| < \delta$. If $\frac{\partial u}{\partial x_1}(\bar{x}_1, \bar{x}_2) \neq 0$, then there is a $\delta > 0$ such that $u(x_1, x_2) = c$ has a unique solution $x_1 = g(x_2)$ for $|x_2 - \bar{x}_2| < \delta$.*

Notice that if $\frac{\partial u}{\partial x_2}(\bar{x}_1, \bar{x}_2) = 0$, then the level curve is parallel to the x_2-axis, and thus we cannot expect the equation $u(x_1, x_2) = c$ to define x_2 as a function of x_1 (a corresponding function $x_2 = g(x_1)$ would then have infinite slope at $x_1 = \bar{x}_1$).

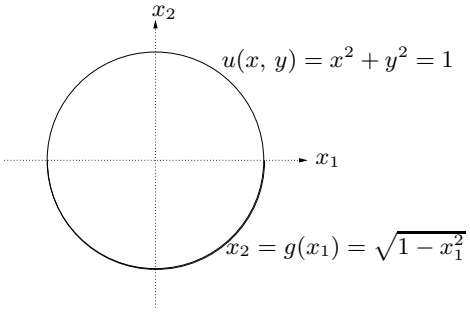


FIGURE 218.3. $x_2 = -\sqrt{1 - x_1^2}$ giving one piece of the level curve $u(x_1, x_2) = x_1^2 + x_2^2 = 1$.

218.3 Level Curves and the Gradient

Differentiating both sides of (218.1), we get using the Chain rule

$$\frac{d}{dt}u(g(t)) = \nabla u(x) \cdot g'(t) = \frac{\partial u}{\partial x_1}(g(t))g_1'(t) + \frac{\partial u}{\partial x_2}(g(t))g_2'(t) = 0.$$

Since $g'(t) = (g_1'(t), g_2'(t))$ is the direction of the tangent of the curve $g(t)$, this means that the direction $g'(t)$ of a level curve of a function $u : \mathbb{R}^2 \to \mathbb{R}$ is orthogonal to the gradient $\nabla u(g(t))$. Recall that the gradient $\nabla u(x)$ points in the direction of the steepest ascent of the function $u(x)$ at x, and the direction perpendicular to the gradient (the direction of the level curve) is a direction in which u stays constant, see Fig. 218.4. Moving along

a level curve the function stays constant, and moving in the direction of the gradient the function increases as quickly as possible!

Since the gradient $\nabla(\bar{x})$ is a normal to the tangent to the level curve through $\bar{x}$, we can write the equation for the tangent to the level curve through $\bar{x}$ in the form $\nabla u(\bar{x}) \cdot (x - \bar{x}) = 0$.

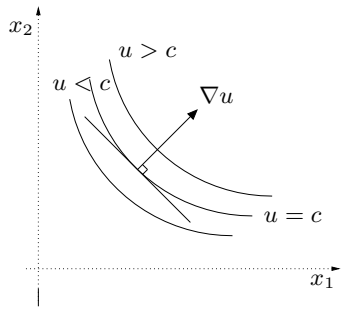


FIGURE 218.4. The gradient of $\nabla u(x)$ of a function $u : \mathbb{R}^2 \to \mathbb{R}$ is orthogonal to the level curve of u through x.

We summarize:

Theorem 218.2 *The gradient $\nabla u(g(t))$ of a function $u : \mathbb{R}^2 \to \mathbb{R}$ is orthogonal to the tangent $g'(t)$ of a level curve $g : I \to \mathbb{R}$. We can write the equation for the tangent to the level curve through $\bar{x}$ in the form $\nabla u(\bar{x}) \cdot (x - \bar{x}) = 0$.*

EXAMPLE 218.3. Consider the function $u(x_1, x_2) = x_1^2 + x_2^2$ with circular level curves $g(t) = (g_1(t), g_2(t))$ satisfying $g_1^2(t) + g_2^2(t) = c^2$. We have $\nabla u(x) = (2x_1, 2x_2)$ and differentiating $g_1^2(t) + g_2^2(t) = c^2$ with respect to t we get $0 = 2g_1(t)g_1'(t) + 2g_2(t)g_2'(t) = \nabla u(g(t)) \cdot g'(t)$ as expected. Alternatively, parameterizing a level curve $g(t)$ satisfying $g_1^2(t) + g_2^2(t) = c^2$ by $g(t) = c(\cos(t), \sin(t))$, we have $g'(t) = c(-\sin(t), \cos(t)) = (-x_2(t), x_1(t))$ with $x = g(t)$. We check that $\nabla u(g(t)) \cdot g'(t) = 2(x_1(t), x_2(t)) \times (-x_2(t), x_1(t)) = 0$.

EXAMPLE 218.4. If $u : \mathbb{R}^2 \to \mathbb{R}$ is of the form $u(x_1, x_2) = f(x_1) - x_2$, where $f : \mathbb{R} \to \mathbb{R}$, then $\nabla u(x) = (f'(x_1), -1)$. A level curve $u(g(t)) = c$ can be parameterized by $g(t) = (t, f(t) - c)$, and $g'(t) = (1, f'(t))$. Clearly, $\nabla u(g(t)) \cdot g'(t) = (f'(t), -1) \cdot (1, f'(t)) = 0$.

218.4 Level Surfaces

A *level surface* of a function $u : \mathbb{R}^3 \to \mathbb{R}$ is a surface $g : Q \to \mathbb{R}^3$, where Q is a subset of $\mathbb{R}^2$, such that

$$u(g(y)) = c \quad \text{for } y \in Q, \tag{218.2}$$

where c is a constant. A level surface is also called an *isosurface*. The points on a level surface $g(t)$ satisfying (218.1) all have the same function value $u(g(y)) = c$.

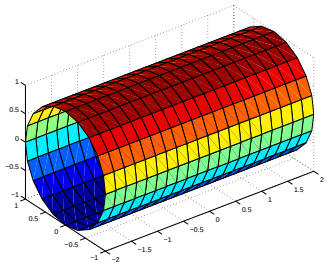


FIGURE 218.5. A piece of the level surface $u(x_1, x_2, x_3) = x_1^2 + x_3^2 = 1$.

218.5 Local Existence of Level Surfaces

The local existence of level surfaces follows from the following special case of the Implicit Function theorem. We find that the level surface is parameterized as $g(y_1, y_2) = (y_1, y_2, f(y_1, y_2))$, $g(y_1, y_3) = (y_1, f(y_1, y_2), y_3)$ or $g(y_2, y_3) = (f(y_2, y_3), y_2, y_3)$ with some function $f : \mathbb{R}^2 \to \mathbb{R}$, depending on which partial derivative is non-zero.

Theorem 218.3 *Assume $u : \mathbb{R}^3 \to \mathbb{R}$ has continuous partial derivatives and $u(\bar{x}_1, \bar{x}_2, \bar{x}_3) = c$, where c is a constant. If $\partial u/\partial x_3 \neq 0$, then there is a $\delta > 0$ such that $u(x_1, x_2, x_3) = c$ has a unique solution $x_3 = f(x_1, x_2)$ for $\|(x_1, x_2) - (\bar{x}_1, \bar{x}_2)\| < \delta$. If $\partial u/\partial x_2 \neq 0$, then there is a $\delta > 0$ such that $u(x_1, x_2, x_3) = c$ has a unique solution $x_2 = g(x_1, x_3)$ for $\|(x_1, x_3) - (\bar{x}_1, \bar{x}_3)\| < \delta$. If $\partial u/\partial x_1 \neq 0$, then there is a $\delta > 0$ such that $u(x_1, x_2, x_3) = c$ has a unique solution $x_1 = g(x_2, x_3)$ for $\|(x_2, x_3) - (\bar{x}_2, \bar{x}_3)\| < \delta$.*

218.6 Level Surfaces and the Gradient

Differentiating both sides of (218.2) with respect to y_1 and y_2, where $y = (y_1, y_2)$, we get using the Chain rule

$$\frac{\partial}{\partial y_i} u(g(y)) = \nabla u(g(y)) \cdot g'_{,i}(y) = 0, \quad i = 1, 2,$$

where we use the notation

$$g'_{,i}(y) = \frac{\partial}{\partial y_i} g(y).$$

We use the comma in $g'_{,i}$ to indicate differentiation with respect to x_i, while g_i will denote component i of $g = (g_1, g_2, g_3)$. We recall that the tangent plane (linearization) of $g(y)$ at $\bar{x} = g(\bar{y})$ is given by $(y_1, y_2) \to g(\bar{y}) + (y_1 - \bar{y}_1)g'_{,1}(\bar{y}) + (y_2 - \bar{y}_2)g'_{,2}(\bar{y})$, and we conclude that $\nabla u(g(\bar{y}))$ is orthogonal to the tangent plane of the level surface through $\bar{x} = g(\bar{y})$. We say that $\nabla u(g(\bar{y}))$ is *orthogonal to the level surface* $u(x) = c$ through $\bar{x} = g(\bar{y})$, or that $\nabla u(g(\bar{y}))$ is a *normal to the level surface* $u(x) = c$ at $\bar{x} = g(\bar{y})$, see Fig. 218.6. Since $\nabla u(\bar{x})$ thus is a normal to the tangent plane at x, the equation for the tangent plane to a level surface through $\bar{x}$ can also be written $\nabla u(\bar{x}) \cdot (x - \bar{x}) = 0$.

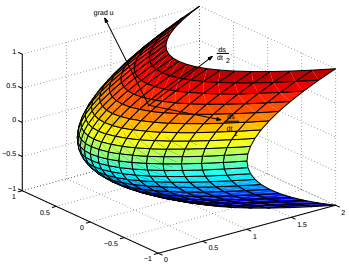


FIGURE 218.6. The gradient $\nabla u(x) = (2x_1, -1, 2x_3)$ of $u(x_1, x_2, x_3) = x_1^2 + x_3^2 - x_2$ is orthogonal to a level surface $(x_1, x_3) \to g(x_1, x_3) = (x_1, x_1^2 + x_3^2 + c, x_3)$ since $g'_1 = (1, 2x_1, 0)$ and $g'_3 = (0, 2x_3, 1)$.

We summarize:

Theorem 218.4 *The gradient $\nabla u(\bar{x})$ of a function $u : \mathbb{R}^3 \to \mathbb{R}$, is orthogonal to the tangent plane $(y_1, y_2) \to g(\bar{y}) + (y_1 - \bar{y}_1)g'_{,1}(\bar{y}) + (y_2 - \bar{y}_2)g'_{,2}(\bar{y})$ of a level surface $y \to x = g(y)$, where $\bar{x} = g(\bar{y})$. The equation for the tangent plane of a level surface through $\bar{x}$ can also be written $\nabla u(\bar{x}) \cdot (x - \bar{x}) = 0$.*

EXAMPLE 218.5. Consider the function $u(x) = x_1^2 + x_2^2 + x_3^2$ with the level surfaces $g(y)$ satisfying $g_1^2(y) + g_2^2(y) + s_3^2(y) = c^2$ representing

spheres centered at the origin with radii c. The gradient $\nabla(x) = 2x$ is evidently orthogonal to a tangent plane of a level surface at x.

EXAMPLE 218.6. If $u : \mathbb{R}^3 \to \mathbb{R}$ is of the form $u(x_1, x_2, x_3) = f(x_1, x_2) - x_3$, where $f : \mathbb{R} \to \mathbb{R}$, then $\nabla u(x) = (f'_{,1}(x_1, x_2), f'_{,2}(x_1, x_2), -1)$. A level surface $u(g(y)) = c$ can be parameterized by $g(y) = (y_1, y_2, f(y_1, y_2) - c)$, and $g'_1(y) = (1, 0, f'_{,1}(y))$ and $g'_2(y) = (0, 1, f'_{,2}(y))$. Clearly, $\nabla u(g(y)) \cdot g'_{,i}(y) = 0$ for $i = 1, 2$.

Chapter 218 Problems

218.1. Sketch the following surfaces in $\mathbb{R}^3$: (a) $\Gamma = \{x : x_1^2 + x_2^2 = x_3\}$, (b) $\Gamma = \{x : x_1^2 + x_2^2 = x_3^2\}$, (c) $\Gamma = \{x : x_1^2 + x_2^2 = -x_3^2\}$, (d) $\Gamma = \{x : x_1^2 + 2x_2^2 + 3x_3^2 = 6\}$. Determine the tangent planes to the surfaces at various points.

218.2. Find parametrization of the curves for the intersections of the surfaces in the previous exercise with the plane $x_3 = 1$.

218.3. Show that the surface $\Gamma = \{x : x_1^2 + 2x_2^2 + 3x_3^2 + x_1x_3^3 = 7\}$ can be expressed in the form $x_3 = g(x_1, x_2)$ close to $(1, 1, 1)$.

218.4. Compute the gradients of the following functions $f : \mathbb{R}^3 \to \mathbb{R}$: (a) $f(x) = x_1^n(x_2^n + x_3^n)$, (b) $f(x) = |x|$, (c) $f(x) = |x|^2$, (d) $f(x) = 1/|x|$, (e) $f(x) = \exp(x_1x_2x_3)$.

218.5. For each of the functions in the previous exercise, determine the equation for the tangent plane to the level surface $f(x) = f(1, 1, 1)$ at $x = (1, 1, 1)$.

218.6. Determine the equation for the tangent plane at $x = (1, 2, 3)$ for the following surfaces: (a) $x_3 = \frac{3}{2}x_1x_2$, (b) $x_1^2 + x_2^2 + x_3^2 = 14$, (c) $x_2 = \sin(2\pi x_1) + 2\cos(2\pi x_3)$.

218.7. Determine the tangent plane and normal vector to the ellipse $x_1^2 + 3x_2^2 = 10$ at $x = (1, \sqrt{3})$.

218.8. Let $f : Q \to \mathbb{R}$, where $Q = [0, 1] \times [0, 1]$ is the unit square, satisfy $f(x) = 0$ for x on the boundary of Q. Prove under convenient assumptions that there is a point $y \in Q$ such that $\nabla f(y) = 0$.

219

Linearization and Stability of Initial Value Problems

The logos of somewome to that base anything, when most characteristically mantissa minus, comes to nullum in the endth: orso, here is nowet badder than the sin of Aha with his cosin Lil, verswaysed on coversvised, and all that's consecants and cotangincies... (Finnegans Wake, James Joyce)

219.1 Introduction

We continue the study of the general initial value problem (212.1), now focussing on the *stability* of solutions, which is a measure of the *sensitivity of solutions to perturbations in given data*. This is a fundamentally important aspect of the behavior of solutions, which we touched upon in Chapter *The general initial value problem*, and which we now consider more closely.

We consider an autonomous problem of the form

$$\dot{u}(t) = f(u(t)) \quad \text{for } 0 < t \leq T,\ u(0) = u^0, \tag{219.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$ is a given bounded Lipschitz continuous function, $u^0 \in \mathbb{R}^d$ is a given initial value, and we seek a solution $u : [0,T] \to \mathbb{R}^d$, where we think of $[0,T]$ as a given time interval. To study the stability of a given solution $u(t)$ to small perturbations in given data, e.g. in the given initial data u^0, we will consider an associated *linearized problem* that arises upon linearizing the function $v \to f(v)$ around the solution $u(t)$.

219.2 Stationary Solutions

We consider first the simplest case of a *stationary solution* $u(t) = \bar{u}$ for $0 \leq t \leq T$, that is a solution $u(t)$ of (219.1) that is independent of time t. Since $\dot{u}(t) = 0$ if $u(t)$ is independent of time, $u(t) = \bar{u}$ is a stationary solution if $f(\bar{u}) = 0$ and $u^0 = \bar{u}$, where $\bar{u} = (\bar{u}_1, ..., \bar{u}_d) \in \mathbb{R}^d$. The equation $f(\bar{u}) = 0$ corresponds to a system of d equations $f_i(\bar{u}_1, ..., \bar{u}_d) = 0$, $i = 1, ..., d$, in d unknowns $\bar{u}_1, ..., \bar{u}_d$, where the f_i are the components of f. We studied such systems of equations in Chapter *Vector-valued functions of several real variables*. Here, we assume the existence of a stationary solution $u(t) = \bar{u}$ so that $\bar{u} \in \mathbb{R}^d$ satisfies the equation $f(\bar{u}) = 0$. In general, there may be several roots $\bar{u}$ of the equation $f(v) = 0$ and thus there may be several stationary solutions. We also refer to a stationary solution $u(t) = \bar{u}$ as an *equilibrium solution*.

EXAMPLE 219.1. The stationary solutions $\bar{u}$ of the Crash model

$$\begin{cases} \dot{u}_1 + \nu u_1 - \kappa u_1 u_2 = \nu & t > 0, \\ \dot{u}_2 + 2\nu u_2 - \nu u_2 u_1 = 0 & t > 0, \end{cases} \tag{219.2}$$

of the form $\dot{u} = f(u)$ with $f(u) = (-\nu u_1 + \kappa u_1 u_2 + \nu, -2\nu u_2 + \nu u_2 u_1)$, are $\bar{u} = (1, 0)$ and $\bar{u} = (2, \frac{\nu}{\kappa})$.

219.3 Linearization at a Stationary Solution

We shall now study perturbations of a given stationary solution under small perturbations of initial data. We thus assume $f(\bar{u}) = 0$ and denote the corresponding equilibrium solution by $\bar{u}(t)$ for $t > 0$, that is $\bar{u}(t) = \bar{u}$ for $t > 0$. We consider the initial value problem (219.1) with $u^0 = \bar{u} + \varphi^0$, where $\varphi^0 \in \mathbb{R}^d$ is a given small perturbation of the initial data $\bar{u}$. We denote the corresponding solution by $u(t)$ and focus attention on the corresponding perturbation in the solution, that is $\psi(t) = u(t) - \bar{u}(t) = u(t) - \bar{u}$. We want to derive a differential equation for the perturbation $\psi(t)$, and to this end we linearize f at $\bar{u}$ and write

$$f(u(t)) = f(\bar{u} + \psi(t)) = f(\bar{u}) + f'(\bar{u})\psi(t) + e(t),$$

where $f'(\bar{u})$ is the Jacobian of $f : \mathbb{R}^d \to \mathbb{R}^d$ at $\bar{u}$ and the error term $e(t)$ is quadratic in $\psi(t)$ (and thus is very small if $\psi(t)$ is small). Since $f(\bar{u}) = 0$ and $u(t)$ satisfies (219.1), we have

$$\dot{\psi}(t) = \frac{d}{dt}(\bar{u} + \psi(t))) = f(u(t)) = f'(\bar{u})\psi(t) + e(t).$$

Neglecting the quadratic term $e(t)$, we are led to a linear initial value problem,

$$\dot{\varphi}(t) = f'(\bar{u})\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0, \tag{219.3}$$

where $\varphi(t)$ is an approximation of the perturbation $\psi(t) = u(t) - \bar{u}$ up to a second order term. We refer to (219.3) as the *linearized problem* associated to the stationary solution $\bar{u}$ of (219.1). Since $f'(\bar{u})$ is a constant $d \times d$ matrix, we can express the solution to (219.3) using the matrix exponential as

$$\varphi(t) = \exp(tA)\varphi^0 \quad \text{for } 0 < t \leq T, \tag{219.4}$$

where $A = f'(\bar{u})$. We thus have a formula that describes the evolution of perturbation $\varphi(t)$ starting from an initial perturbation $\varphi(0) = \varphi^0$. Depending on the nature of the matrix $\exp(tA)$, the perturbation may increase or decrease with time, reflecting a stronger or lesser sensitivity of the solution $u(t)$ to perturbations in initial data and therefore different stability features of the given problem.

We know that if A is diagonalizable, so that $A = B\Lambda B^{-1}$ where B is a non-singular $d \times d$ matrix and Λ is a diagonal matrix with the eigenvalues $\lambda_1, ..., \lambda_d$ of A on the diagonal, then

$$\varphi(t) = B\exp(t\Lambda)B^{-1}\varphi^0 \quad \text{for } t \geq 0. \tag{219.5}$$

We see that each component of $\varphi(t)$ is a linear combination of $\exp(t\lambda_1)$,..., $\exp(t\lambda_d)$ and the sign of the real part $\text{Re}\,\lambda_i$ of λ_i determines if the corresponding term grows or decays exponentially. If some $\text{Re}\,\lambda_i > 0$, then we have exponential growth of certain perturbations, which indicates that the corresponding stationary solution $\bar{u}$ is *unstable*. On the other hand, if all $\text{Re}\,\lambda_i \leq 0$, then we would expect $\bar{u}$ to be *stable*.

These considerations are qualitative in nature, and to be more precise we should base judgements of stability or instability on quantitative estimates of perturbation growth. In the diagonalizable case, (219.5) implies in the Euclidean vector and matrix norms that

$$\|\varphi(t)\| \leq \|B\|\|B^{-1}\| \max_{i=1,\ldots,d} \exp(t\lambda_i)\|\varphi^0\|. \tag{219.6}$$

We see that the maximal perturbation growth is governed by the maximal exponential factors $\exp(t\lambda_i)$ as well as the factors $\|B\|$ and $\|B^{-1}\|$ related to the transformation matrix B. If the transformation matrix B is orthogonal, then $\|B\| = \|B^{-1}\| = 1$, and the perturbation growth is governed solely by the exponential factors $\exp(t\lambda_i)$. We give this case special attention:

219.4 Stability Analysis when $f'(\bar{u})$ Is Symmetric

If $A = f'(\bar{u})$ is symmetric so that $A = Q\Lambda Q^{-1}$ with Q orthogonal and Λ a diagonal matrix with real diagonal elements λ_i, then

$$\|\varphi(t)\| \leq \max_{i=1,\ldots,d} \exp(t\lambda_i)\|\varphi^0\|. \tag{219.7}$$

In particular, if all eigenvalues $\lambda_i \leq 0$ then perturbations $\varphi(t)$ cannot grow with time, and we say that the solution $\bar{u}$ is stable. On the other hand, if some eigenvalue $\lambda_i > 0$ and the corresponding eigenvector is g_i then $\varphi(t) = \exp(t\lambda_i)g_i$ solves the linearized initial value problem (219.3) with $\varphi^0 = g_i$, and evidently the particular perturbation $\varphi(t)$ grows exponentially. We then say that the solution $\bar{u}$ is *unstable*. Of course, the size of the positive eigenvalues influence the perturbation growth, so that if $\lambda_i > 0$ is small, then then growth is slow and the instability is mild. Likewise, if λ_i is small negative, then the exponential decay is slow.

219.5 Stability Factors

We may express the stability features of a particular perturbation φ^0 through a *stability factor* $S(T, \varphi_0)$ defined as follows:

$$S(T, \varphi_0) = \max_{0\leq t\leq T} \frac{\|\varphi(t)\|}{\|\varphi^0\|}.$$

where $\varphi(t)$ solves the linearized problem (219.3) with initial data φ^0. The stability factor $S(T, \varphi_0)$ measures the maximal growth of the norm of $\varphi(t)$ over the time interval $[0, T]$ versus the norm of the initial value φ_0.

We can now seek to capture the overall stability features of a stationary solution $\bar{u}$ by maximization over all different perturbations:

$$S(T) = \max_{\varphi^0\neq 0} S(T, \varphi_0).$$

If the stability factor $S(T)$ is large, then some perturbations grow very much over the time interval $[0, T]$, which indicates a strong sensitivity to perturbations or *instability*. On the other hand, if $S(T)$ is of moderate size then the perturbation growth is moderate, which signifies *stability*. Using the Euclidean matrix norm, we can also express $S(T)$ as

$$S(T) = \max_{0\leq t\leq T} \|\exp(tA)\|.$$

EXAMPLE 219.2. If $A = f'(\bar{u})$ is symmetric with eigenvalues $\lambda_1, ..., \lambda_d$, then

$$S(T) = \max_{i=1,...,d} \max_{0\leq t\leq T} \exp(t\lambda_i).$$

In particular, if all $\lambda_i \leq 0$, then $S(T) = 1$.

EXAMPLE 219.3.

The initial value problem for a pendulum takes the form

$$\begin{aligned} \dot{u}_1 = u_2, \quad \dot{u}_2 = -\sin(u_1) \quad \text{for } t > 0, \\ u_1(0) = u_{01},\ u_2(0) = u_{02}, \end{aligned}$$

corresponding to $f(u) = (u_2, -\sin(u_1)$ and the equilibrium solutions are $\bar{u} = (0,0)$ and $\bar{u} = (\pi, 0)$. We have

$$f'(\bar{u}) = \begin{pmatrix} 0 & 1 \\ -\cos(\bar{u}_1) & 0 \end{pmatrix},$$

and the linearized problem at $\bar{u} = (0,0)$ thus takes the form

$$\dot{\varphi}(t) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \varphi(t) \equiv A_0\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0,$$

with solution

$$\varphi_1(t) = \varphi_1^0 \cos(t) + \varphi_2^0 \sin(t), \quad \varphi_2(t) = -\varphi_1^0 \sin(t) + \varphi_2^0 \cos(t).$$

It follows by a direct computation (or using that $\begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}$ is an orthogonal matrix), that for $t > 0$

$$\|\varphi(t)\|^2 = \|\varphi_0\|^2,$$

and thus the norm $\|\varphi(t)\|$ of a solution $\varphi(t)$ of the linearized equations is constant in time, which means that the stability factor $S(T) = 1$ for all $T > 0$. We conclude that if the norm of a perturbation is small initially, it will stay small for all time. This means that the equilibrium solution $\bar{u} = (0,0)$ is *stable*. More precisely, if the pendulum is perturbed initially a little from its bottom position, the pendulum will oscillate back and forth around the bottom position with constant amplitude. This fits our direct experimental experience of course.

Note that the linearized operator A_0 is non-symmetric; the eigenvalues of A_0 are purely imaginary $\pm i$, which says that $\|\varphi(t)\| = \|\varphi_0\|$, that is a perturbation neither grows nor decays. Another way to derive this fact is to use the fact that A_0 is *antisymmetric*, that is $A_0^\top = -A_0$, which shows that $(A_0\varphi, \varphi) = (\varphi, A_0^\top \varphi) = -(\varphi, A_0\varphi) = -(A_0\varphi, \varphi)$, and thus $(A_0\varphi, \varphi) = 0$, where $(\cdot,\cdot)$ is the $\mathbb{R}^2$ scalar product. It follows from the equation $\dot{\varphi} = A_0\varphi$ upon multiplication by φ that $0 = (\dot{\varphi}, \varphi) = \frac{1}{2}\frac{d}{dt}(\varphi, \varphi) = \frac{1}{2}\frac{d}{dt}\|\varphi\|^2$, which proves that $\|\varphi(t)\|^2 = \|\varphi_0\|^2$.

The linearized problem at $\bar{u} = (\pi, 0)$ reads

$$\dot{\varphi}(t) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \varphi(t) \equiv A_\pi\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0,$$

with symmetric matrix A_π with eigenvalues ± 1. Since one eigenvalue is positive, the stationary solution $\bar{u} = (\pi, 0)$ is unstable. More precisely, the solution is given by

$$\varphi_1 = \frac{\varphi_1^0}{2}(e^t + e^{-t}) + \frac{\varphi_2^0}{2}(e^t - e^{-t}), \quad \varphi_2 = \frac{\varphi_1^0}{2}(e^t - e^{-t}) + \frac{\varphi_2^0}{2}(e^t + e^{-t}),$$

and due to the exponential factor e^t, perturbations will grow exponentially in time, and thus an initially small perturbation will become large as soon as $t \geq 10$ say. Physically, this means that if the pendulum is perturbed initially a little from its top position, the pendulum will eventually move away from the top position, even if the initial perturbation is very small. This fact of course has direct experimental evidence: to balance a pendulum with the weight in the top position is tricky business. Small perturbations quickly grow to large perturbations and the equilibrium solution $(\pi, 0)$ of the pendulum is unstable.

EXAMPLE 219.4. The linearization of the Crash model (219.2) at the equilibrium solution $\bar{u} = (1, 0)$, takes the form

$$\dot{\varphi}(t) = \begin{pmatrix} -\nu & \kappa \\ 0 & -\nu \end{pmatrix} \varphi(t) \equiv A_{\nu,\kappa}\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0, \quad (219.8)$$

The solution is given by $\varphi_2(t) = \varphi_2^0 \exp(-\nu t)$, and $\varphi_1(t) = t\kappa \exp(-\nu t)\varphi_2^0 + \exp(-\nu t)\varphi_1^0$. Clearly, $\varphi_2(t)$ decays monotonically to zero and so does $\varphi_1(t)$ if $\kappa = 0$. But, if $\kappa \neq 0$ then $\varphi_1(t)$ reaches the following value, assuming for simplicity that $\varphi_{01} = 0$,

$$\varphi_1(\nu^{-1}) = \nu^{-1}\kappa \exp(-1)\varphi_2^0,$$

which contains the factor ν^{-1} that is large if ν is small. In other words, the stability factor $S(\nu^{-1}) \sim \nu^{-1}$, which is large if ν is small. Eventually, however, $\varphi_1(t)$ decays to zero. As a result, the equilibrium solution $(1, 0)$ is stable only to small perturbations, since we saw in the Chapter The Crash model that $(1, 0)$ is unstable to perturbations above a certain threshold depending on λ. Note that here the Jacobian $f'(\bar{u}) = A_{\nu,\kappa}$ has a double eigenvalue $-\nu$, but $A_{\nu,\kappa}$ is non-symmetric and the space of eigenvectors is one-dimensional and is spanned by $(1, 0)$. As a result, the term $t\kappa \exp(-\nu t)\varphi_2^0$ with linear growth in t appears; thus in this highly non-symmetric problem (if ν is small), large perturbation growth $\sim \nu^{-1}$ is possible although all eigenvalues are non-positive.

The matrix $A_{\nu,\kappa}$ is an example of a *non-normal* matrix. A non-normal matrix A is a matrix such that $A^\top A \neq AA^\top$. A non-normal matrix may or may not be diagonalizable, and if diagonalizable so that $A = B\Lambda B^{-1}$, we may have $\|B\|$ or $\|B^{-1}$ large, resulting in large stability factors in the corresponding linearized problem, as we just saw (cf. Problem 219.5).

The linearization at the equilibrium solution $\bar{u} = (2, \frac{\nu}{\kappa})$ takes the form

$$\dot{\varphi}(t) = \begin{pmatrix} 0 & 2\kappa \\ \frac{\nu^2}{\kappa} & 0 \end{pmatrix} \varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0. \quad (219.9)$$

The eigenvalues of the Jacobian are $\pm\sqrt{2}\nu$ and the solution is a linear combination of $\exp(\sqrt{2}\nu t)$ and $\exp(-\sqrt{2}\nu t)$ and thus has one exponentially growing part with growth factor $\exp(\sqrt{2}\nu t)$. The equilibrium solution $u = (2, \frac{\nu}{\kappa})$ is thus unstable.

219.6 Stability of Time-Dependent Solutions

We now seek to extend the scope to linearization and linearized stability for a time-dependent solution $\bar{u}(t)$ of (219.1). We want to study solutions of the form $u(t) = \bar{u}(t) + \psi(t)$, where $\psi(t)$ is a perturbation. Using $\frac{d}{dt}\bar{u} = f(\bar{u})$ and linearizing f at $\bar{u}(t)$, we obtain

$$\frac{d}{dt}(\bar{u} + \psi)(t) = f(\bar{u}(t)) + f'(\bar{u}(t))\psi(t) + e(t),$$

with $e(t)$ quadratic in $\psi(t)$. This leads to the linearized equation

$$\dot{\varphi}(t) = A(t)\varphi(t) \quad \text{for } t > 0, \varphi(0) = \varphi_0, \tag{219.10}$$

where $A(t) = f'(\bar{u}(t))$ is an $d \times d$ matrix that now depends on t if $\bar{u}(t)$ depends on t. We have no analytical solution formula to this general problem and thus although the stability properties of the given solution $\bar{u}(t)$ are expressed through the solutions $\varphi(t)$ of the linearized problem (219.10), it may be difficult to analytically assess these properties. We may define stability factors $S(T, \varphi_0)$ and $S(T)$ just as above, and we may say that a solution $\bar{u}(t)$ is stable if $S(T)$ is moderately large, and unstable if $S(T)$ is large. To determine $S(T)$ in general, we have to use numerical methods and solve (219.10) with different initial data φ^0. We return to the computation of stability factors in the next chapter on adaptive solvers for initial value problems.

219.7 Sum Up

The question of stability of solutions to initial value problems is of fundamental importance. We can give an affirmative answer in the case of a stationary solution with corresponding symmetric Jacobian. In this case a positive eigenvalue signifies instability, with the instability increasing with increasing eigenvalue, and all eigenvalues non-positive means stability. The case of an anti-symmetric Jacobian also signifies stability with the norm of perturbations being constant in time. If the Jacobian is non-normal we have to watch out and remember that just looking at the sign of the real part of eigenvalues may be misleading: in the non-normal case algebraic growth may in fact dominate slow exponential decay for finite time. In these cases and also for time-dependent solutions, an analytical stability analysis may be out of reach and the desired information about stability may be obtained by numerical solution of the associated linearized problem.

Chapter 219 Problems

219.1. Determine the stationary solutions to the system

$$\begin{aligned}\dot{u}_1 &= u_2(1-u_1^2),\\ \dot{u}_2 &= 2-u_1u_2,\end{aligned}$$

and study the stability of these solutions.

219.2. Determine the stationary solutions to the following system (Minea's equation) for different values of $\delta > 0$ and γ,

$$\begin{aligned}\dot{u}_1 &= -u_1-\delta(u_2^2+u_3^2)+\gamma,\\ \dot{u}_2 &= -u_2-\delta u_1u_2,\\ \dot{u}_3 &= -u_3-\delta u_1u_3,\end{aligned}$$

and study the stability of these solutions.

219.3. Determine the stationary solutions of the system (219.1) with (a) $f(u) = (u_1(1-u_2), u_2(1-u_1))$, (b) $f(u) = (-2(u_1-10)+u_2\exp(u_1), -2u_2-u_2\exp(u_1))$, (c) $f(u) = (u_1+u_1u_2^2+u_1u_3^2, -u_1+u_2-u_2u_3+u_1u_2u_3, u_2+u_3-u_1^2)$, and study the stability of these solutions.

219.4. Determine the stationary solutions of the system (219.1) with (219.1) with (a) $f(u) = (-1001u_1+999u_2, 999u_1-1001u_2)$, (b) $f(u) = (-u_1+3u_2+5u_3, -4u_2+6u_3, u_3)$, (c) $f(u) = (u_2, -u_1-4u_2)$, and study the stability of these solutions.

219.5. Analyze the stability of the following variant of the linearized problem (219.8) with $\epsilon > 0$ small,

$$\dot{\varphi}(t) = \begin{pmatrix}-\nu & \kappa\\ \epsilon & -\nu\end{pmatrix}\varphi(t) \equiv A_{\nu,\kappa,\epsilon}\varphi(t) \quad \text{for } t>0, \quad \varphi(0)=\varphi^0, \tag{219.11}$$

by diagonalizing the matrix $\equiv A_{\nu,\kappa,\epsilon}$. Note that the diagonalization degenerates as ϵ tends to zero (that is, the two eigenvectors become parallel). Check if $A_{\nu,\kappa,\epsilon}$ is a normal or non-normal matrix.

220
Adaptive Solvers for IVPs

On two occasions I have been asked (by members of Parliament), "Pray, Mr Babbage, if you put into the machine wrong figures, will the right answer come out?". I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question. (Babbage (1792-1871))

220.1 Introduction

In this chapter, we discuss the important issue of *adaptive error control* for numerical methods for initial value problems. This is the subject of automated choice of the time step with the purpose of controlling the numerical error to within a given tolerance level. The basic idea is to combine *feed-back* information from the computation concerning the *residual* of the computed solution and the results of auxiliary computations of *stability factors*. We focus first on on the cG(1) method and then comment on the backward Euler method, also referred to as dG(0), the discontinuous Galerkin method with piecewise constants.

We also discuss the application of cG(1) and dG(0) to a class of so-called *stiff* IVPs typically arising in chemical reaction modeling.

220.2 The cG(1) Method

We recall that cG(1), the continuous Galerkin method with polynomials of order 1, for the initial value problem $\dot{u}(t) = f(u(t))$ for $t > 0$, $u(0) = u^0$, with $f : \mathbb{R}^d \to \mathbb{R}^d$, takes the form

$$U(t_n) = U(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(U(t))\,dt, \quad n = 1, 2, ..., \tag{220.1}$$

where $U(t)$ is continuous piecewise linear with nodal values $U(t_n) \in \mathbb{R}^d$ at an increasing sequence of discrete time levels $0 = t_0 < t_1 < ...$, and $U(0) = u^0$. If we evaluate the integral in (220.1) with the midpoint quadrature rule, we obtain the Midpoint method:

$$U(t_n) = U(t_{n-1}) + k_n f\left(\frac{U(t_n) + U(t_{n-1})}{2}\right), \quad n = 1, 2, ..., \tag{220.2}$$

where $k_n = t_n - t_{n-1}$ is the time step. The cG(1)-method is the first in a family of cG(q)-methods with $q = 1, 2, ...,$, where the solution is approximated by continuous piecewise polynomials of order q. The Galerkin "orthogonality" of cG(1) is expressed by the fact that the method can be formulated

$$\int_{t_{n-1}}^{t_n} (\dot{U}(t) - f(U(t))) \cdot v\,dt = 0, \quad n = 1, 2, ..., \tag{220.3}$$

for all $v \in \mathbb{R}^d$. This says that the *residual*

$$R(U(t) = \dot{U}(t) - f(U(t)), \quad t \in [0, T], \tag{220.4}$$

of the continuous piecewise linear approximate solution $U(t)$ is *orthogonal* to the constant functions $v(t) = v \in \mathbb{R}^d$ on each subinterval (t_{n-1}, t_n). The residual $\dot{u}(t) - f(u(t))$ of the exact solution is zero since $\dot{u}(t) = f(u(t))$, while the residual of $R(U(t))$ of the approximate solution $U(t)$ is non-zero in general. Similarly, in cG(q) the residual is orthogonal on (t_{n-1}, t_n) to polynomials of degree $q - 1$. Note that (220.1) is a vector equation that reads

$$U_i(t_n) = U_i(t_{n-1}) + \int_{t_{n-1}}^{t_n} f_i(U(t))\,dt, \quad n = 1, 2, ..., \; i = 1, ..., d,$$

as can be seen from (220.3) upon setting $v = e_i$, $i = 1, ..., d$.

We will now study the problem of *automatic step-size control* with the purpose of keeping the error

$$\|u(T) - U(T)\| \leq TOL,$$

where $T = t_N$ is a final time and TOL is a given tolerance, while using as few time steps as possible. The objective is the same as that of computing an integral over an interval $[0, T]$ using numerical quadrature to a certain tolerance using as few quadrature points as possible. This is exactly the problem we meet in the case of a scalar initial value problem $\dot{u}(t) = f(u(t), t)$ with $f(u(t), t) = f(t)$.

We shall derive an *a posteriori error estimate* in which the final error $\|u(T)-U(T)\|$ is estimated in terms of the residual $R(U(t) = \dot{U}(t)-f(U(t))$ and certain *stability factors* that measure the *accumulation* of the numerical errors introduced in each time step.

The a posteriori error estimate takes the form

$$\|u(T) - U(T)\| \leq S_c(T) \max_{0\leq t\leq T} \|k(t)R(U(t)\|, \tag{220.5}$$

where $k(t) = k_n = t_n - t_{n-1}$ for $t \in [t_{n-1}, t_n)$ and where the stability factor $S_c(T)$ is defined as follows. We consider the linearized problem

$$-\dot{\varphi}(t) = A^\top(t)\varphi(t) \quad \text{for } 0 < t < T,\ \varphi(T) = \varphi^0, \tag{220.6}$$

where

$$A(t) = \int_0^1 f'(su(t) + (1 - s)U(t))\, ds.$$

We note that replacing $u(t)$ by $U(t)$ gives the following approximate formula for $A(t)$,

$$A(t) \approx f'(U(t)),$$

assuming $U(t)$ is close to $u(t)$. We conclude that $A(t)$ is close to the Jacobian $f'(u(t))$ of $f(v)$ at $v = u(t)$ if $U(t)$ is a reasonable approximation of $u(t)$ Note that the dual $A^\top(t)$ of $A(t)$ occurs in (220.6). Note further that the linearized dual problem (220.6) runs *backward* in time since the initial value $\varphi(T) = \varphi^0$ is specified at time $t = T$. We are now ready to introduce the following stability factors:

$$\begin{aligned} S_d(T) &= \max_{\varphi^0\in\mathbb{R}^d} \frac{\|\varphi(t)\|}{\|\varphi^0\|}, \\ S_c(T) &= \max_{\varphi^0\in\mathbb{R}^d} \frac{\int_0^t \|\dot{\varphi}(s)\|\, ds}{\|\varphi^0\|}, \end{aligned} \tag{220.7}$$

where φ solves (220.6). We note that the stability factors measure different features of the the dual solution φ. The stability factor $S_d(t)$ measures the maximal perturbation growth over the time interval $[0, T]$. We met this factor in the previous chapter. We shall see that this factor is tailored to measure the effect of an error in the initial data u^0 and the "d" in S_d refers to "data". The stability factor $S_c(t)$ measures the integral of $\|\dot{\varphi}\|$ over $[0, T]$ and is geared to evaluate the error in cG(1) and the "c" in S_c refers to "computation".

We shall give the proof of (220.5) below, first in a very simple case with $n = 1$ and $f(u(t)) = au(t)$ with a a constant and then in the general case. The proofs are very similar. Before plunging into the proofs, we shall try to digest the a posteriori error estimate, and see how it can be used to design an adaptive algorithm aiming at controlling the final error $\|u(T) - U(T)\|$ on a given tolerance level with as few time steps as possible.

The stability factors $S_c(T)$ and $S_d(T)$ can be computed by numerically solving the linearized dual problem (220.6) with $\varphi^0 = e_i$ for $i = 1, ..., d$. If d is large, then we may reduce the variation of the initial data by limiting the error control to certain components only, or by trying to choose φ^0 parallel to $u(T) - U(T)$, which we approximate as $U_h(T) - U_H(T)$ with $U_h(T)$ and $U_H(T)$ being approximations computed with two different tolerances.

220.3 Adaptive Time Step Control for cG(1)

We recall the basic error estimate (220.5):

$$\|u(T) - U(T)\| \leq S_c(T) \max_{0 \leq t \leq T} \|k(t)R(t)\|, \tag{220.8}$$

where $R(t) = \dot{U}(t) - f(U(t))$ and we assume that the stability factor $S(T)$ has been computed or estimated. We will return to this issue below. To achieve $\|u(T) - U(T)\| \leq TOL$, we use (220.5) to choose the time steps $k_n = t_n - t_{n-1}$ so that

$$k(t) = k_n \approx \frac{TOL}{S_c(T)R_n} \quad \text{for } t \in [t_{n-1}, t_n), \tag{220.9}$$

where

$$R_n = \max_{t_{n-1} \leq t \leq t_n} \|\dot{U}(t) - f(U(t))\|$$

is the residual on the time interval $[t_{n-1}, t_n)$. Note that the residual R_n is computable from the computed solution $U(t)$ and if $S_c(T)$ is known, timestepalg gives an equation for the time step $k_n = t_n - t_{n-1}$, where t_{n-1} already known. As with adaptive numerical quadrature, (220.9) yields a nonlinear equation for the time step $k_n = t_n - t_{n-1}$ that we can seek to solve using some form of trial-and-error strategy or by prediction, e.g. replacing R_n by R_{n-1}.

220.4 Analysis of cG(1) for a Linear Scalar IVP

We shall now prove an a posteriori error estimate for cG(1) for a a linear scalar IVP of the form

$$\dot{u}(t) = au(t) + f(t) \quad \text{for } t > 0, u(0) = u^0, \tag{220.10}$$

where a is a constant and $f(t)$ is a given function. The analysis is based on representing the error in terms of the solution $\varphi(t)$ of the following dual problem:

$$\begin{cases} -\dot{\varphi} = a\varphi & \text{for } T > t \geq 0, \\ \varphi(T) = e(T), \end{cases} \tag{220.11}$$

where $e = u - U$. Note again that (220.11) runs "backwards" in time starting at time t_N and that the time derivative term $\dot{\varphi}$ has a minus sign. We start from the identity

$$|e(T)|^2 = |e(T)|^2 + \int_0^T e\,(-\dot{\varphi} - a\varphi)\,dt,$$

and integrate by parts to get the following representation of $|e(T)|^2$,

$$|e(T)|^2 = \int_0^T (\dot{e} - ae)\varphi\,dt + e(0)\varphi(0),$$

where we allow $U(0)$ to be different from $u(0)$, corresponding to an error in the initial value $u(0)$. Since u solves the differential equation (220.10), that is $\dot{u} + au = f$, we have

$$\dot{e} - au = \dot{u} - au - \dot{U} + aU = f - \dot{U} + aU,$$

and thus we obtain the following representation of the error $|e(T)|^2$ in terms of the residual $R(U) = \dot{U} - aU - f$ and the dual solution φ ,

$$|e(T)|^2 = \int_0^T (f + aU - \dot{U})\varphi\,dt + e(0)\varphi(0) = -\int_0^{t_N} R(U)\varphi\,dt + e(0)\varphi(0). \tag{220.12}$$

Next, we use the Galerkin orthogonality of cG(1),

$$\int_{t_{n-1}}^{t_n} R(U)\,dt = 0 \quad \text{for } n = 1, 2, ...,$$

to rewrite (220.12) as

$$e(T)^2 = -\int_0^T R(U)(\varphi - \bar{\varphi})\,dt + e(0)\varphi(0), \tag{220.13}$$

where $\bar{\varphi}$ is the mean-value of φ over each time interval, that is

$$\bar{\varphi}(t) = \frac{1}{k_n}\int_{t_{n-1}}^{t_n} \varphi(s)\,ds \quad \text{for } t \in [t_{n-1}, t_n).$$

We shall now use

$$\int_{I_n} |\varphi - \bar{\varphi}|\,dt \leq k_n \int_{I_n} |\dot{\varphi}|\,dt,$$

which follows by integration from the facts that

$$\varphi(t) - \bar{\varphi}(t) = \frac{1}{k_n} \int_{t_{n-1}}^{t_n} (\varphi(t) - \varphi(s))\, ds,$$

and

$$|\varphi(t) - \varphi(s)| \le \int_s^t |\dot{\varphi}(\sigma)|\, d\sigma \le \int_{t_{n-1}}^{t_n} |\dot{\varphi}(\sigma)|\, d\sigma \quad \text{for } s, t \in [t_{n-1}, t_n].$$

Thus, (220.13) implies

$$\begin{aligned} |e(T)|^2 &\le \sum_{n=1}^{N} R_n \int_{I_n} |\varphi - \bar{\varphi}| dt + |e(0)||\varphi(0)| \\ &\le \sum_{n=1}^{N} k_n\, R_n \int_{I_n} |\dot{\varphi}| dt + |e(0)||\varphi(0)|, \end{aligned} \tag{220.14}$$

where

$$R_n = \max_{t_{n-1} \le t \le t_n} |R(U(t))|.$$

Bringing out the max of $k_n R_n$ over n, we get

$$|e(T)|^2 \le \max_{1 \le n \le N} k_n R_n \int_0^{t_N} |\dot{\varphi}|\, dt + |e(0)||\varphi(0)|.$$

Recalling that $\varphi(T) = e(T)$ and using the definitions of $S_c(t_N)$ and $S_d(t_N)$, we get the following final estimate,

$$|e(T)| \le S_c(T) \max_{0 \le t \le T} |k(t) R(U(t))| + S_d(T)|e(0)|.$$

The stability factors $S_c(T)$ and $S_d(T)$ measure the effects of the accumulation of error in the approximation. To give the analysis a quantitative meaning, we have to give a quantitative bound of this factor. The following lemma gives an estimate for $S_c(T)$ and $S_d(T)$ in the cases $a \le 0$ and the case $a \ge 0$ with possibly vastly different stability factors. We notice that the solution $\varphi(t)$ of (220.11) is given by the explicit formula

$$\varphi(t) = e(T) \exp(a(T - t)).$$

We see that if $a \le 0$, then the solution $\varphi(t)$ decays as t decreases from T, and the case $a \le 0$ is thus the "stable case". If $a > 0$ then the exponential factor $\exp(aT)$ enters, and depending on the size of a this case is "unstable". More precisely, we conclude directly from the explicit solution formula that

Lemma 220.1 *The stability factors $S_c(T)$ and $S_d(T)$ satisfy if $a > 0$,*

$$S_d(T) \le \exp(aT), \quad S_c(T) \le \exp(aT), \tag{220.15}$$

and if $a \le 0$, then

$$S_d(T) \le 1, \quad S_c(T) \le 1. \tag{220.16}$$

220.5 Analysis of cG(1) for a General IVP

The extension of the a posteriori error analysis to a general IVP $\dot{u} = f(u)$ with $f : \mathbb{R}^d \to \mathbb{R}^d$ goes as follows. We recall that the linearized dual problem takes the form

$$-\dot{\varphi}(t) = A^\top(t)\varphi(t) \quad \text{for } 0 < t < T,\ \varphi(T) = e(T), \tag{220.17}$$

with

$$A(t) = \int_0^1 f'(su(t) + (1-s)U(t))\,ds,$$

where $u(t)$ is the exact solution and $U(t)$ the approximate solution. We now use the fact that

$$\begin{aligned} A(t)e(t) &= \int_0^1 f'(su(t) + (1-s)U(t))e(t)\,ds \\ &= \int_0^1 \frac{d}{ds} f(su(t) + (1-s)U(t))\,ds = f(u(t)) - f(U(t)), \end{aligned} \tag{220.18}$$

where we used the Chain rule and the Fundamental Theorem of Calculus. We start from the identity

$$\|e(T)\|^2 = \|e(T)\|^2 + \int_0^T e \cdot (-\dot{\varphi} - A^\top\varphi)\,dt,$$

and integrate by parts to get the error representation,

$$\|e(T)\|^2 = \int_0^T (\dot{e} - Ae) \cdot \varphi\,dt + e(0) \cdot \varphi(0),$$

where we allow $U(0)$ to be different from $u(0)$, corresponding to an error in the initial value $u(0)$. Since u solves the differential equation $\dot{u} - f(u) = 0$, (220.18) implies

$$\dot{e} - Ae = \dot{u} - f(u) - \dot{U} + f(U) = -\dot{U} + f(U),$$

and thus we obtain the following representation of the error $\|e(T)\|^2$ in terms of the residual $R(U) = \dot{U} - f(U)$ and the dual solution φ,

$$\|e(T)\|^2 = -\int_0^{t_N} R(U)\varphi\,dt + e(0)\varphi(0). \tag{220.19}$$

From this point, the proof proceeds just as in the scalar case considered above and we thus obtain the following a posteriori error estimate

$$\|e(T)\| \le S_c(T) \max_{0\le t\le T} \|k(t)R(U(t))\| + S_d(T)\|e(0)\|,$$

which can be used a basis for adaptive time step control as described above. The stability factors $S_c(T)$ and $S_d(T)$ may be estimated by solving the dual problem with suitable initial data. The proof of the a posteriori error estimate shows that the stability factors may be defined by

$$\begin{aligned} S_d(T) &= \frac{\|\varphi(t)\|}{\|e(T)\|}, \\ S_c(T) &= \frac{\int_0^t \|\dot{\varphi}(s)\|\,ds}{\|e(T)\|}, \end{aligned} \tag{220.20}$$

where φ solves the linearized dual problem with initial data $\varphi(T) = e(T)$. As indicated, to compute the stability factors $S_d(T)$ and $S_c(T)$, we may solve the dual problem with some estimation of $e(T)$ obtained by solving the initial value problem with two tolerances and approximating $e(T)$ by the difference of the corresponding approximate solutions. Alternatively, choosing $\varphi(T) = e_i$, we obtain a posteriori error control for error component $e_i(T)$. If d is not large, we may this way control all components of the error, and if d is large, we may choose a couple different i at random.

The size of the stability factors indicate the degree of stability of the solution $u(t)$ being computed. If the stability factors are large, the residuals $R(U(t)$ and $e(0)$ have to be made correspondingly smaller by choosing smaller time steps and the computational problem is more demanding.

220.6 Analysis of Backward Euler for a General IVP

We now derive an a posteriori error estimate for the backward Euler method for the IVP (219.1):

$$U(t_n) = U(t_{n-1}) + k_n f(U(t_n)), \quad n = 1, 2, ..., N, \quad U(0) = u^0.$$

We associate a function $U(t)$ defined on $[0, T]$ to the function values $U(t_n)$, $n = 0, 1, ..., N$, as follows:

$$U(t) = U(t_n) \quad \text{for } t \in (t_{n-1}, t_n].$$

In other words, $U(t)$ is piecewise constant on $[0, T]$ and takes the value $U(t_n)$ on $(t_{n-1}, t_n]$, and thus takes a jump from the value $U(t_{n-1})$ to the value $U(t_n)$ at the time level t_{n-1}.

We can now write the backward Euler method as,

$$U(t_n) = U(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(U(t))\,dt,$$

or equivalently

$$U(t_n) \cdot v = U(t_{n-1}) \cdot v + \int_{t_{n-1}}^{t_n} f(U(t)) \cdot v \, dt, \tag{220.21}$$

for all $v \in \mathbb{R}^d$. This method os also referred to as dG(0), that is the *discontinuous Galerkin method of order zero, corresponding to approximating the exact solution by a piecewise constant function $U(t)$ satisfying the orthogonality condition* (220.21).

We are now ready to derive an a posteriori error estimate following the same strategy as for the cG(1) method. We start from the identity

$$\|e(T)\|^2 = \|e(T)\|^2 + \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} e \cdot (-\dot{\varphi} - A^\top \varphi) \, dt,$$

and integrate by parts on each subinterval (t_{n-1}, t_n) to get the following error representation,

$$\|e(T)\|^2 = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (\dot{e} - Ae) \cdot \varphi \, dt - \sum_{n=2}^{N-1} (U(t_n) - U(t_{n-1}))\varphi(t_{n-1}),$$

where the last term results from the jumps of $U(t)$ at the the nodes $t = t_{n-1}$ and we assume $U(0) = u(0)$ for simplicity. Since u solves the differential equation $\dot{u} - f(u) = 0$, (220.18) and the fact that $\dot{U}$ on (t_{n-1}, t_n) imply

$$\dot{e} - Ae = \dot{u} - f(u) - \dot{U} + f(U) = -\dot{U} + f(U) = f(U) \quad \text{on } (t_{n-1}, t_n),$$

and thus we obtain

$$\|e(T)\|^2 = -\sum_{n=2}^{N-1} (U(t_n) - U(t_{n-1}))\varphi(t_{n-1}) + \int_0^{t_N} f(U)\varphi \, dt.$$

Using (220.21) with $v = \bar{\varphi}$, the mean value of φ as above, we get

$$\|e(T)\|^2 = -\sum_{n=2}^{N-1} (U(t_n) - U(t_{n-1})) \cdot (\varphi(t_{n-1}) - \bar{\varphi}(t_{n-1})) + \sum_{n=1}^{n} \int_{t_{n-1}}^{t_n} f(U)(\varphi - \bar{\varphi}) \, dt.$$

We note that

$$\int_{t_{n-1}}^{t_n} f(U)(\varphi - \bar{\varphi}) \, dt = 0,$$

since $f(U(t))$ is constant on $(t_{n-1}, t_n]$, and $\bar{\varphi}$ is the mean value of φ, and thus the error representation takes the final form

$$\|e(T)\|^2 = -\sum_{n=2}^{N-1} (U(t_n) - U(t_{n-1})) \cdot (\varphi(t_{n-1}) - \bar{\varphi}(t_{n-1})).$$

Using

$$\|\varphi(t_{n-1}) - \bar{\varphi}(t_{n-1})\| \leq \int_{t_{n-1}}^{t_n} \|\dot{\varphi}(t)\|\, dt,$$

we obtain the following a posteriori error estimate for the backward Euler method ,

$$\|e(T)\| \leq S_c(T) \max_{1\leq n\leq N} \|U(t_n)) - U(t_{n-1}))\|. \tag{220.22}$$

Note the very simple form of this estimate involving the jumps $\|U(t_n)) - U(t_{n-1}))\|$ playing the role the residual. The a posteriori error estimate (220.22) can be used as a basis for an algorithm for adaptive time step control of the following form: for $n = 1, 2, ...$, choose k_n so that

$$\|U(t_n)) - U(t_{n-1}))\| \approx \frac{TOL}{S_c(T)}.$$

220.7 Stiff Initial Value Problems

A *stiff* initial value problem $\dot{u} = f(u)$ may be characterized by the fact that the stability factors $S_d(T)$ and $S_c(T)$ are of moderate size even for large T, while the norm of the linearized operator $f'(u(t))$ is large, that is the Lipschitz constant L_f is very large. Such initial value problems are common for example in models of chemical reaction with reactions on a range of time scales from slow to fast. Typical solutions include so-called *transients* where the fast reactions make the solution change quickly over a short (initial) time interval, after which the fast reactions are "burned out" and the slow reactions make the solution change on a longer time scale.

The prototype of a stiff initial value problem has the form

$$\dot{u} = f(u) \equiv -Au \quad \text{for } t > 0,\ u(t) = u^0 = (u_i^0), \tag{220.23}$$

where A is a constant symmetric positive semidefinite $d \times d$ matrix with non-negative eigenvalues λ_i ranging from zero to large positive values. Accordingly, the norm of the matrix A is large and L_f is large. By diagonalization, we may reduce to the case when A is a diagonal matrix with non-negative diagonal elements λ_i, in which case the solution $u(t) = (u_i(t))$ is given by

$$u_i(t) = \exp(-\lambda_i\, t)u_i^0 \quad \text{for } t > 0, \tag{220.24}$$

with $u^0 = (u_i^0)$. This explicit solution formula shows that a component $u_i(t)$ corresponding to a large positive eigenvalue λ_i decays very quickly to zero, while a component with a small eigenvalue stays almost constant for a long time and eventually decays to zero. The sign of the eigenvalues is evidently crucial: if some λ_i was negative, then the corresponding solution component would explode exponentially more or less quickly depending on the size of λ_i. In particular, (220.24) with the λ_i non-negative implies

$$\|u(t)\| \leq \|u^0\| \quad \text{for } t > 0, \tag{220.25}$$

which indicates a form of stability with stability factor equal to 1 in the sense that the norm of the solution does not increase in time.

The dual problem corresponding to (220.23) takes the form

$$-\dot{\varphi} + A\varphi = 0 \quad \text{for } T > t > 0,\ \varphi(T) = \psi,$$

with ψ given data at time $t = T$. As a counterpart of (220.25), we conclude that $S_d(T) \leq 1$. We can similarly show that $S_c(T)$ grows very slowly with increasing T. We sum up: (220.23) represents a stiff problem; stability factors are of moderate size even for large T while the norm of the (linearized) operator A is large.

From numerical point of view, stiff problems may seem particularly friendly since the stability factors grow very slowly with time, but there is one hook that has attracted a lot of attention in the literature on numerical methods for initial value problems, namely the failure of an explicit method like the forward Euler method. We write this method for the equation $\dot{u} = -Au$ in the form

$$U^n = U^{n-1} - k_n A U^{n-1}$$

with U^n an approximation of $u(t_n)$ and $0 = t_0 < t_1 < \ldots$ an increasing sequence of time levels, and $k_n = t_n - t_{n-1}$. If A is diagonal with diagonal elements $\lambda_i \geq 0$, then

$$U_i^n = (1 - k_n\lambda_i)U_i^{n-1}$$

and if λ_i is large positive, then $|1 - k_n\lambda_i|$ may be much larger than 1 unless the time step k_n is sufficiently small ($k_n \leq 2/|\lambda_i|$ for all i) and the the numerical solution will then quickly explode to infinity, while the corresponding exact solution quickly decays to zero. The explicit Euler method will thus give completely wrong results unless sufficiently small time steps are used. This may lead to very inefficient time-stepping since after the transients have died out, the solution may vary only slowly and large time steps would be desirable. We note that the time step limit $k_n \leq 2/|\lambda_i|$ for all i, is set by the largest eigenvalue $\max \lambda_i$, while the time long-time scale is set by the smallest eigenvalue $\min \lambda_i$, so that if the quotient $\max \lambda_i / \min \lambda_i$ is large (which signifies a stiff problem), then explicit Euler would be inefficient outside transients.

On the other hand, the dG(0), or implicit Euler method,

$$U^n + k_n A U^n = U^{n-1}$$

with

$$U_i^n = (1 + k_n \lambda_i)^{-1} U_i^{n-1}$$

will be stable and work very well without step size limitation because $1 + k_n \lambda_i \geq 1$ for all $\lambda_i \leq 0$.

For the cG(1)-method, we will have

$$U_i^n = \frac{1 - k_n \lambda_i}{1 + k_n \lambda_i} U_i^{n-1}$$

and stability prevails because

$$|\frac{1 - k_n \lambda_i}{1 + k_n \lambda_i}| \leq 1$$

for all $\lambda_i \geq 0$.

We conclude that both dG(0) and cG(1) may be used for stiff problems, but both these methods are implicit and require the solution of system of equations at each time step. More precisely, dG(0) for a problem of the form $\dot{u} = f(u)$ takes the form

$$U^n - k_n f(U^n) = U^{n-1}.$$

At each time step we have to solve an equation of the form $v - k_n f(v) = U^{n-1}$ with U^{n-1} given. To this end we may try a damped fixed point iteration in the form

$$v^{(m)} = v^{(m-1)} - \alpha(v^{(m-1)} - k_n f(v^{(m-1)}) - U^{n-1}),$$

with α some suitable matrix (or constant in the simplest case). Choosing $\alpha = I$, and iterating once with $v^0 = 0$ corresponds to the explicit Euler method. Convergence of the fixed point iteration requires that

$$\|I + k_n \alpha f'(v)\| < 1$$

for relevant values of v, which could force α to be small (e.g. in the stiff case with $f'(v)$ having large negative eigenvalues) and result in slow convergence. A first try could be to choose α to be a diagonal matrix with $\alpha_i = (f'_{ii})(v^{m-1}))^{-1}$ (corresponding to *diagonal scaling*) and hope that the number of iterations would not be too large. In some cases more efficient iterative solvers would have to be used.

220.8 On Explicit Time-Stepping for Stiff Problems

We just learned that explicit time-stepping for stiff problems require small time steps outside transients and thus may be inefficient. We shall now indicate a way to get around this limitation through a process of stabilization, where a large time step is accompanied by a couple of small time steps. The resulting method has similarities with the control system of a modern (unstable) jet fighter like the Swedish JAS Gripen, the flight of which is controlled by quick small flaps of a pair of small extra wings ahead of the main wings, or balancing a stick vertically on the finger tips if we want a more domestic application.

We shall now explain the basic (simple) idea of the stabilization and present some examples, as illustrations of fundamental aspects of adaptive IVP-solvers and stiff problems. Thus to start with, suppose we apply explicit Euler to the scalar problem

$$\begin{aligned} \dot{u}(t) + \lambda u(t) &= 0 \quad \text{for } t > 0, \\ u(0) &= u^0, \end{aligned} \tag{220.26}$$

with $\lambda > 0$ taking first a large time step K satisfying $K\lambda > 2$ and then m small time steps k satisfying $k\lambda < 2$, to get the method

$$U^n = (1 - k\lambda)^m (1 - K\lambda) U^{n-1}, \tag{220.27}$$

altogether corresponding to a time step of size $k_n = K + mk$. Here K gives a large unstable time step with $|1 - K\lambda| > 1$ and k is a small time step with $|1 - k\lambda| < 1$. Defining the polynomial function $p(x) = (1 - \theta x)^m (1 - x)$, where $\theta = \frac{k}{K}$, we can write the method (220.27) in the form

$$U^n = p(K\lambda) U^{n-1}.$$

For stability we need

$$|p(K\lambda)| \leq 1, \quad \text{that is } |1 - k\lambda|^m (K\lambda - 1) \leq 1,$$

or

$$m \geq \frac{\log(K\lambda - 1)}{-\log|1 - k\lambda|} \approx 2\log(K\lambda), \tag{220.28}$$

with $c = k\lambda \approx 1/2$ for definiteness.

We conclude that m may be quite small even if $K\lambda$ is large, since the logarithm grows so slowly, and then only a small fraction of the total time would be spent on stabilizing time-stepping with the small time steps k.

To measure the efficiency gain we introduce

$$\alpha = \frac{1 + m}{K + km} \in (1/K, 1/k),$$

which is the number of time steps per unit interval with stabilized explicit Euler method, and by (220.28)) we have

$$\alpha \approx \frac{1+2\log(K\lambda)}{K+\log(K\lambda)/\lambda} \approx 2\lambda\frac{\log(K\lambda)}{K\lambda} \ll 2\lambda, \tag{220.29}$$

for $K\lambda \gg 1$. On the other hand, the number of time steps per unit interval for the usual explicit Euler is

$$\alpha_0 = 1/k = \lambda/2, \tag{220.30}$$

choosing a maximum time step $k = 2/\lambda$.

The cost reduction factor using the stabilized explicit Euler method would thus be

$$\frac{\alpha}{\alpha_0} \approx \frac{4\log(K\lambda)}{K\lambda}$$

which can be quite significant for large values of $K\lambda$.

We now present some examples using an adaptive cg(1) IVP-solver in stabilized explicit form with just a few iterations in each time step, which allows large time steps. In all problems we note the initial transient, where the solution components change quickly, and the oscillating nature of the time step sequence outside the transient with large time steps followed by some small stabilizing time steps.

EXAMPLE 220.1. We apply the indicated method to the scalar problem equation (220.26) with $u^0 = 1$ and $\lambda = 1000$, and display the result in Figure 220.1. The cost reduction factor with comparison to a standard explicit method is large: $\alpha/\alpha_0 \approx 1/310$.

EXAMPLE 220.2. We now consider the 2×2 diagonal system

$$\begin{aligned} \dot{u}(t) + \begin{pmatrix} 100 & 0 \\ 0 & 1000 \end{pmatrix} u(t) = 0 \quad \text{for } t > 0, \\ u(0) = u^0, \end{aligned} \tag{220.31}$$

with $u^0 = (1, 1)$. There are now two eigenmodes with large eigenvalues that have to be stabilized. The cost reduction is $\alpha/\alpha_0 \approx 1/104$.

EXAMPLE 220.3. The is the so-called HIRES problem ("High Irradiance RESponse") from plant physiology which consists of the following

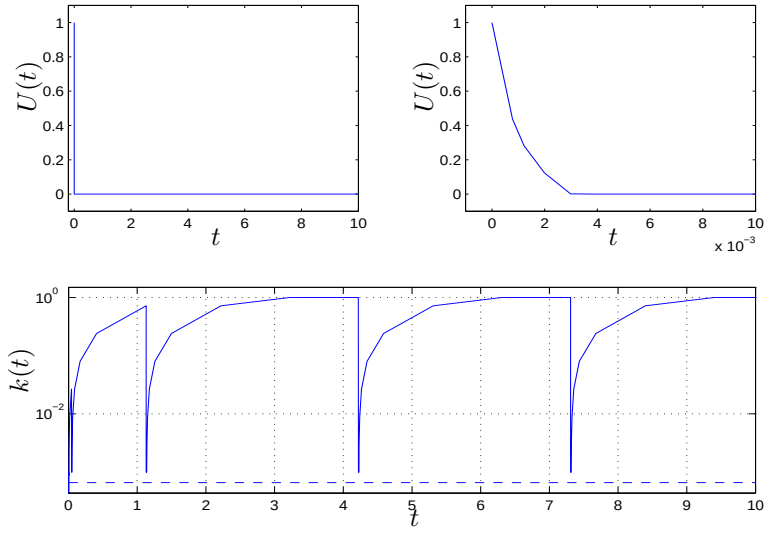


FIGURE 220.1. Solution and time step sequence for eq. (220.26), $\alpha/\alpha_0 \approx 1/310$.

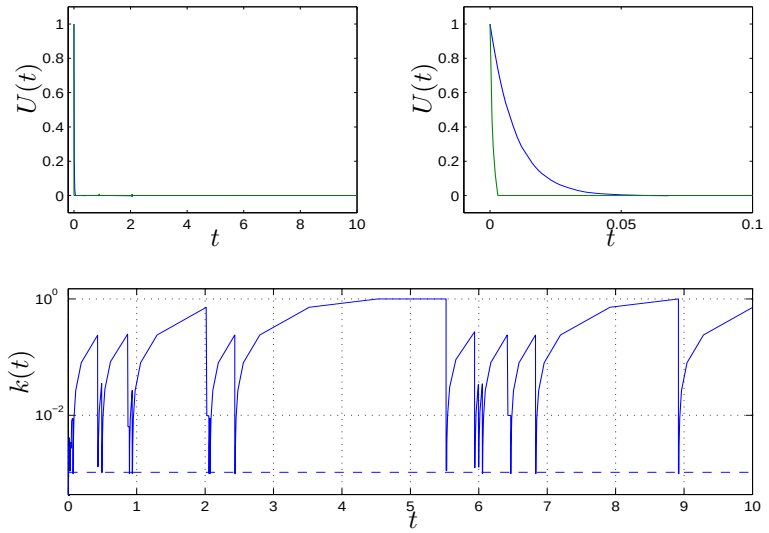


FIGURE 220.2. Solution and time step sequence for eq. (220.31), $\alpha/\alpha_0 \approx 1/104$.

eight equations:

$$\begin{cases} \dot{u}_1 &= -1.71u_1 + 0.43u_2 + 8.32u_3 + 0.0007, \\ \dot{u}_2 &= 1.71u_1 - 8.75u_2, \\ \dot{u}_3 &= -10.03u_3 + 0.43u_4 + 0.035u_5, \\ \dot{u}_4 &= 8.32u_2 + 1.71u_3 - 1.12u_4, \\ \dot{u}_5 &= -1.745u_5 + 0.43u_6 + 0.43u_7, \\ \dot{u}_6 &= -280.0u_6u_8 + 0.69u_4 + 1.71u_5 - 0.43u_6 + 0.69u_7, \\ \dot{u}_7 &= 280.0u_6u_8 - 1.81u_7, \\ \dot{u}_8 &= -280.0u_6u_8 + 1.81u_7, \end{cases} \tag{220.32}$$

together with the initial condition $u^0 = (1.0, 0, 0, 0, 0, 0, 0, 0.0057)$. We present the solution and the time step sequence in Figure 220.3. The cost is now $\alpha \approx 8$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/33$.

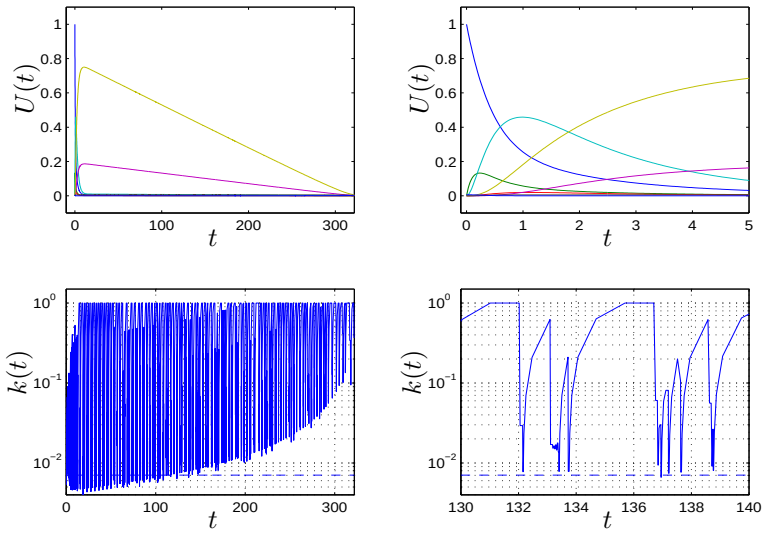


FIGURE 220.3. Solution and time step sequence for eq. (220.32), $\alpha/\alpha_0 \approx 1/33$.

EXAMPLE 220.4. The "Chemical Akzo-Nobel" problem consists of the following six equations:

$$\begin{cases} \dot{u}_1 &= -2r_1 + r_2 - r_3 - r_4, \\ \dot{u}_2 &= -0.5r_1 - r_4 - 0.5r_5 + F, \\ \dot{u}_3 &= r_1 - r_2 + r_3, \\ \dot{u}_4 &= -r_2 + r_3 - 2r_4, \\ \dot{u}_5 &= r_2 - r_3 + r_5, \\ \dot{u}_6 &= -r_5, \end{cases} \tag{220.33}$$

where $F = 3.3 \cdot (0.9/737 - u_2)$ and the reaction rates are given by $r_1 = 18.7 \cdot u_1^4 \sqrt{u_2}$, $r_2 = 0.58 \cdot u_3 u_4$, $r_3 = 0.58/34.4 \cdot u_1 u_5$, $r_4 = 0.09 \cdot u_1 u_4^2$ and $r_5 = 0.42 \cdot u_6^2 \sqrt{u_2}$. We integrate over the interval $[0, 180]$ with initial condition $u^0 = (0.437, 0.00123, 0, 0, 0, 0.367)$. Allowing a maximum time step of $k_{\max} = 1$ (chosen arbitrarily), the cost is only $\alpha \approx 2$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/9$. The actual gain in a specific situation is determined by the quotient between the large time steps and the small damping time steps, as well as the number of small damping steps that are needed. In this case the number of small damping steps is small, but the large time steps are not very large compared to the small damping steps. The gain is thus determined both by the stiff nature of the problem and the tolerance (or the size of the maximum allowed time step).

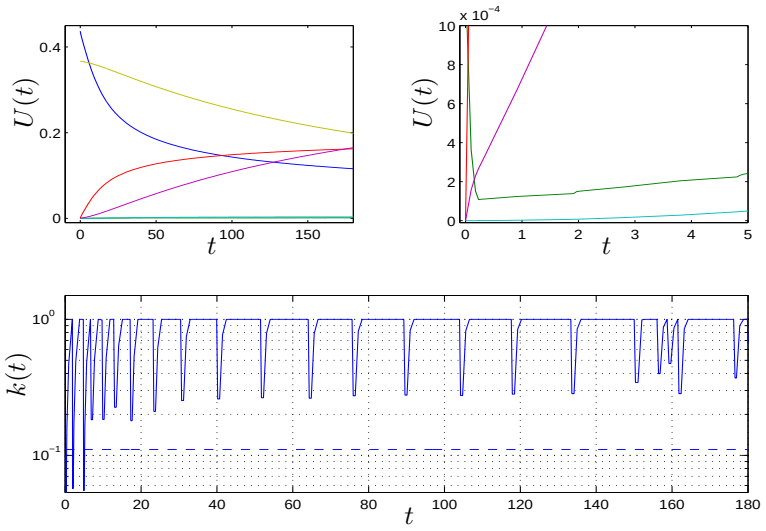


FIGURE 220.4. Solution and time step sequence for eq. (220.33), $\alpha/\alpha_0 \approx 1/9$.

EXAMPLE 220.5. We consider now Van der Pol's equation:

$$\ddot{u} + \mu(u^2 - 1)\dot{u} + u = 0,$$

which we write as

$$\begin{cases} \dot{u}_1 &= u_2, \\ \dot{u}_2 &= -\mu(u_1^2 - 1)u_2 - u_1. \end{cases} \tag{220.34}$$

We take $\mu = 1000$ and solve on the interval $[0, 10]$ with initial condition $u^0 = (2, 0)$. The time step sequence behaves as desired with only a small portion of the time spent on taking small damping steps. The cost is now $\alpha \approx 140$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/75$.

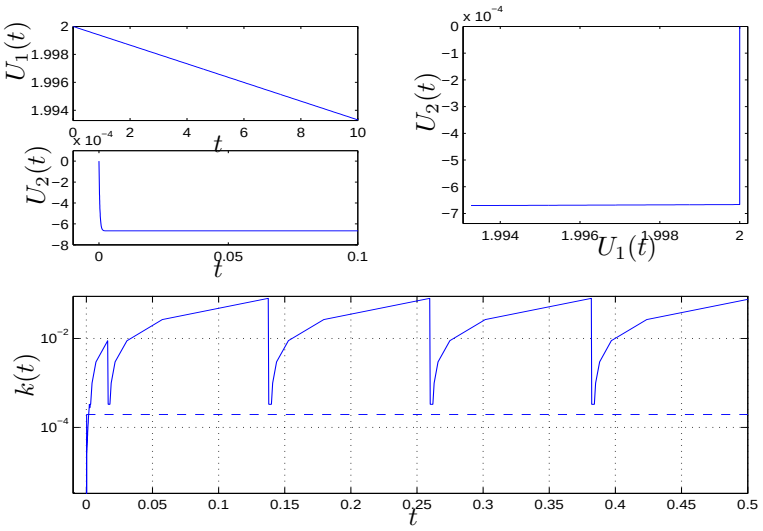


FIGURE 220.5. Solution and time step sequence for eq. (220.34), $\alpha/\alpha_0 \approx 1/75$.

Chapter 220 Problems

220.1. Compute the stability factors $S_d(T)$ and $S_c(T)$ for the linear scalar IVP $\dot{u}(t) = -\lambda(t)u(t)$ for $t > 0$, $u(0) = u^0$, where $\lambda(t)$ depends on time t and (a) $\lambda(t) \geq 0$, (b) $\lambda(t) < 0$.

220.2. Compute $S_d(T)$ and $S_c(T)$ for the linear 2×2 system $\dot{u}_1 = u_2, \quad \dot{u}_2 = -u_1$ for $t > 0$, $u(0) = u^0$.

220.3. Implement adaptive IVP-solvers based on dG(0) and cG(1) and apply the solvers to different problems.

220.4. Show that the a posteriori error estimate for cG(1) may be written on the form $\|e(T)\| \leq S_c(T)\max_{0\leq t\leq T}\|k(t)(f(U(t)) - \bar{f}(U(t)))\| + S_d(T)\|e(0)\|$, where $\bar{f}(U(t))$ is the mean-value of $f(U(t))$ over each time interval.

220.5. Show that choosing in the dual problem $\varphi(T) = e_i$ gives control of error component $e_i(T)$.

220.6. Develop explicit versions of dG(0) and cG(1) based on fixed point iteration at each time step. Show that with diagonal scaling such an explicit method may work very well for some stiff problems.

221
Optimization

1. All living beings are driven by passion to seek maximal Pleasure. 2. There is Pleasure of the Body and Pleasure of the Soul. In the Pleasure of the Soul, the Body cannot take part, whereas the Pleasure of the Body is equally shared by the Soul. (the 2 first of the 14 basic principles of *Anthropologica physica*, by King Karl XII of Sweden, 1717)

221.1 Introduction

In this chapter, we expand on some basic aspects of optimization touched upon in the previous chapter in connection to minimization. Optimization is very rich subject and we shall return to other aspects below. The issues we discuss here are connected to the very basics of Calculus and are considered as "deep" and understandable only by the very best math majors. You may test your own reaction to the discussions presented, and if you get the expected feeling of confusion, don't worry, just proceed to the next chapter. If on the other hand, against all odds, you get the feeling of grasping the main ideas, then you may congratulate yourself for being more gifted for mathematics than you thought!

In our modern world, *optimization* is a code word. To *optimize* is to use available resources as efficiently as possible, or to find the best of available alternatives. In our private lives, we may want our car to use as little fuel as possible, to buy an item at lowest possible price, to use as little effort

as possible to clean the house, or to get maximal enjoyment out of the vacation trip.

In automatized production, the leading principle is always to optimize and seek to use as little energy, material and human resources as possible to produce a certain amount of goods. A basic idea in our capitalistic system is that in the long run the most efficient mode of production will win the market.

A basic problem of optimization is to find the maximum or minimum value of a given function $f : \Omega \to \mathbb{R}$ defined on some set of numbers Ω. Typically, Ω may be a domain in $\mathbb{R}^d$ with $d = 1, 2, 3, ...$, that may be bounded or unbounded, or Ω may be a finite set such as the set of natural numbers $1, 2..., 100$. More precisely, finding a *minimum point* $\bar{x}$ in Ω amounts to finding a point $\bar{x} \in \Omega$ such that

$$f(x) \geq f(\bar{x}) \quad \text{for all } x \in \Omega, \tag{221.1}$$

and we then say that $f(\bar{x})$ is the *minimum value* of $f : \Omega \to \mathbb{R}$. Note that there may be several minimum points, but of course there may be only one minimum value. If in an Olympic 100 meter race, three runners share the best time of 9.99 seconds, then all the three runners may share the gold medal. However, there cannot be two runners with different final times who both get a gold medal.

We now consider the problem of finding the minimum value and corresponding minimum point(s) of a given function $f : \Omega \to \mathbb{R}$. We may distinguish between the following two cases: (a) Ω is a domain of $\mathbb{R}^d$ with infinitely many points, as when Ω is the unit disc $\{x \in \mathbb{R}^d : \|x\| \leq 1\}$; (b) Ω contains finitely many points, as for example $\Omega = \{1, 2, 3, ..., 10\}$. The case (a) is "continuous" and (b) is "discrete". The two cases are not fully disjoint; there may be a gradual passage from "discrete" to "continuous" as the number of elements in Ω increases. In the case Ω is discrete with finitely many points, we may find the minimum value and corresponding minimum point(s) of $f : \Omega \to \mathbb{R}$ by different algorithms for *sorting*. If Ω is "continuous" with infinitely many points, sorting may be impossible and different algorithms that use information from the derivative of $f(x)$ in variations of steepest descent are often used.

221.2 Sorting if Ω Is Finite

If Ω is a finite set of numbers, for example if $\Omega = \{1, 2, ..., 9, 10\}$, then we just make a list of the corresponding 10 function values $f(1)$, $f(2)$,...,$f(10)$ and by sorting these values according to magnitude in increasing order, we can find the minimum value $f(\bar{x})$ and the corresponding argument $\bar{x}$. Of course, we don't have to sort all the numbers according to magnitude to find the smallest one. We just have to find the first element in the list

sorted according to magnitude in increasing order. Repeating this process we can sort all the numbers in the given list of numbers.

EXAMPLE 221.1. For example, suppose that $\Omega = \{1, 2, ..., 9, 10\}$ and $f(1) = 143$, $f(2) = 538$, $f(3) = 67$, $f(4) = 964$, $f(5) = 287$, $f(6) = 64$, $f(7) = 123$, $f(8) = 333$, $f(9) = 63$, $f(10) = 88$. By direct inspection, we see that the minimum point is $\bar{x} = 9$ and the minimum value is $f(\bar{x}) = 63$.

While sorting sounds simple, it turns out to be an interesting problem to do sorting efficiently when there are a large number of values to be sorted. So there are different algorithms for sorting and sorting algorithms hold a prominent place in computer science. A simple algorithm for finding the minimum m of N numbers $f(1), \cdots, f(N)$ goes as follows:

1. Set $m = f(1)$ and $\bar{x} = 1$
2. For $x = 2, \cdots, N$, if $f(x) < m$ set $m = f(\bar{x})$ and $\bar{x} = x$.

The minimum value is then $m = f(\bar{x})$ and the minimum point is $x = \bar{x}$. The algorithm is based on successive comparison of pairs of numbers (if $f(x) < m$ then we update m and $\bar{x}$ and set $m = f(\bar{x})$ and $\bar{x} = x$). The number of comparisons in the indicated algorithm is apparently $N - 1$.

Repeating the algorithm with $f(\bar{x})$ eliminated, we can get a complete sorting according to magnitude using $(N-1) + (N-2) + \cdots + 1 \approx \frac{1}{2}N^2$ comparisons.

221.3 What if Ω Is Not Finite?

If Ω is an interval of real numbers, for example $\Omega = [0, 1]$, then Ω contains infinitely many points and sorting the values $f(x)$ with $x \in \Omega$ of a given function $f : \Omega \to \mathbb{R}$ by pairwise comparison appears impossible in practice because we cannot perform infinitely many comparisons. Of course, in practice we replace Ω by a finite set of numbers, for example by using a single precision floating point representation of the points in Ω. So in principle, we can then apply the above sorting strategy. But, the procedure will be computationally intensive. With seven digits we would have 10^7 values $f(x)$ to compare, which using the above algorithm requires on the order of 10^7 comparisons to find the minimum. If the interval Ω is larger and the desired precision higher, then the number of comparisons would be correspondingly larger. The total computational cost would also involve as a multiplicative factor the cost of evaluating the function value $f(x)$ for a given x, which itself could require many arithmetic operations. The total cost in direct comparison thus may be prohibitively large.

We now seek efficient algorithms to handle the case that Ω is a domain of $\mathbb{R}^d$ interval and the function $f(x)$ is Lipschitz continuous with a Lipschitz

constant L. In this case, the function values $f(x)$ cannot change more than the argument of x changes times L. If we want to find the minimum value up to a certain tolerance TOL, then we need to do approximately $(L/TOL)^d$ comparisons if the diameter of Ω is of order one. Depending on the choice of the tolerance TOL and L this may be acceptable or not.

If the function $f(x)$ is differentiable then we may restrict the search further using information from the derivative, as we shall see below.

221.4 Existence of a Minimum Point

How can we be sure that there in fact is a minimum point? We discuss the proof of the following basic theorem addressing this question below.

Theorem 221.1 *If $f : \Omega \to \mathbb{R}$ is Lipschitz continuous, where Ω is a closed and bounded subset of $\mathbb{R}^d$, then there is a minimum point $\bar{x} \in \Omega$ such that $f(\bar{x}) \leq f(x)$ for all $x \in \Omega$.*

The assumption that Ω is closed and bounded is essential to guarantee existence of the minimum, as the following examples show.

EXAMPLE 221.2. The function $f : (0,1) \to \mathbb{R}$ with $f(x) = x$ does not have a minimum point in $(0,1)$. In this case $\Omega = (0,1)$ is not closed.

EXAMPLE 221.3. The function $f : [1,\infty) \to \mathbb{R}$ with $f(x) = 1/x$ does not have a minimum point in $[1,\infty)$. In this case $\Omega = [1,\infty)$ is not bounded.

Note however that a function $f : \Omega \to \mathbb{R}$ may have a minimum even if Ω is unbounded. In particular, if $f(x)$ increases to infinity as $\|x\|$ increases, then we can effectively reduce the search for a minimum to a bounded set.

EXAMPLE 221.4. The function $f : [0,\infty)$ given by $f(x) = x^2 - 2x$ attains a minimum value $f(1) = -1$; since $f(x) \geq 0$ for $x \geq 2$, we may restrict the search for a minimum to $[0,2]$.

221.5 The Derivative Is Zero at an Interior Minimum Point

We assume that $f : \Omega \to \mathbb{R}$ is a given Lipschitz continuous differentiable function, where Ω is a domain in $\mathbb{R}^d$. We shall now prove that if $\bar{x}$ is an *interior minimum point* of $f : \Omega \to \mathbb{R}$, that is $\bar{x}$ is a minimum point and the ball $\{x \in \mathbb{R}^d : \|x - \bar{x}\| < \delta\}$ is included in Ω for some positive number δ, then $f'(\bar{x}) = \nabla f(\bar{x}) = 0$, where $f' = \nabla f$ is the gradient of f. This follows by writing

$$f(x) = f(\bar{x}) + f'(\bar{x}) \cdot (x - \bar{x}) + E_f(x, \bar{x})$$

with $|E_f(x,\bar{x})| \leq K_f(\bar{x})\|x-\bar{x}\|^2$. If now $f'(\bar{x}) \neq 0$, we may choose $x = \bar{x} - \epsilon f'(\bar{x}) \in \Omega$ with $\epsilon > 0$ and estimate to get

$$\begin{aligned} f(x) &\leq f(\bar{x}) - \epsilon\|f'(\bar{x})\|^2 + \epsilon^2 K_f(\bar{x})\|f'(\bar{x})\|^2 \\ &= f(\bar{x}) - \epsilon\|f'(\bar{x})\|^2(1-\epsilon K_f(\bar{x})) < f(\bar{x}). \end{aligned}$$

For ϵ sufficiently small, we get a contradiction to the assumption that $\bar{x}$ is a minimum point. We have proved the following basic result, see Fig. 221.1 and Fig. 221.2.

Theorem 221.2 *Suppose $f : \Omega \to \mathbb{R}$ has a minimum point at an interior point $\bar{x}$ in Ω, and suppose that $f : \Omega \to \mathbb{R}$ is differentiable at $\bar{x}$. Then $f'(\bar{x}) = 0$.*

Using this result, we may search for interior minimum points among the zeros of the derivative $f'(x)$ in Ω. To find these zeros we may use some algorithm for computing roots, like Fixed Point Iteration or Newton or the Bisection algorithm. There is thus a strong connection between algorithms for finding interior minimum points of $f : \Omega \to \mathbb{R}$ and algorithms for computing roots of $f'(x) = 0$.

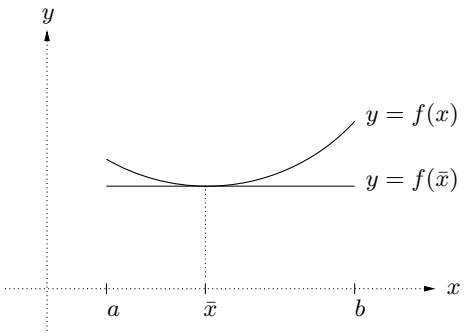


FIGURE 221.1. $f'(\bar{x}) = 0$ at an interior minimum point $\bar{x}$.

Note that if the minimum point $\bar{x}$ of $f : \Omega \to \mathbb{R}$ is not interior to Ω, i.e. $\bar{x}$ lies on the boundary of Ω, then the derivative $f'(\bar{x})$ may be non-zero, see Fig. 221.3.

EXAMPLE 221.5. Suppose we want to minimize $f : \Omega \to \mathbb{R}$ with $\Omega = [0,2]$ and $f(x) = x^2 - 2x$. Since Ω is closed and bounded and $f(x)$ is Lipschitz continuous, we know that there is a minimum point $\bar{x} \in [0,2]$. If $\bar{x}$ is interior to $[0,2]$, that is if $0 < \bar{x} < 2$, then $f'(\bar{x}) = 2\bar{x} - 2 = 0$ and thus $\bar{x} = 1$. We compare the value $f(1) = -1$ to the values $f(0) = 0$ and $f(2) = 0$ on the boundary of $[0,2]$ and conclude that $f(1) = -1$ is the minimum value and $\bar{x} = 1$ the corresponding minimum point.

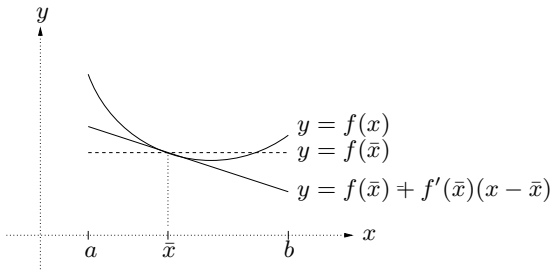


FIGURE 221.2. $f'(\bar{x}) < 0$ implies that $f(x) < f(\bar{x})$ for x close to $\bar{x}$ with $\bar{x} > x$, that is, $\bar{x}$ cannot be a minimum point.

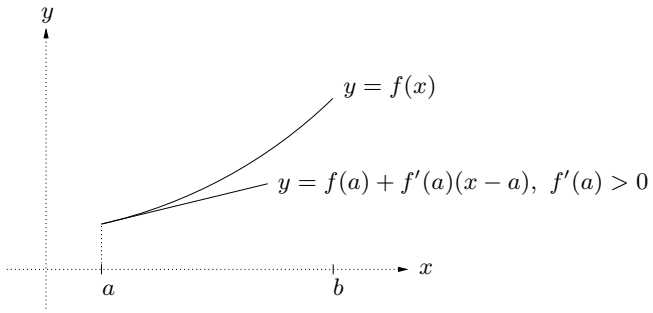


FIGURE 221.3. $f'(\bar{x})$ may be non-zero at a minimum $\bar{x}$ on the boundary.

EXAMPLE 221.6. Suppose we want to minimize $f : Q \to \mathbb{R}$ with $f(x) = f(x_1, x_2) = x_1^2 + x_2^2 - 2x_1 - x_2$ on a closed square $Q = [0, 2] \times [0, 2]$, see Fig. 221.4. We know that there is a minimum point in Q. We first compute the interior points $\hat{x}$ where $f'(\hat{x}) = 0$. Since $f'(x) = (2x_1 - 2, 2x_2 - 1)$, $\hat{x} = (1, 0.5)$ with the function value $f(1, 0.5) = -1.25$. It remains to study the variation of $f(x)$ on the boundary of Q to see if we find a value smaller than -1.25. We do this by considering each piece of the boundary separately. On the part $x_2 = 0$, we have $f(x) = x_1^2 - 2x_1$ with $x_1 \in [0, 2]$, and we see arguing as in the previous example that the minimum value is $f(1, 0) = -1$. On the part $x_2 = 2$, we have $f(x_1, 2) = x_1^2 - 2x_1 + 3$ with minimum $f(1, 2) = 2$. On the part $x_1 = 0$, we have $f(0, x_2) = x_2^2 - x_2$ with minimum $f(0, 0.5) = -0.25$, and on the part $x_1 = 2$, we have $f(2, x_2) = x_2^2 - x_2$ with minimum $f(2, 0.5) = -0.25$. We conclude that the minimum point is the interior point $\bar{x} = (1, 0.5)$ and that the minimum value is $f(1, 0.5) = -1.25$.

EXAMPLE 221.7. You are asked to design a box (without top) of a given volume using as little material as possible. Letting the sides of the box be x_1, x_2 and x_3, the volume is $x_1 x_2 x_3 = V$ and the surface to

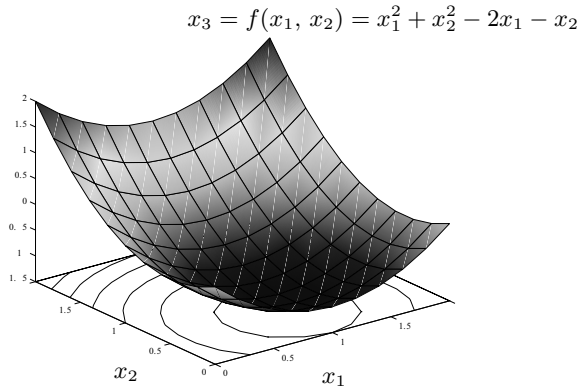


FIGURE 221.4. Minmizing $f(x) = x_1^2 + x_2^2 - 2x_1 - x_2$ on $Q = [0, 2] \times [0, 2]$

be minimized is $x_1x_2 + 2x_1x_3 + 2x_2x_3$. Eliminating x_3 gives

$$f(x_1, x_2) = x_1x_2 + 2V(\frac{1}{x_1} + \frac{1}{x_2}),$$

which is to be minimized over $\Omega = [0, \infty) \times [0, \infty)$. Seeking points $\hat{x}$ with $f'(\hat{x}) = (0, 0)$, we find $\hat{x}_1 = \hat{x}_2 = (2V)^{1/3}$ with the corresponding height $\hat{x}_3 = \frac{1}{2}(2V)^{1/3}$, and the area

$$f(\hat{x}) = (2V)^{2/3} + 2(2V)^{2/3}.$$

Comparing with (x_1, x_2) with x_1 or x_2 very large or small give large values to $f(x_1, x_2)$ and thus the minimum point is $\hat{x}$. The solution is a box with square bottom and height half of the width.

We also remark that a minimum value may be attained at an interior point where the given function is *nondifferentiable*. For example, the minimum value of the function $f(x) = |x - 1|$ on $[0, 2]$ is attained at $\bar{x} = 1$ with minimum value $f(\bar{x}) = f(1) = 0$. This type of interior minimum points must be considered separately. Thus, to find all possible minimum points we have to consider the points $\bar{x}$ for which $f'(\bar{x}) = 0$, and in addition to these the end points of the domain of definition and interior points where $f(x)$ is not differentiable, see Fig. 221.5.

221.6 The Role of the Hessian

We know that if $\bar{x}$ is an interior minimum point of a function $f : \Omega \to \mathbb{R}$, then $f'(\bar{x}) = 0$. But it is not true in general that if $f'(\bar{x}) = 0$, then $\bar{x}$ is a minimum point. A point $\bar{x}$ with $f'(\bar{x}) = 0$ may e.g. be a *maximum point*, or

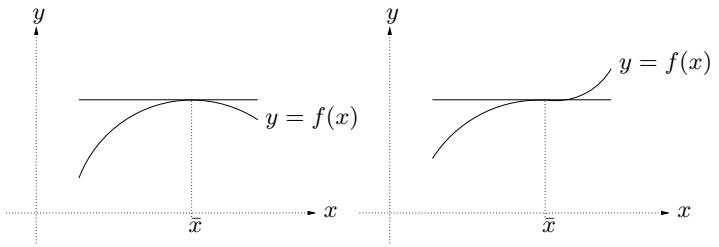


FIGURE 221.5. $f'(\bar{x}) = 0$ may correspond also to a maximum point or an inflection point.

an *inflection point*, see Fig. 221.5. If the Hessian H of $f : \Omega \to$ is positive definite close to $\bar{x}$ and $f'(\bar{x}) = 0$, then we have by Taylor's theorem

$$f(x) = f(\bar{x}) + \frac{1}{2}(x - \bar{x})^\top H(y) \cdot (x - \bar{x}) > f(\bar{x})$$

for x close to $\bar{x}$ and some y between x and $\bar{x}$, and thus $\bar{x}$ is a *local minimum point*.

We recall that an $n \times n$ matrix A is said to be positive definite if

$$v^\top Av > 0$$

for all non-zero $v \in \mathbb{R}^n$. The Spectral Theorem implies that A is positive definite if and only if the eigenvalues of A are positive.

EXAMPLE 221.8. If $A = (a_{ij})$ is a symmetric 2×2 matrix, then A is positive definite if

$$a_{11}a_{22} - a_{12}^2 > 0 \quad \text{and } a_{11} > 0.$$

This follows by completing squares in

$$v^\top Av = a_{11}v_1^2 + a_{22}v_2^2 + 2a_{12}v_1v_2.$$

221.7 Minimization Algorithms: Steepest Descent

We discuss briefly how to find candidates for minimum points of a given function $f : \Omega \to \mathbb{R}$, where Ω is a domain in $\mathbb{R}^d$. We assume that $f : \Omega \to \mathbb{R}$ is Lipschitz continuous and differentiable on Ω.

In the *steepest descent method*, we construct a sequence $\{x_i\}$ in $\mathbb{R}^d$ that hopefully converges to a (local) minimum point by means of the iteration

$$x_{i+1} = x_i - \alpha_i f'(x_i), \tag{221.2}$$

where α_i is a positive parameter. Since $\alpha_i > 0$, if $f'(x_i) > 0$ then $x_{i+1} < x_i$, and if $f'(x_i) < 0$ then $x_{i+1} > x_i$. This means that if $f'(x_i) > 0$, so

that $f(x)$ is increasing at $x = x_i$, then taking $x_{i+1} < x_i$ should result in $f(x_{i+1}) < f(x_i)$, and thus x_{i+1} should be closer to a minimum point than x_i. A similar argument applies in the case $f'(x_i) < 0$.

It is clear that the choice of the parameter α_i is important. If α_i is too small, then the convergence will be slow, and if α_i is too large, the sequence x_i may start to oscillate.

Note that we may view the gradient method (221.2) for minimization of $f(x)$ as Fixed Point Iteration for computing a root of $f'(x) = 0$.

If steepest descent leads to the boundary Γ of Ω, then we may replace the steepest descent iteration by the *projected gradient method* defined

$$x_{i+1} = x_i - \alpha_i P f'(x_i),$$

where $Pf'(x_i)$ is the projection of $f'(x_i)$ onto the tangent plane to Γ at $x_i \in \Gamma$.

The general idea is thus to find roots of $f'(x) = 0$ using steepest descent for the minimization of $f(x)$ or equivalently fixed point iteration for $f'(x) = 0$. Once the roots of $f'(x) = 0$ have been determined, the minimization is reduced to a search on the boundary of Ω and the interior zeros of $f'(x)$.

221.8 Existence of a Minimum Value and Point

We return to proof of the fundamental result which says that if $f : \Omega \to \mathbb{R}$ is Lipschitz continuous and Ω is a closed and bounded domain of $\mathbb{R}^d$, then there is a minimum point $\bar{x} \in \Omega$ with corresponding minimum value $f(\bar{x})$. We carry out the proof in for $d = 1$ so that $\Omega = [a, b]$ is a bounded closed interval. The proof in the case $d > 1$ is similar.

We shall prove that a Lipschitz continuous function $f : [a, b] \to \mathbb{R}$ on a closed and bounded interval $[a, b]$ has a minimum point by "constructing" a minimum point using the Bisection algorithm. We shall see that the "construction" is controversial at one step. Trying to resolve this issue yields added insight into the nature of minimization algorithms.

Normally, the proof we present here is considered so "difficult" that it is given only in "advanced" senior undergraduate or beginning graduate courses. With our good preparation on the Bisection algorithm and the nature of real numbers, we can plunge into the proof, and we will see that it is "easy" up to the non-constructive aspects.

We first recall that the Lipschitz continuity of $f(x)$ and the fact that $[a, b]$ is bounded implies that $f : [a, b] \to \mathbb{R}$ is bounded from above and below. In particular, there is some $m \in \mathbb{R}$ such that

$$f(x) \geq m \quad \text{for all } x \in [a, b]. \tag{221.3}$$

We say that m is a *lower bound* of $f : [a, b] \to \mathbb{R}$ if (221.3) holds. Clearly, there are many lower bounds since if m is a lower bound, any number $\underline{m} < m$ is also a lower bound.

In the proof, we shall use the concept of *greatest lower bound* defined as follows: we say that $\overline{m}$ is a greatest lower bound of $f : [a, b] \to \mathbb{R}$ if

$$f(x) \geq \overline{m} \quad \text{for all } x \in [a, b] \tag{221.4}$$

$$\text{for all } M > \overline{m} \quad \text{there is some } x \in [a, b] \quad \text{such that } f(x) < M. \tag{221.5}$$

In words, $\overline{m}$ is a greatest lower bound of $f : [a, b] \to \mathbb{R}$ if $\overline{m}$ is a lower bound of $f : [a, b] \to \mathbb{R}$ and any number bigger than $\overline{m}$ is not a lower bound for $f : [a, b] \to \mathbb{R}$. The concept of greatest lower bound has played an important role in the development of Calculus during the 20th century.

The proof now proceeds in two steps:

Step 1: Existence of a Greatest Lower Bound $\overline{m}$ of $f : [a, b] \to \mathbb{R}$

We shall prove the existence of a greatest lower bound $\overline{m}$ by using the Bisection method. Let m be a lower bound of $f : [a, b] \to \mathbb{R}$, whose existence was established above. Set $y_0 = m$ and $Y_0 = f(b)$ and define $\hat{y}_1 = \frac{1}{2}(y_0 + Y_0) = \frac{1}{2}(m + f(b))$. Note that $y_0 \leq \hat{y}_1 \leq Y_0$. If $f(x) \geq \hat{y}_1$ for all $x \in [a, b]$, then set $y_1 = \hat{y}_1$ and $Y_1 = Y_0$. If not, then there is an $x \in [a, b]$ such that $f(x) < \hat{y}_1$, and we set $y_1 = m$ and $Y_1 = \hat{y}_1$. We have now passed from the pair (y_0, Y_0), or interval (y_0, Y_0), to the interval (y_1, Y_1). By construction, $f(x) \geq y_i$ for all $x \in [a, b]$ and $i = 0, 1$ and there is some $x \in [a, b]$ such that $f(x) < Y_i$ unless Y_0 or Y_1 is already a greatest lower bound.

Repeating this process, we get two sequences $\{y_i\}$ and $\{Y_i\}$ such that for $i = 0, 1, 2, ...,$,

$$\begin{gathered} y_i < Y_i, \quad y_{i+1} \geq y_i \quad Y_{i+1} \leq Y_i, \\ 0 < Y_i - y_i = 2^{-i}(Y_0 - m), \\ f(x) \geq y_i \quad \text{for all } x \in [a, b], \\ \text{there is an } x \in [a, b] \quad \text{such that } f(x) < Y_i, \end{gathered}$$

or some Y_i is a greatest lower bound. As in Chapter $\sqrt{2}$, we see that the sequences $\{y_i\}$ and $\{Y_i\}$ are Cauchy sequences and both converge to one real number, which we denote by $\overline{m}$. The number $\overline{m}$ is the greatest lower bound of $f : [a, b] \to \mathbb{R}$ since $\overline{m}$ satisfies the following two conditions:

$$\begin{gathered} (f(x) \geq \overline{m} \quad \text{for all } x \in [a, b], \\ \text{for any } M > \overline{m} \text{ there is an } x \in [a, b] \quad \text{such that } f(x) < M. \end{gathered}$$

We have now proved the existence of a greatest lower bound to the Lipschitz continuous function $f : [a, b] \to \mathbb{R}$ on the closed and bounded interval $[a, b]$. Note that this result also holds if (a, b) is a bounded open interval. We have thus not yet used the fact that $[a, b]$ is closed.

Step 1: Existence of a Minimum Point

We now construct a convergent sequence $\{x_i\}$ with $x_i \in [a,b]$ and

$$\lim_{i\to\infty} f(x_i) = \overline{m}.$$

Setting $\bar{x} = \lim_{i\to\infty} x_i$, we have $f(\bar{x}) = \overline{m}$ and thus $\bar{x}$ is a minimum point and we are done.

To construct $\{x_i\}$ we again use the Bisection algorithm as follows: set $x_0 = a$, and $X_0 = b$, and define $\hat{x}_1 = \frac{1}{2}(x_0 + X_0)$. If $f(x) > \overline{m}$ for all x such that $\hat{x}_1 < x \leq X_1$, then we set $x_1 = x_0$ and $X_1 = \hat{x}_1$. If not, we set $x_1 = \hat{x}_1$ and $X_1 = X_0$. Repeating the process, we obtain a convergent sequence $\{x_i\}$ with limit $\bar{x}$ and by construction we have $f(\bar{x}) = \overline{m}$. Note that to guarantee that $\overline{x} \in [a,b]$, we need $[a,b]$ to be closed. We note that the minimum value (of course) is equal to the greatest lower bound.

We summarize in the following theorem:

Theorem 221.3 (Existence of minimum point) *Suppose $f : I \to \mathbb{R}$ is Lipschitz continuous and $I = [a,b]$ is a closed and bounded interval. Then there is a point $\bar{x} \in [a,b]$, where $f : I \to \mathbb{R}$ assumes a minimum value $\bar{m}$, that is, $f(x) \geq \bar{m}$ for all $x \in [a,b]$, and $f(\bar{x}) = \bar{m}$.*

In the proof of this theorem we used the Bisection algorithm twice. Setting $y = f(x)$, we may say that we first used the Bisection algorithm in the variable y to prove existence of a greatest lower bound $\overline{m}$ and then in the variable x to prove existence of a minimum point $\bar{x}$ satisfying $f(\bar{x}) = \overline{m}$.

221.9 Existence of Greatest Lower Bound

If we examine the proof of existence of a greatest lower bound to the Lipschitz continuous function $f : I \to \mathbb{R}$, we see that the crucial fact behind the proof is that $f : [a,b] \to \mathbb{R}$ is bounded below, that is there is a real number m such that $f(x) \geq m$ for all $x \in [a,b]$. We can interpret this in terms of a property of the range $R(f) = \{y : y = f(x) \text{ for some } x \in D(f) = [a,b]\}$, namely

$$y \geq m \quad \text{for all } y \in R(f).$$

This says that the set $R(f)$ is *bounded below*.

More generally, we say that a set A of real numbers is bounded from below if there is a real number m such that $y \geq m$ for all $y \in A$. Using the same argument as just used in the case $A = R(f)$, we obtain the following fundamental property of real numbers.

Theorem 221.4 (Existence of greatest lower bound) *Suppose A is a set of real numbers which is bounded from below, that is, there is a real number m such that $x \geq m$ for $x \in A$. Then the set A has a greatest lower*

bound $\overline{m} \in \mathbb{R}$ *satisfying* $x \geq \overline{m}$ *for all* $x \in A$ *and for all* $M > \overline{m}$ *there is an* $x \in A$ *such that* $x < M$.

221.10 Constructibility of a Minimum Value and Point

We now discuss to what extent the above existence proof is constructive. There are two issues: (i) construction of the greatest lower bound, which is the same as the minimum value, and (ii) construction of a minimum point.

In the application of the Bisection algorithm in (i), we have to check if

$$f(x) \geq \hat{y}_i \quad \text{for all } x \in [a, b],$$

while in the application in (ii), we have to check if

$$f(x) > \overline{m} \text{ for all } x \text{ such that } \hat{x}_1 < x \leq X_1.$$

Both checks appear to involve *infinitely many* values of x. In the worst case this would require infinitely many comparisons. The number may be reduced if $f(x)$ is differentiable by using information concerning $f'(x)$. For example, the sign of $f'(x)$ indicates if $f(x)$ is increasing or decreasing which may be used to reduce the amount of comparison.

Thus, depending on the nature of the given function $f : I \to \mathbb{R}$, the given proof of existence of a minimum value and minimum point may be more or less constructive in nature.

Is it possible to make the proof fully constructive? We expect this to be possible if we accept to determine the minimum value up to a tolerance $TOL > 0$. Suppose then that the function $f(x)$ is Lipschitz continuous with Lipschitz constant L. We can then reduce all comparisons to a discrete grid of points of mesh size $\frac{1}{L}TOL$ between neighboring points.

To sum up, if $f : I \to \mathbb{R}$ is Lipschitz continuous and $[a, b]$ is bounded, then it is possible to determine the minimum value $f : I \to \mathbb{R}$ up to a given tolerance with a finite number of operations.

To determine an interior minimum point amounts to finding a root of $f'(x) = 0$ and thus the constructibility of a minimum point can be reduced to the constructibility of a root of $f'(x) = 0$. We discussed the cost of computing roots in Chapters *Fixed Point Iteration* and *Newton's method*.

221.11 A Decreasing Bounded Sequence Converges!

Suppose $\{x_i\}$ is a bounded decreasing sequence, that is $x_1 \geq x_2 \geq \cdots \geq x_n \geq x_{n+1} \geq ...$, and $x_n \geq m$ for all n for some number m. Then the set of all numbers x_n is bounded below, and thus has a greatest lower bound $\bar{m}$.

We shall prove that $\lim_{n\to\infty} x_n = \bar{m}$. By the definition of greatest lower bound, for all $\epsilon > 0$ there is an x_N such that $\bar{m} \leq x_N \leq \bar{m} + \epsilon$. Since $x_n \leq x_N$ for $n \geq N$, and $x_n \geq \bar{x}$, it follows that $\bar{m} \leq x_n \leq \bar{m} + \epsilon$ for all $n \geq N$, which proves the desired result. We summarize in the following theorem which is a cornerstone of the analysis of functions of a real variables.

Theorem 221.5 *Suppose $\{x_i\}_{i=1}^{\infty}$ is a decreasing sequence that is bounded below or an increasing sequence that is bounded above. Then $\{x_i\}_{i=1}^{\infty}$ is convergent.*

Chapter 221 Problems

221.1. Find the maximum and minimum values of the function $f(x_1, x_2) = x_1^2 + 2x_2^2 - x_1$ on the unit disc $x_1^2 + x_2^2 \leq 1$.

221.2. Find the point of the plane $3x_1 + 4x_2 - x_3 = 26$ which is closest to the origin.

221.3. Find the shape of a box (with top included) which for given surface area has maximal volume.

221.4. Seek minimum and maximum values of the following functions:

(a) $f(x_1, x_2) = (1 + x_1^2 + x_2^2)^{-1}$ for $(x_1, x_2) \in \mathbb{R}^2$, (b) $f(x_1, x_2) = x_1 x_2$ for $x_1^2 + x_2^2 \leq 1$, (c) $f(x_1, x_2, x_3) = x_1 + x_2 + x_3$ for $x_1^2 + x_2^2 + x_3^2 \leq 1$.

221.5. Show that the function $x_1^4 + x_2^4 + x_3^4 - 4x_1x_2x_3$ has a minimum point at $(x_1, x_2.x_3) = (1, 1, 1)$.

221.6. Find the triangle of largest area that can be inscribed in a given circle.

221.7. Find the point on the curve $x_2 = x_1^2$ which is closest to the point $(0, 1)$.

221.8. Determine the constants a_0 and a_1 which minimize for a given function $f : [0, 1] \to \mathbb{R}$, the integral

$$\int_0^1 (f(x) - a_0 - a_1 x)^2 \, dx.$$

221.9. Find the maximum value of $x_1 + x_2 + ... + x_n$ subject to the condition $x_1^2 + x_2^2 + ... + x_n^2 \leq 1$.

221.10. A *stationary point* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is a point $x \in \mathbb{R}^n$ such that $f'(x) = 0$. Determine if any of the stationary points of the following functions is a maximum or minimum point: (a) $f(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 - x_1 - x_2 + x_3 + 1$, (b) $f(x_1, x_2, x_3) = x_1^2 + x_2^2 + 2x_3^2 + 4x_1 - x_2 + x_3 + 5$, (c) $f(x_1, x_2, x_3) = \cos(x_1) + \cos(x_2) + \cos(x_3)$.

222

The Divergence, Rotation and Laplacian

.. Stokes was a very important formative influence on subsequent generations of Cambridge men, including Maxwell. With Green, who in turn had influenced him, Stokes followed the work of the French, especially Lagrange, Laplace, Fourier, Poisson and Cauchy. This is seen most clearly in his theoretical studies in optics and hydrodynamics; but it should also be noted that Stokes, even as an undergraduate, experimented incessantly. Yet his interests and investigations extended beyond physics, for his knowledge of chemistry and botany was extensive, and often his work in optics drew him into those fields. (Parkinson)

Appointed professor of mathematics at the Ecole Polytechnique in 1809 Ampére held posts there until 1828. Ampére and Cauchy shared the teaching of analysis and mechanics and there was a great contrast between the two with Cauchy's rigorous analysis teaching leading to great mathematical progress but found extremely difficult by students who greatly preferred Ampére's more conventional approach to analysis and mechanics. (O'Connor and Robertson)

222.1 Introduction

We saw previously that the gradient of a function of several variables is a practically useful *differential operator*. In this chapter, we introduce some other useful operators, including the *divergence*, *rotation* and the *Laplacian*, together with the gradient play a fundamental role in mathematical

modeling in science and engineering. We first define the operators in $\mathbb{R}^2$ and then in $\mathbb{R}^3$, noting that the rotation takes somewhat different forms in $\mathbb{R}^2$ and $\mathbb{R}^3$.

FIGURE 222.1. Napoleon to Laplace (1749-1827): "You have written this huge book on the system of the world without once mentioning the Author of the Universe". Laplace to Napoleon: "Sire, I had no need of this hypothesis"

222.2 The Case of $\mathbb{R}^2$

We recall that the *gradient* of a function $u : \mathbb{R}^2 \to \mathbb{R}$, denoted grad u or ∇u, is the vector-valued function formed by the first order partial derivatives of u, i.e.

$$\text{grad } u = \nabla u = \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2} \right).$$

The *divergence* of a vector function $u = (u_1, u_2) : \mathbb{R}^2 \to \mathbb{R}^2$, denoted $\text{div}\, u$ or $\nabla \cdot u$, is the scalar function defined by

$$\text{div}\, u = \nabla \cdot u = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2}.$$

Formally, we have

$$\nabla \cdot u = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}) \cdot (u_1, u_2)$$

where we may think of $(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2})$ "as a vector" and let the dot indicate a "scalar product". This idea applies to all the formulas below involving ∇ combined with the operators $\cdot$ and $\times$.

The *rotation* of a vector function $u : \mathbb{R}^2 \to \mathbb{R}^2$, denoted by rot u or $\nabla \times u$, is the scalar function

$$\text{rot}\, u = \nabla \times u = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}) \times (u_1, u_2).$$

If $u : \mathbb{R}^2 \to \mathbb{R}$ is a scalar function, then $\text{rot}\, u = \nabla \times u$ is defined as the vector function

$$\text{rot}\, u = \nabla \times u = \left(\frac{\partial u}{\partial x_2}, -\frac{\partial u}{\partial x_1} \right).$$

The different appearances of $\text{rot}\, u = \nabla \times u$, with u a scalar or $u = (u_1, u_2)$ a vector function will be explained when we pass to $\mathbb{R}^3$ below. For now, it may be helpful to recall the different appearances of $a \times b$ with $a, b \in \mathbb{R}^2$ or $a, b \in \mathbb{R}^3$.

The following identities follow directly from the definitions for any function u:

$$\begin{aligned} \nabla \cdot (\nabla \times u) &= \text{div}\ (\text{rot}\ u) = 0, \qquad (u : \mathbb{R}^2 \to \mathbb{R}^2) \\ \nabla \times (\nabla u) &= \text{rot}\ (\text{grad}\ u) = 0, \qquad (u : \mathbb{R}^2 \to \mathbb{R}). \end{aligned} \tag{222.1}$$

Finally, the *Laplacian* Δu of a function $u : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$\Delta u = \nabla \cdot (\nabla u) = \text{div}\ (\text{grad}\ u) = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2},$$

where $\frac{\partial^2 u}{\partial x_i^2} = \frac{\partial}{\partial x_i}(\frac{\partial u}{\partial x_i})$.

222.3 The Laplacian in Polar Coordinates

In polar coordinates $x = (x_1, x_2) = (r\cos(\theta), r\sin(\theta))$ with $r \geq 0$ and $0 \leq \theta < 2\pi$, the Laplacian takes the form

$$\Delta u = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r} \right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}. \tag{222.2}$$

This follows by a routine computation using that the Jacobian of the mapping $x = (r\cos(\theta), r\sin(\theta))$, in the notation (217.9) is given by

$$\frac{d(x_1, x_2)}{d(r, \theta)} = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix},$$

so that

$$\frac{d(r, \theta)}{d(x_1, x_2)} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta)/r & \cos(\theta)/r \end{pmatrix},$$

and thus by the Chain rule

$$\frac{\partial}{\partial x_1} = \cos(\theta)\frac{\partial}{\partial r} - \frac{\sin(\theta)}{r}\frac{\partial}{\partial \theta} \quad \text{and} \quad \frac{\partial}{\partial x_2} = \sin(\theta)\frac{\partial}{\partial r} + \frac{\cos(\theta)}{r}\frac{\partial}{\partial \theta}.$$

222.4 Some Basic Examples

The function $u : \mathbb{R}^2 \to \mathbb{R}^2$ given by $u(x) = \frac{1}{2}(x_1, x_2)$, satisfies

$$\nabla \cdot u(x) = 1.$$

The function $v : \mathbb{R}^2 \to \mathbb{R}^2$ given by $v(x) = \frac{1}{2}(-x_2, x_1)$, satisfies

$$\nabla \times v(x) = 1.$$

The function $w : \mathbb{R}^2 \to \mathbb{R}$ given by $w(x) = \frac{1}{4}(x_1^2 + x_2^2)$, satisfies

$$\Delta w = 1.$$

We plot these basic examples in Fig. 222.2

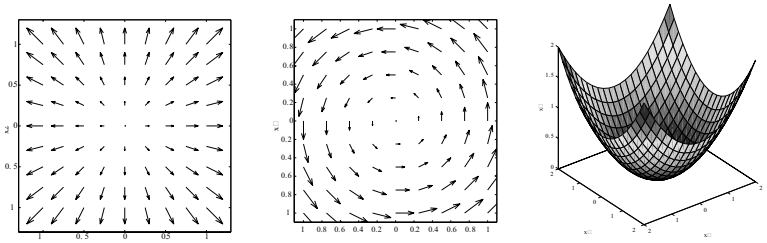


FIGURE 222.2. Basic examples satisfying $\nabla \cdot u = 1$, $\nabla \times v = 1$ and $\Delta w = 1$, respectively.

We see that $u(x)$ "explodes", $v(x)$ "rotates" and $w(x)$ is a "hump".

222.5 The Laplacian Under Rigid Coordinate Transformations

It follows from the form of the Laplacian in polar coordinates, that the Laplacian is invariant under rotations and translations in $\mathbb{R}^2$, i.e. so-called *rigid transformations* of the form

$$\begin{aligned}\tilde{x}_1 &= \cos(\alpha)x_1 + \sin(\alpha)x_2 + a_1\\ \tilde{x}_2 &= -\sin(\alpha)x_1 + \cos(\alpha)x_2 + a_2,\end{aligned}$$

where (x_1, x_2) are the old coordinates and $(\tilde{x}_1, \tilde{x}_2)$ the new ones. In other words, the Laplacian takes exactly the same form in the two coordinate systems:

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = \frac{\partial^2 u}{\partial \tilde{x}_1^2} + \frac{\partial^2 u}{\partial \tilde{x}_2^2}.$$

This fact is reflected in the observation that the Laplace operator typically occurs in *isotropic* models that have the same properties in all directions.

222.6 The Case of $\mathbb{R}^3$

The *gradient* of a function $u : \mathbb{R}^3 \to \mathbb{R}$, denoted grad u or ∇u, is the vector-valued function formed by the set of first order partial derivatives of u, i.e.

$$\text{grad } u = \nabla u = \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \frac{\partial u}{\partial x_3} \right).$$

For a vector function $u : \mathbb{R}^3 \to \mathbb{R}^3$, the divergence div u is a scalar function defined by

$$\text{div } u = \sum_{i=1}^{3} \frac{\partial u_i}{\partial x_i},$$

and rot u is the vector function

$$\text{rot } u = \nabla \times u = \left(\frac{\partial u_3}{\partial x_2} - \frac{\partial u_2}{\partial x_3}, \frac{\partial u_1}{\partial x_3} - \frac{\partial u_3}{\partial x_1}, \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} \right).$$

We now explain the relation of the operator of rotation $\nabla\times$ in $\mathbb{R}^3$ to the operator of rotation $\nabla\times$ in $\mathbb{R}^2$ introduced above. Consider first a function $u : \mathbb{R}^3 \to \mathbb{R}^3$ of the form $u = (u_1, u_2, 0)$ with u_1 and u_2 being independent of x_3 so that effectively $u_i : \mathbb{R}^2 \to \mathbb{R}$ with $u_i = u_i(x_1, x_2)$ for $i = 1, 2.$. We have

$$\nabla \times u = (0, 0, \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}) = (0, 0, \nabla \times (u_1, u_2)).$$

Secondly, if $u : \mathbb{R}^3 \to \mathbb{R}^3$ has the form $u = (0, 0, u_3)$ with u_3 independent of x_3, so that effectively $u_3 : \mathbb{R}^2 \to \mathbb{R}$, then

$$\nabla \times u = (\frac{\partial u_3}{\partial x_2}, -\frac{\partial u_3}{\partial x_1}, 0) = (\nabla \times u_3, 0).$$

We conclude that $\nabla \times u$ for $u : \mathbb{R}^2 \to \mathbb{R}$ and $\nabla \times u$ for $u : \mathbb{R}^2 \to \mathbb{R}^2$, may be viewed as special cases of $\nabla \times u$ for $u : \mathbb{R}^3 \to \mathbb{R}^3$.

The *Laplacian* Δu of a function $u : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$\Delta u = \nabla \cdot (\nabla u) = \text{div (grad } u) = \sum_{i=1}^{3} \frac{\partial^2 u}{\partial x_i^2}.$$

By direct computation we verify the following identities:

$$\begin{aligned} \nabla \cdot (\nabla \times u) &= 0, \\ \nabla \times (\nabla u) &= 0, \\ \nabla \times (\nabla \times u) &= -\Delta u + \nabla(\nabla \cdot u). \end{aligned} \tag{222.3}$$

222.7 Basic Examples, Again

The function $u : \mathbb{R}^3 \to \mathbb{R}^3$ given by $u(x) = \frac{1}{3}x$, satisfies

$$\nabla \cdot u(x) = 1.$$

The function $v : \mathbb{R}^3 \to \mathbb{R}^3$ given by $v(x) = \frac{1}{2}(-x_2, x_1, 0)$, satisfies

$$\nabla \times v(x) = (0, 0, 1).$$

The function $w : \mathbb{R}^3 \to \mathbb{R}$ given by $w(x) = \frac{1}{6}\|x\|^2$, satisfies

$$\Delta w = 1.$$

We plot these basic examples in Fig. 222.3

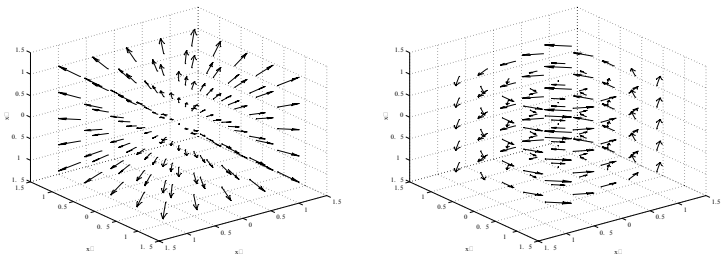


FIGURE 222.3. Basic examples in $\mathbb{R}^3$ satisfying $\nabla \cdot u = 1$, $\nabla \times v = 1$.

We see again that $u(x)$ "explodes", $v(x)$ "rotates" along the x_3 axis while the "hump" $w(x)$ is difficult to visualize.

222.8 The Laplacian in Spherical Coordinates

In *spherical coordinates*.

$$x = (x_1, x_2, x_3) = (r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta), r\cos(\varphi)),$$

where $r \geq 0$, $0 \leq \theta < 2\pi$ and $0 \leq \varphi < \pi$, the Laplacian is given by

$$\Delta u = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{r^2\sin(\theta)}\frac{\partial}{\partial \theta}\left(\sin(\theta)\frac{\partial u}{\partial \theta}\right) + \frac{1}{r^2\sin^2(\theta)}\frac{\partial^2 u}{\partial \varphi^2}. \quad (222.4)$$

The Laplacian is invariant under orthogonal coordinate transformations in $\mathbb{R}^3$.

EXAMPLE 222.1. Consider the velocity field generated by rotation around a vector $\omega \in \mathbb{R}^3$ with angular speed $\|\omega\|$, that is the vector field

$$v(x) = \omega \times x.$$

We compute

$$\begin{aligned}\nabla \times v(x) &= \nabla \times (\omega_2 x_3 - \omega_3 x_2, \omega_3 x_1 - \omega_1 x_3, \omega_1 x_2 - \omega_2 x_1) \\ &= (2\omega_1, 2\omega_2, 2\omega_3) = 2\omega.\end{aligned}$$

We conclude that the rotation $\nabla \times v(x)$ of a velocity field $v(x)$ generated by a rotation according to a given vector ω is equal to 2ω. This motivates the name of the differential operator $\nabla\times$ as the "rotation".

EXAMPLE 222.2. A basic formula of electromagnetics expressing Ampère's law states that the *magnetic field* H generated by a unit electrical current flowing through the x_3-axis in the positive direction, is given by

$$H(x) = H(x_1, x_2, x_3) = \frac{1}{2\pi}\frac{(-x_2, x_1, 0)}{x_1^2 + x_2^2} \quad \text{for } x_1^2 + x_2^2 > 0. \qquad (222.5)$$

We compute

$$\nabla \times H(x) = \frac{1}{2\pi}(0, 0, \frac{\partial}{\partial x_1}\frac{x_1}{x_1^2 + x_2^2} - \frac{\partial}{\partial x_2}\frac{-x_2}{x_1^2 + x_2^2}) = 0 \quad \text{for } x_1^2 + x_2^2 > 0.$$

Thus $\nabla \times H(x) = 0$ for $x_1^2 + x_2^2 > 0$, which is just *Amperes's Law* $\nabla \times H = J$, where J is the current density, noting that $J(x) =$ for $x_1^2 + x_2^2 > 0$, i.e. outside the x_3-axis. Amperes's Law is one of *Maxwell's equations*. Below we shall show how to interpret the equation $\nabla \times H(x) = J(x)$ for $x_1^2 + x_2^2 = 0$ and motivate the factor $\frac{1}{2\pi}$ in (222.5).

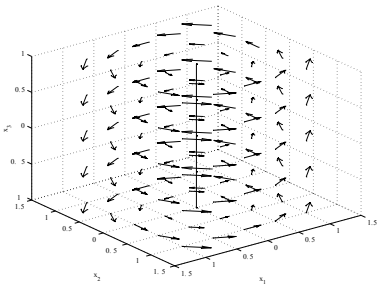


FIGURE 222.4. The magnetic field around a current through the x_3-axis

Chapter 222 Problems

222.1. Let $F = (5x_1 - 3x_1x_2 + x_3^2, \sin(x_1)\cos(x_1) + x_1, \sin(x_1)\exp(x_1x_2))$. With $x = (1, 2, 3)$, compute (a) $\nabla \cdot F$, (b) $\nabla \times F$, (c) $\nabla(\nabla \cdot F)$, (d) $\nabla \times (\nabla \times F)$.

222.2. Interpret the expression $(\nabla \times \nabla)u$ in a reasonable way and show that $(\nabla \times \nabla)u = 0$ for any u. Compare with $\nabla \times (\nabla \times u)$.

222.3. Show that for appropriate function u and v

1. $\nabla(uv) = (\nabla u)v + u(\nabla v)$,
2. $\nabla \cdot (uv) = (\nabla u) \cdot v + u(\nabla \cdot v)$,
3. $\nabla \times (uv) = (\nabla u) \times v + u(\nabla \times v)$,
4. $\nabla \cdot (u \times v) = v \cdot (\nabla \times u) - u \cdot (\nabla \times v)$,
5. $\nabla \times (u \times v) = (v \cdot \nabla)u - (\nabla \cdot u)v - (u \cdot \nabla)v + (\nabla \cdot v)u$,
6. $\nabla(u \cdot v) = (u \cdot \nabla)v + (v \cdot \nabla)u + u \times (\nabla \times v) + v \times (\nabla \times u)$.

222.4. Compute $\nabla(r \cdot F(r))$ where $r = \|x\|$.

222.5. Prove that the velocity field $v(x) = \omega \times x$, where $\omega \in \mathbb{R}^3$ is a given vector, satisfies $\nabla \cdot v(x) = 0$. Interpret the result in fluid mechanical terms.

222.6. Prove directly using the Chain rule that the Laplacian in $\mathbb{R}^2$ and $\mathbb{R}^3$ is invariant under rigid coordinate transformations.

222.7. Prove (222.3), (222.2) and (222.4).

222.8. Show that if $u : \mathbb{R}^2 \to \mathbb{R}$, then $\nabla \times (\nabla \times u) = \text{rot (rot u)} = -\Delta u$.

222.9. Show that the function $u : \mathbb{R}^2 \to \mathbb{R}$ given by $u(x) = c_1 \log(\|x\|) + c_2$ with c_1 and c_2 constants, is a solution of the Laplace equation $\Delta u(x) = 0$ in $\mathbb{R}^2$ for $x \neq 0$.

222.10. Prove that the function $u : \mathbb{R}^3 \to \mathbb{R}$ given by $u(x) = c_1\|x\|^{-1} + c_2$, with c_1 and c_2 constants, is a solution of Laplace's equation $\Delta u(x) = 0$ in $\mathbb{R}^3$ for $x \neq 0$.

222.11. Show that the divergence is invariant under rigid coordinate transformations. Does the rotation have the same property?

> All the effects of Nature are only the mathematical consequences of a small number of immutable laws. (Laplace)

223
Curve Integrals

> We can scarcely believe that Ampère really discovered the law of action by means of the experiments which he describes. We are led to suspect, what, indeed, he tells us himself, that he discovered the law by some process which he has not shown us, and that when he had afterwards built up a perfect demonstration he removed all traces of the scaffolding by which he had raised it. (Maxwell about Ampères *Memoir on the Mathematical Theory of Electrodynamic Phenomena, Uniquely Deduced from Experience*)

223.1 Introduction

In this chapter we introduce the concept of an *integral over a curve* or *curve integral*, and develop some applications including *arc length*, *work* and *line integrals*. We start with plane curves parameterized by functions $s : I \to \mathbb{R}^2$, where $I = [a, b]$ is an interval of the real line $\mathbb{R}$. We then generalize to curves in $\mathbb{R}^n$ parameterized by functions $s : I \to \mathbb{R}^n$ with $n \geq 2$.

223.2 The Length of a Curve in $\mathbb{R}^2$

Let Γ be a curve in $\mathbb{R}^2$ given by the function $s : I \to \mathbb{R}^2$, where $I = [a, b]$ is an interval of $\mathbb{R}$, that is, $\Gamma = \{s(t) \in \mathbb{R}^2 : t \in I\}$, or $\Gamma = s(I)$, see Fig. 223.1.

We now try to determine the *length* of Γ. We shall see that this leads to the introduction of the notion of an integral over a curve or a curve integral.

To define the length of a curve, we view the curve Γ as the being made up of little pieces of Γ. If the little pieces are sufficiently small, we can get away with approximating them by straight segments, and the length of a straight piece of curve is easy to compute. To find the total length of Γ, we will sum the lengths of all the little pieces forming Γ. We will find the integral is useful for this purpose.

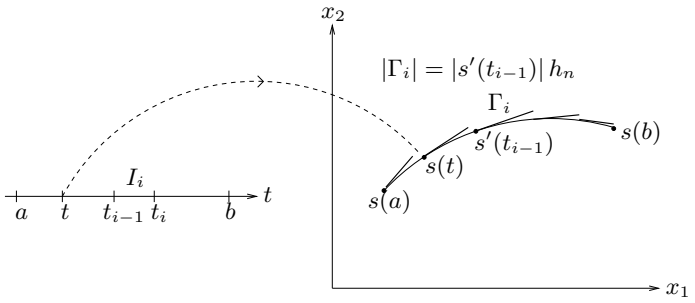


FIGURE 223.1. The total length of curve is the sum of the lengths of little pieces of the curve.

Let $a = t_0 < t_1 < ... < t_n = b$ be a subdivison of I into intervals $I_i = (t_{i-1}, t_i]$. Consider the following linear approximation of the mapping $s(t)$ restricted to the subinterval I_i, see Fig. 223.1,

$$\bar{s}(t) = s(t_{i-1}) + (t - t_{i-1})s'(t_{i-1}).$$

The mapping $\bar{s}$ maps I_i onto the line segment Γ_i of length

$$\|s'(t_{i-1})\|(t_i - t_{i-1}),$$

and it is thus natural to use

$$L_n(\Gamma) = \sum_{i=1}^{n} \|s'(t_{i-1})\|(t_i - t_{i-1})$$

as an approximation of the length of Γ. Assuming that $\|s'(t)\|$ is Lipschitz continuous on I and assuming that $\max_i(t_i - t_{i-1})$ tends to zero as n tends to infinity, we can use the usual arguments to show that $\{L_n(\Gamma)\}_{n=1}^{\infty}$ is a Cauchy sequence and thus converges to a limit, which we denote by $L(\Gamma)$. We define this limit to be the *length* of Γ:

$$L(\Gamma) = \int_I \|s'(t)\|\, dt. \tag{223.1}$$

This formula expresses the length of a curve $\Gamma = s(I)$ as an integral over the parameter domain I of Γ with the modulus $\|s'(t)\|$ of the derivative

of the representing function $s : I \to \mathbb{R}^2$ as a weight. Formally, we have $ds = \|s'(t)\|dt$, where ds represents the increase of the length of the curve corresponding to an increase dt of the parameter t; the function $\|s'(t)\|$ gives the local "change of scale" between the "element of curve length" ds; and the "parameter element" dt, see Fig. 223.1. We are thus led to write

$$L(\Gamma) = \int_\Gamma ds = \int_I \|s'(t)\|\, dt.$$

We will return to this notation in the next section.

EXAMPLE 223.1. We compute the length of the circumference Γ of a circle of radius 1 centered at the origin. The curve Γ is given by the function $s : [0, 2\pi) \to \mathbb{R}^2$ with $s(t) = (\cos(t), \sin(t))$ and $0 \le t < 2\pi$. We have $s'(t) = (-\sin(t), \cos(t))$ and $\|s'(t)\| = 1$, and thus

$$L(\Gamma) = \int_0^{2\pi} \|s'(t)\|\, dt = \int_0^{2\pi} dt = 2\pi.$$

We conclude that the the length of the circumference of a circle of radius 1 is equal to 2π (no big surprise). We check the result using a different parametrization. The upper semi-circle Γ_+ of Γ can be parameterized by $s : [-1, 1] \to \mathbb{R}^2$ given by $s(t) = (t, \sqrt{1-t^2})$ with $-1 \le t \le 1$. We have

$$s'(t) = (1, -\frac{t}{\sqrt{1-t^2}}), \quad \|s'(t)\| = \frac{1}{\sqrt{1-t^2}},$$

and thus

$$\begin{aligned} L(\Gamma) &= 2L(\Gamma_+) \\ &= \int_{-1}^1 \frac{1}{\sqrt{1-t^2}}\, dt = 2\big[\arcsin(t)\big]_{-1}^1 = 2(\frac{\pi}{2} - (-\frac{\pi}{2})) = 2\pi. \end{aligned}$$

223.3 Curve Integral

Let $\Gamma = s(I)$ be a curve in $\mathbb{R}^2$ given by the function $s : I \to \mathbb{R}^2$, where $I = [a, b]$ is an interval of $\mathbb{R}$, and let $u : \Gamma \to \mathbb{R}$ be a function defined on Γ. We assume that the tangent $s' : I \to \mathbb{R}^2$ and the function $u : \Gamma \to \mathbb{R}$ are both Lipschitz continuous, which guarantees that $\|s'(t)\|$ and $u(s(t))$ are both Lipschitz continuous on I. We define the *integral of* u *over* Γ by

$$\int_\Gamma u\, ds \equiv \int_\Gamma u(x)\, ds(x) \equiv \int_a^b u(s(t))\|s'(t)\|\, dt.$$

Formally, we have $ds = ds(x) = \|s'(t)\|\, dt$, where $x = s(t)$.

EXAMPLE 223.2. If Γ is an interval $[a,b]$ on the x_1-axis given by $s(t)=(t,0)$, $a\le t\le b$, then $s'(t)=(1,0)$, $\|s'(t)\|=1$, and

$$\int_\Gamma u\,ds=\int_a^b u(x_1,0)\,dx_1=\int_a^b u(t,0)\,dt.$$

EXAMPLE 223.3. Let $\Gamma=s(I)$ be the semicircle given by $s(t)=(\cos(t),\sin(t))$, $0\le t\le\pi$, and $u(x)=u(x_1,x_2)=x_1^2$. Using $\|s'(t)\|=1$, we get

$$\int_\Gamma u\,ds=\int_0^\pi \cos^2(t)\,dt=\frac{1}{2}\int_0^\pi(1-\cos(2t))\,dt=\frac{\pi}{2}.$$

223.4 Reparameterization

An important observation is that the value of a curve integral is independent of the parameterization of the curve. To see this, consider two different parameterizations $s:[a,b]\to\Gamma$ and $\sigma:[c,d]\to\Gamma$ of a curve Γ in $\mathbb{R}^2$. Associate to each $\tau\in[c,d]$ the unique value $t\in[a,b]$ such that $s(t)=\sigma(\tau)$, which defines $t=t(\tau)$) as a function of τ (assuming that the curve does not cross itself), so that $\sigma(\tau)=s(t(\tau))$, see Fig. 223.2.

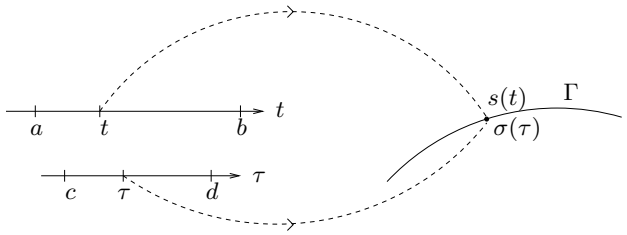


FIGURE 223.2. Reparametrization of a curve.

We now use the formula for change of integration variables and the fact that by the Chain rule

$$\sigma'(\tau)=\frac{d\sigma}{d\tau}=\frac{ds}{dt}\frac{dt}{d\tau}=s'(t)\frac{dt}{d\tau},$$

to see that, assuming $\frac{dt}{d\tau}\ge 0$,

$$\int_a^b u(s(t))\|s'(t)\|\,dt=\int_c^d u(s(t(\tau)))\|s'(t(\tau))\|\frac{dt}{d\tau}\,d\tau$$
$$=\int_c^d u(\sigma(\tau))\|\sigma'(\tau)\|\,d\tau.$$

This shows that the curve integral

$$\int_\Gamma u\,ds = \int_\Gamma u\,d\sigma$$

is independent of the parametrization $s : [a, b] \to \Gamma$ or $\sigma : [c, d] \to \Gamma$ of Γ.

EXAMPLE 223.4. We reparameterize the semicircle Γ in the previous example by $s(t) = (t, \sqrt{1-t^2})$ with $-1 \le t \le 1$ and get with $u(x) = x_1^2$, integrating by parts

$$\begin{aligned}\int_\Gamma u\,ds &= \int_{-1}^1 t\frac{t}{\sqrt{1-t^2}}\,dt = \left[-t\sqrt{1-t^2}\right]_{-1}^1 + \int_{-1}^1 \sqrt{1-t^2}\,dt \\ &= \int_{-\pi}^0 \sqrt{1-\cos^2(\theta)}(-\sin(\theta))\,d\theta = \int_0^\pi \sin^2(\theta)\,d\theta = \frac{\pi}{2}.\end{aligned}$$

223.5 Work and Line Integrals

Let $F : \mathbb{R}^2 \to \mathbb{R}^2$ be a vector function representing a variable force, or a *force field*, defined in $\mathbb{R}^2$, and let Γ be a curve in $\mathbb{R}^2$ given by $s : [a, b] \to \mathbb{R}^2$ starting at $A = s(a)$ and ending at $B = s(b)$. Consider a particle acted upon by the force F moving along Γ from A to B, see Fig. 223.3. The projection $F_s(s(t))$ of the force $F(s(t))$ on the direction $s'(t)$ of the tangent to $s(t)$ is equal to

$$F_s(s(t)) = F(s(t)) \cdot s'(t)\frac{1}{\|s'(t)\|}. \tag{223.2}$$

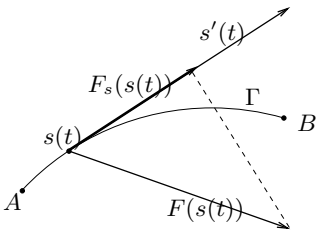


FIGURE 223.3. Force field F and curve Γ, and projection of F onto $s'(t)$.

Using the idea that "the work is equal to the projection of the force in the direction of the displacement $\times$ displacement", the *work* performed by the force $F(s(t))$ as the particle moves from $s(t_{i-1})$ to $s(t_i)$ is

$$\begin{aligned}&F(s(t_i)) \cdot s'(t_i)\frac{1}{\|s'(t_i)\|}\|s(t_i) - s(t_{i-1})\| \\ &\approx F(s(t_i)) \cdot s'(t_i)\frac{1}{\|s'(t_i)\|}\|s'(t_i)\|(t_i) - t_{i-1}) = F(s(t_i)) \cdot s'(t)(t_i - t_{i-1}).\end{aligned}$$

As above $a = t_0 < t_1 < ...t_{i-1} < t_i < ... < t_n = b$ is an increasing sequence of discrete time levels, where we think of the time steps $t_i - t_{i-1}$ as tending to zero. We are now led to define the *total work* $W(F,\Gamma)$ as the particle moves from $A = s(a)$ to $B = s(b)$ along Γ, as

$$W(F,\Gamma) = \int_a^b F(s(t)) \cdot s'(t)\, dt.$$

Setting $ds = s'(t)\, dt$, we also write

$$\int_\Gamma F \cdot ds = \int_a^b F(t) \cdot s'(t)\, dt,$$

which we call a *line integral*. To sum up, we have

$$\begin{aligned} W(F,\Gamma) = \int_\Gamma F \cdot ds &= \int_a^b F(t) \cdot s'(t)\, dt \\ &= \int_a^b (F_1(t)s_1'(t) + F_2(t)s_2'(t))\, dt. \end{aligned} \tag{223.3}$$

Alternatively, we can write

$$W(F,\Gamma) = \int_\Gamma F_s\, ds = \int_a^b F_s \|s'(t)\|\, dt = \int_\Gamma F \cdot ds, \tag{223.4}$$

with F_s being the projection of F onto $s'(t)$ according to (223.2).

EXAMPLE 223.5. Assume that $F(x) = (x_2, -x_1)$ and let Γ be given by $s(t) = (\cos(t), \sin(t))$, $0 \le t < 2\pi$. We have

$$\begin{aligned} W(F,\Gamma) = \int_\Gamma F \cdot ds &= \int_0^{2\pi} (\sin(t), -\cos(t)) \cdot (-\sin(t), \cos(t))\, dt \\ &= -\int_0^{2\pi} dt = -2\pi. \end{aligned}$$

223.6 Work and Gradient Fields

There is an important special case. If $F = \nabla\varphi$, that is the force field F is the *gradient field* of a *potential* $\varphi(x)$, then the Chain rule implies

$$\begin{aligned} W(F,\Gamma) = \int_\Gamma F \cdot ds &= \int_a^b \nabla\varphi(s(t)) \cdot s'(t)\, dt \\ &= \int_a^b \frac{d}{dt}\varphi(s(t))\, dt = \varphi(B) - \varphi(A). \end{aligned}$$

We conclude that if the force field F is the gradient field $F = \nabla\varphi$ of a potential $\varphi(x)$, then the work performed by F along a curve Γ from A to B is equal to the difference $\varphi(B) - \varphi(A)$ of the values of the potential φ at the end point B and the starting point A. In other words, the work is independent of the curve from A to B. In particular, if the curve is *closed* so that $B = s(b) = s(a) = A$, then the work is zero.

Below we consider the problem of finding conditions guaranteeing that a given force $F(x)$ is the gradient of a potential so that $F(x) = \nabla\varphi(x)$ for some scalar function $\varphi(x)$.

EXAMPLE 223.6. As a basic application, we consider the attractive *gravitational force* $F(x) = \nabla\varphi(x)$ with $\varphi(x) = 1/\|x\|$ being the *Newtonian potential*, corresponding to a unit mass at the the origin, that is

$$F(x) = -\frac{1}{\|x\|^2}\frac{x}{\|x\|},$$

with normalization of the gravitational constant to one. We note that $F(x)$ is directed towards the origin and obeys the inverse square law: $\|F(x)\| = \|x\|^{-2}$. We have

$$W(F,\Gamma) = \frac{1}{\|B\|} - \frac{1}{\|A\|},$$

which corresponds to the work performed as a unit mass moves in the gravitational field from a distance $\|A\|$ to the distance $\|B\|$ from the origin. In particular, if $\|A\| = \infty$, then $W(F,\Gamma) = 1/\|B\|$. We conclude that the work required to "lift" a particle of unit mass from a distance r of an attracting gravitational field of unit strength at the origin to an infinite distance is equal to $1/r$.

223.7 Using the Arclength as a Parameter

Note that if $u(x) = 1$ for all $x \in \Gamma$, then

$$\int_\Gamma ds = \int_\Gamma 1\, ds = \int_\Gamma u(x)\, ds(x) = \int_a^b \|s'(t)\|\, dt$$

is the length of the curve $\Gamma = s(I)$ with $I = [a,b]$. In particular,

$$\sigma(\bar{t}) = \int_a^{\bar{t}} \|s'(t)\|\, dt$$

is the *arclength* of the part of the curve from $s(a)$ to $s(\bar{t})$. The Fundamental Theorem of Calculus implies

$$\sigma'(\bar{t}) = \|s'(\bar{t})\|. \tag{223.5}$$

We may now choose the arclength $\sigma = \sigma(t)$ as the parameter instead of t since to each t, there is a unique arclength $\sigma(t)$ and vice versa. This gives a reparameterization of $s(t) = \bar{s}(\sigma)$ with

$$\|\bar{s}'(\sigma)\| = \|\frac{ds}{dt}\|\|\frac{dt}{d\sigma}| = \|s'(t)\|\frac{1}{|\sigma'(t)|} = \frac{\|s'(t)\|}{\|s'(t)\|} = 1.$$

We conclude that if the arclength σ to used to parameterize the curve $s : I \to \mathbb{R}^2$, then $\|s'(\sigma)\| = 1$ and, see Fig. 223.4,

$$L(\Gamma) = \int_\Gamma ds = \int_0^{L(\Gamma)} d\sigma.$$

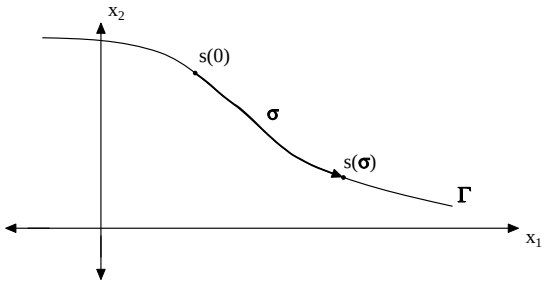


FIGURE 223.4. A curve Γ parameterized by arclength σ.

223.8 The Curvature of a Plane Curve

The *curvature* of a curve $s : [a, b] \to \mathbb{R}^2$ measures of how quickly the curve bends as we move along the curve. It is defined by

$$\kappa = \frac{d\theta}{d\sigma},$$

where θ is the polar angle of the tangent vector $s' = (s_1', s_2')$ defined by $\theta(t)) = \tan^{-1}(s_2'/s_1')$ and σ is arclength. In the case of a straight line, the polar angle $\theta(t)$ is constant and the curvature is zero, see Fig. 223.5.

The arc length $\sigma(t)$ satisfies, recalling (223.5), $\frac{d\sigma}{dt} = |s'|$, and thus $\frac{dt}{d\sigma} = |s'|^{-1}$. The chain rule implies

$$\kappa(t) = \frac{d\theta}{dt}\frac{dt}{d\sigma} = \frac{\theta'(t)}{\|s'(t)\|}.$$

Computing $\theta'(t)$, we find that

$$\kappa(t) = \frac{s_1'(t)s_2''(t) - s_1''(t)s_2'(t)}{\left(s_1'(t)^2 + s_2'(t)^2\right)^{3/2}}.$$

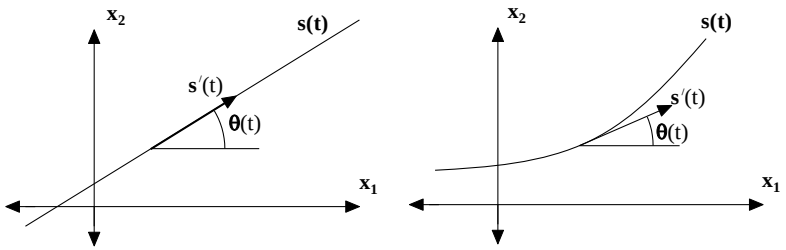


FIGURE 223.5. The polar angle θ of the tangent vector of a straight line is constant as shown on the right. The tangent vector of a curve that bends, like the example on the left, has a different polar angle at each point.

In particular if the curve is parameterized by $s(x_1) = (x_1, f(x_1))$, where $f : \mathbb{R} \to \mathbb{R}$ has two continuous derivatives, then the curvature at the point $(x_1, f(x_1))$, is given by

$$\kappa(x_1) = \frac{f''(x_1)}{\left(1 + (f'(x_1))^2\right)^{3/2}}.$$

We define the *circle of curvature* at a point $P = s(t)$ on a curve $s : [a, b] \to \mathbb{R}^2$, as the circle of radius $|\kappa|^{-1}(t)$ (assuming $\kappa \neq 0$) that shares the same tangent line as Γ at P and points to the left of T if $\kappa > 0$ and to the right if $\kappa < 0$, see Fig. 223.6. The *radius of curvature* at P is $|\kappa|^{-1}(t)$.

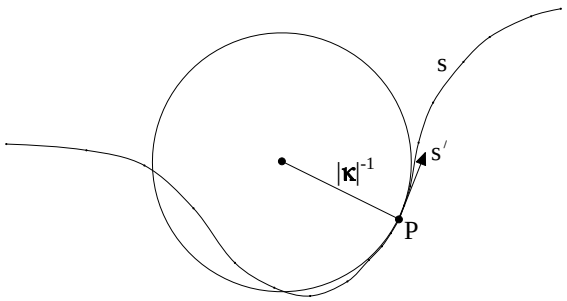


FIGURE 223.6. The circle of curvature of Γ at P.

223.9 Extension to Curves in $\mathbb{R}^n$

The definitions of integrals over curves and line integrals directly extend to curves in $\mathbb{R}^n$ represented by $s : [a, b] \to \mathbb{R}^n$ with $n \geq 2$.

EXAMPLE 223.7. Consider the circular helix Γ in $\mathbb{R}^3$ given by $s(t) = (\cos(t), \sin(t), t)$, $0 \le t \le 20\pi$, and let $u(x) = x_3^2$. We have since $s'(t) = (-\sin(t), \cos(t), 1)$ and thus $\|s'(t)\| = \sqrt{2}$,

$$\int_\Gamma u\, ds = \int_0^{20\pi} t^2\sqrt{2}\, dt = \frac{\sqrt{2}}{3}(20\pi)^3.$$

Chapter 223 Problems

223.1. (a) Compute the length of (a) the *catenary* (hanging chain curve) given by $s(t) = (t, \cosh(t))$ with $-1 \le t \le 1$, (b) the *circular helix* $s(t) = (\cos(t), \sin(t), t)$ with $0 \le t \le 4\pi$, (c) the *cycloid* $s(t) = (t - \sin(t), 1 - \cos(t))$ with $0 \le t \le 2\pi$, (d) the *semi-cubical parabola* $s(t) = (t^3, t^2)$ with $0 \le t \le 2$, (e) the *four-cusped hypocycloid or astroid* $s(t) = (\cos^3(t), \sin^3(t))$.

223.2. Let Γ be the circular helix $s(t) = (\cos(t), \sin(t), t)$ with $t \in [0, 2\pi)$. Compute the value of the curve integral $\int_\Gamma u\ ds$ for (a) $u(x) = 1$, (b) $u(x) = x_3$, (c) $u(x) = x_1x_2x_3$.

223.3. Compute the curve integral $\int_\Gamma x_1x_2\ ds$, where (a) Γ is the part of the unit circle in the x_1x_2-plane from $(1, 0, 0)$ to $(0, 1, 0)$, (b) Γ is the part of the unit square in the x_1x_2-plane from $(1, 0, 0)$ to $(0, 1, 0)$. (c) Γ is the shortest path from $(1, 0, 0)$ to $(0, 1, 0)$.

223.4. (a) Compute the line integral $\int_\Gamma x \cdot ds$ where Γ is the unit circle in the x_1x_2-plane. (b) Try other choices of closed curves Γ and evaluate the integral.

223.5. Compute the line integral $\int_\Gamma F \cdot ds$ with Γ the unit circle in the x_1x_2 plane and (a) $F(x) = \frac{(x_1,x_2)}{|x|^2}$, (b) $F(x) = \frac{(-x_2,x_1)}{|x|^2}$. Does the result depend on whether you integrate around the unit circle clockwise or counter-clockwise?

223.6. A particle is moved counter-clockwise around the square $0 \le x_1, x_2 \le 1$, $x_3 = 0$ under the action of the force field $f(x) = ((x_1 - x_2)^2, 2x_2 + x_1^2, x_1)$. Compute the work done.

223.7. Let $f(x) = (2x_1 + x_2, 3x_1 - 2x_2)$. Compute $\int_\Gamma f \cdot ds$ with Γ given by (a) the straight line from $(0, 0)$ to $(1, 1)$, (b) the parabola $x_2 = x_1^2$ from $(0, 0)$ to $(1, 1)$, (c) the curve $x_2 = sin(\pi x_1/2)$ from $(0, 0)$ to $(1, 1)$, (d) the curve $x_2 = x_1^n$ with $n > 0$ from $(0, 0)$ to $(1, 1)$.

223.8. Compute the integral of $u = x_1\, x_2$ over the boundary of the unit square $[0, 1] \times [0, 1]$.

223.9. Find the circle of curvature of $x_2 = x_1^2$ at $x_1 = 0$.

223.10. Find the curvature of the plane curve $\big(R\cos(\theta), R\sin(\theta)\big)$ where R is constant. Conclude that the curvature of a circle of radius R is R^{-1}.

223.11. Verify the two formulas for the curvature.

223.12. (a) Compute the curvature of the curve (x_1, x_1^2). (b) Do the same for (x_1, x_1^3), and then discuss what happens at the inflection point.

223.13. Consider a hanging chain described by a function $y(x)$ with $-1 \le x \le 1$ and $y(-1) = y(1)$. Let for $0 \le x \le 1$, $T(x)$ be the modulus of the chain force at x, and let $s(x)$ be the length of the chain from 0 to x. Derive the vertical equilibrium equation

$$y'(x) = cs(x) = c \int_0^1 \sqrt{1 + (y'(x))^2}\, dx,$$

with c a constant. Show that this equation is satisfied with $y'(x) = \sinh(\frac{x}{c})$, and conclude that $y(x) = c\cosh(\frac{x}{c})$.

223.14. Find the direction of the tangent at the point $(1, 1, 1)$ of the curve cut out on the surface $x_1^2 + x_1^2 x_2 + x_2^2 x_3 + x_3^2 = 0$. Hint: Use implicit differentiation.

223.15. Show that if a plane curve Γ is represented in polar coordinates $(\rho(\theta), \theta)$ with $\rho(theta)$ a function of θ and $a \le \theta \le b$, then $ds^2 = \rho^2\, d\theta^2 + d\rho^2$ and thus

$$L(\Gamma) = \int_a^b (\rho^2 + (\rho')^2)^{1/2}\, d\theta.$$

Compute the the length of the *cardioid* $\rho = (1 - \cos(\theta)$ with $0 \le \theta \le 2\pi$.

223.16. Compute the length of a string which is wound around a circular cylinder with a uniform pitch.

224
Double Integrals

To understand this for sense it is not required that a man should be a geometrician or a logician, but that he should be mad. ["This" is that the volume generated by revolving the region under 1/x from 1 to infinity has finite volume.] (Hobbes 1588-1679)

He was 40 years old before he looked on geometry; which happened accidentally. Being in a gentleman's library, Euclid's Elements lay open, and "twas the 47 El. libri I" [Pythagoras' Theorem]. He read the proposition . "By God", sayd he, "this is impossible:" So he reads the demonstration of it, which referred him back to such a proposition; which proposition he read. That referred him back to another, which he also read. Et sic deinceps, that at last he was demonstratively convinced of that trueth. This made him in love with geometry. (About Thomas Hobbes by John Aubrey 1626-1697)

224.1 Introduction

We have studied the integral

$$\int_0^1 f(x)\,dx,$$

where $f : [0,1] \to \mathbb{R}$ is a Lipschitz continuous function of one variable. We call this a *one-dimensional integral*. We generalize this idea to the *double integral*

$$\int_0^1 \int_0^1 f(x_1,x_2)\,dx_1\,dx_2, \tag{224.1}$$

which has *two integration variables* x_1 and x_2 that run from 0 to 1. Here $f : Q \to \mathbb{R}$ is a Lipschitz continuous function defined on the unit square $Q = [0,1] \times [0,1] = \{x = (x_1, x_2) : 0 \le x_1 \le 1, 0 \le x_2 \le 1\}$, satisfying

$$|f(x) - f(y)| \le L_f \|x - y\| \quad \text{for } x, y \in Q. \tag{224.2}$$

224.2 Double Integrals over the Unit Square

Recall that we define the one dimensional integral as

$$\int_0^1 f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{N} f(x_i^n) h_n, \tag{224.3}$$

where $0 = x_0^n < x_1^n < ... < x_N^n = 1$ is a subdivision of the interval $[0,1]$ with $x_i^n = ih_n$, $i = 1, ..., N$ and $h_n = 2^{-n}$ and $N = 2^n$.

To define the double integral, we let $0 = x_{1,0}^n < x_{1,1}^n < ... < x_{1,N}^n = 1$ and $0 = x_{2,0}^n < x_{2,1}^n < ... < x_{2,N}^n = 1$ be a subdivisions of the interval $[0,1]$ with $x_{1,i}^n = ih_n$, $i = 0, ..., N$, and $x_{2,j}^n = jh_n$, $j = 0, ..., N$, where $h_n = 2^{-n}$ and $N = 2^n$. This corresponds to a subdivision of the unit square $Q = [0,1]\times[0,1]$ into sub-squares $Q_{i,j}^n = I_i^n \times J_j^n$ of area $h_n h_n$, where $I_i^n = (x_{1,i-1}^n, x_{1,i}^n]$ $J_j^n = (x_{2,j-1}^n, x_{2,j}^n]$, where $i, j = 1, ..., N$, see Fig. 224.1.

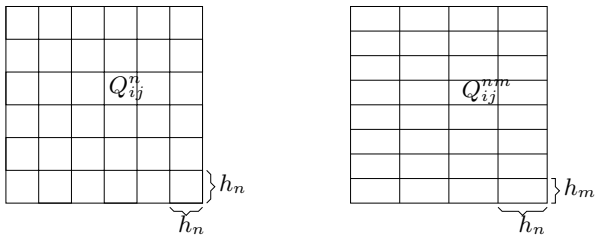


FIGURE 224.1. Partition of the unit square Q into quadratic or rectangular sub-domains Q_{ij}^n or Q_{ij}^{nm}.

We shall prove that the limit $\lim_{n\to\infty} S_n$ exists, where

$$S_n = \sum_{i=1}^{N}\sum_{j=1}^{N} f(x_{1,i}^n, x_{2,j}^n) h_n h_n \tag{224.4}$$

is a *Riemann sum* over all the sub-squares $Q_{i,j}^n$. We define

$$\int_0^1 \int_0^1 f(x_1, x_2)\,dx_1\,dx_2 = \lim_{n\to\infty} \sum_{i=1}^{N}\sum_{j=1}^{N} f(x_{1,i}^n, x_{2,j}^n) h_n h_n. \tag{224.5}$$

We begin by estimating the difference $S_n - S_{n+1}$ with the goal of proving that $\{S_n\}$ is a Cauchy sequence. Each sub-square $Q_{i,j}^n$ consists of the four sub-squares $Q_{2i,2j}^{n+1}$, $Q_{2i-1,2j}^{n+1}$ $Q_{2i,2j-1}^{n+1}$, and $Q_{2i-1,2j-1}^{n+1}$, see Fig. 224.2. We have

$$S_n - S_{n+1} = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} h_n h_n,$$

where, see Fig. 224.2,

$$a_{ij} = f(x_{1,i}^n, x_{2,j}^n) - \frac{1}{4}\big(f(x_{1,2i}^{n+1}, x_{2,2j}^{n+1}) + f(x_{1,2i-1}^{n+1}, x_{2,2j}^{n+1}) + f(x_{1,2i}^{n+1}, x_{2,2j-1}^{n+1}) + f(x_{1,2i-1}^{n+1}, x_{2,2j-1}^{n+1})\big).$$

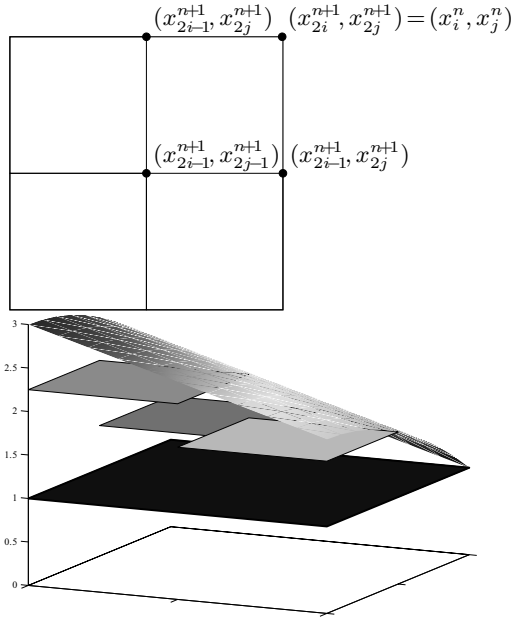


FIGURE 224.2. On the left: Q_{ij}^n and four sub-squares and quadrature point. On the right: A function $f(x_1, x_2)$ and its piecewise constant approximation on Q_{ij}^n and on the four sub-squares.

The Lipschitz continuity condition (224.2) implies

$$|a_{ij}| \le \frac{1}{4} L_f (h_{n+1} + h_{n+1} + \sqrt{2} h_{n+1}) \le L_f h_{n+1},$$

and thus

$$|S_n - S_{n+1}| \le \sum_{i=1}^{N} \sum_{j=1}^{N} |a_{ij}| h_n h_n \le L_f h_{n+1} \sum_{i=1}^{N} \sum_{j=1}^{N} h_n h_n = L_f h_{n+1}.$$

The usual arguments show that for $m > n$,

$$|S_n - S_m| \leq 2L_f h_{n+1} = L_f h_n,$$

which proves that $\{S_n\}$ is a Cauchy sequence and thus converges to a real number. We decide, following our dear friends Leibniz and Cauchy as usual, to denote this real number by

$$\int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2 = \lim_{n\to\infty} S_n = \lim_{n\to\infty} \sum_{i=1}^N \sum_{j=1}^N f(x_{1,i}^n, x_{2,j}^n) h_n h_n.$$

We shall also use the notation

$$\int_Q f(x)\, dx = \int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2.$$

We summarize as follows:

Theorem 224.1 *If $f : [0,1] \times [0,1] \to \mathbb{R}$ is Lipschitz continuous, then the limit*

$$\lim_{n\to\infty} \sum_{i=1}^N \sum_{j=1}^N f(x_{1,i}^n, x_{2,j}^n) h_n h_n,$$

exists, where $h_n = 2^{-n}$ and $N = 2^n$, $x_{1,i}^n = ih_n$, $x_{2,j}^n = jh_n$, and we define

$$\int_Q f(x)\, dx = \int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2 = \lim_{n\to\infty} \sum_{i=1}^N \sum_{j=1}^N f(x_{1,i}^n, x_{2,j}^n) h_n h_n. \tag{224.6}$$

In general, the partitions in x_1 and x_2 can be independent, leading to Riemann sums of the form

$$S_{nm} = \sum_{i=1}^N \sum_{j=1}^M f(x_{1,i}^n, x_{2,j}^m) h_n h_m, \tag{224.7}$$

where $h_n = 2^{-n}$ and $N = 2^n$, $h_m = 2^{-m}$ and $M = 2^m$. This corresponds to a subdivision of Q into sub-squares $Q_{ij}^{nm} = I_i^n \times J_j^m$. The proof above directly generalizes to prove that if $\overline{n} \geq n$ and $\overline{m} \geq m$ then

$$|S_{nm} - S_{\overline{nm}}| \leq L_f \max(h_n, h_m).$$

This proves the following generalization of the previous theorem.

Theorem 224.2 *Suppose $f : [0,1] \times [0,1] \to \mathbb{R}$ is Lipschitz continuous. Then the following limit exists*

$$\lim_{n,m\to\infty} \sum_{i=1}^{N} \sum_{j=1}^{M} f(x_{1,i}^n, x_{2,j}^m) h_n h_m,$$

where $h_n = 2^{-n}$, $N = 2^n$, $h_m = 2^{-m}$, $M = 2^m$, $x_{1,i}^n = ih_n$, $x_{2,j}^m = jh_m$, and

$$\int_Q f(x)\,dx = \int_0^1 \int_0^1 f(x_1, x_2)\,dx_1\,dx_2 = \lim_{n,m\to\infty} \sum_{i=1}^{N} \sum_{j=1}^{M} f(x_{1,i}^n, x_{2,j}^m) h_n h_m. \tag{224.8}$$

224.3 Double Integrals via One-Dimensional Integration

To compute the Riemann sum S_{nm}, we have to perform a summation over all the sub-squares Q_{ij}^{nm} covering Q. The summation may be performed in different orders, row by row, column by column, or in some other order. We thus obtain the following alternative expressions for the double integral of $f(x_1, x_2)$ over Q:

$$\begin{aligned}
\int_0^1 \int_0^1 f(x_1, x_2)\,dx_1\,dx_2 &= \lim_{n,m\to\infty} \sum_{i=1}^{N} \sum_{j=1}^{M} f(x_{1,i}^n, x_{2,j}^m) h_n h_m \\
&= \lim_{n,m\to\infty} \sum_{i=1}^{N} \Big(\sum_{j=1}^{M} f(x_{1,i}^n, x_{2,j}^m) h_m \Big) h_n \\
&= \lim_{n,m\to\infty} \sum_{j=1}^{M} \Big(\sum_{i=1}^{N} f(x_{1,i}^n, x_{2,j}^m) h_n \Big) h_m,
\end{aligned}$$

where $\sum_{i=1}^{N} \sum_{j=1}^{M}$ indicates an arbitrary order of summation, $\sum_{i=1}^{N} \big(\sum_{j=1}^{M} \big)$ summation column by column, and $\sum_{j=1}^{M} \big(\sum_{i=1}^{N} \big)$ summation row by row over the subdomains Q_{ij}^{nm} of Q in the x_1x_2-plane, see Fig. 224.3

We can also perform the limits with respect to n and m independently, and we then arrive at the formula

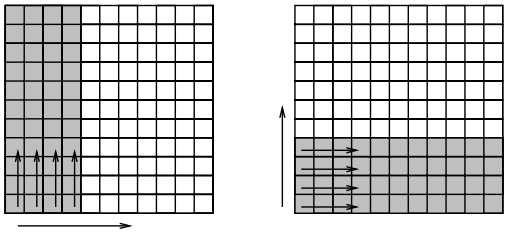


FIGURE 224.3. Different orders of summation

$$\begin{aligned}\int_0^1\int_0^1 f(x_1,x_2)\,dx_1\,dx_2 &= \lim_{n,m\to\infty}\sum_{i=1}^{N}\sum_{j=1}^{M} f(x_{1,i}^n, x_{2,j}^m)h_n h_m\\ &= \lim_{n\to\infty}\sum_{i=1}^{N}\big(\lim_{m\to\infty}\sum_{j=1}^{M} f(x_{1,i}^n, x_{2,j}^m h_m\big)h_n\\ &= \lim_{m\to\infty}\sum_{j=1}^{M}\big(\lim_{n\to\infty}\sum_{i=1}^{N} f(x_{1,i}^n, x_{2,j}^m h_n\big)h_m.\end{aligned}$$

This corresponds to the following formula:

$$\begin{aligned}\int_0^1\int_0^1 f(x_1,x_2)\,dx_1\,dx_2 &= \int_0^1\big(\int_0^1 f(x_1,x_2)\,dx_2\big)\,dx_1\\ &= \int_0^1\big(\int_0^1 f(x_1,x_2)\,dx_1\big)\,dx_2,\end{aligned}$$

or

$$\int_0^1\int_0^1 f(x_1,x_2)\,dx_1\,dx_2 = \int_0^1 g_1(x_1)\,dx_1 = \int_0^1 g_2(x_2)\,dx_2,$$

where

$$g_1(x_1) = \int_0^1 f(x_1,x_2)\,dx_2 = \lim_{m\to\infty}\sum_{j=1}^{M} f(x_1, x_{2,j}^m)h_m$$

and

$$g_2(x_2) = \int_0^1 f(x_1,x_2)\,dx_1 = \lim_{n\to\infty}\sum_{i=1}^{N} f(x_{1,i}^n, x_2)h_n$$

define functions $g_1(x_1)$ and $g_2(x_2)$ of x_1 and x_2 respectively. In other words, the double integral of $f(x_1,x_2)$ over $[0,1]\times[0,1]$ equals the integral of $g_2(x_2)$ over $[0,1]$,

$$\int_0^1\int_0^1 f(x_1,x_2)\,dx_1\,dx_2 = \int_0^1 g_2(x_2)dx_2 = \lim_{n\to\infty}\sum_{j=1}^{M} g_2(x_{2,j}^m)h_m,$$

and equals the integral of $g_1(x_1)$ over $[0,1]$,

$$\int_0^1 \int_0^1 f(x_1,x_2)\,dx_1\,dx_2 = \int_0^1 g_1(x_1)dx_1 = \lim_{n\to\infty}\sum_{i=1}^N g_1(x_{1,i}^n)h_n.$$

We conclude that a double integral can be computed by repeated, or iterated, integration in one dimension. We may summarize this experience as follows:

Theorem 224.3 *If $f:[0,1]\times[0,1]\to\mathbb{R}$ is Lipschitz continuous, then*

$$\int_Q f(x)\,dx = \int_0^1\int_0^1 f(x_1,x_2)\,dx_1\,dx_2 =$$
$$= \int_0^1 \big(\int_0^1 f(x_1,x_2)\,dx_2\big)\,dx_1 = \int_0^1 \big(\int_0^1 f(x_1,x_2)\,dx_1\big)\,dx_2.$$

We can interpret the statement of this theorem as a *change of order of integration* in the sense that integrating with respect to x_1 and then with respect to x_2 gives the same result as integrating first with respect to x_2 and then with respect to x_1. The usual way to evaluate a double integral is to use iterated one-dimensional integration in some order.

Example 224.1.

With $Q=[0,1]\times[0,1]$,

$$\int_Q x_1x_2^3\,dx = \int_0^1\int_0^1 x_1x_2^3\,dx_1dx_2 = \int_0^1 x_1\big(\int_0^1 x_2^3\,dx_2\big)dx_1$$
$$= \int_0^1 x_1\big[\frac{x_2^4}{4}\big]_0^1dx_1 = \frac{1}{4}\int_0^1 x_1dx_1 = \frac{1}{4}\big[\frac{x_1^2}{2}\big]_0^1 = \frac{1}{8}.$$

$$\int_Q x_1x_2^3\,dx = \int_0^1\int_0^1 x_1x_2^3\,dx_1dx_2 = \int_0^1 x_2^3\big(\int_0^1 x_1\,dx_1\big)dx_2$$
$$= \int_0^1 x_2^3\big[\frac{x_1^2}{2}\big]_0^1dx_2 = \frac{1}{2}\int_0^1 x_2^3dx_2 = \frac{1}{2}\big[\frac{x_2^4}{4}\big]_0^1 = \frac{1}{8}.$$

Alternatively, we may first integrate with respect to x_1 and then with respect to x_2 to get,

$$\int_Q x_1x_2^3\,dx = \int_0^1\int_0^1 x_1x_2^3\,dx_1dx_2 = \int_0^1 x_2^3\big(\int_0^1 x_1\,dx_1\big)dx_2$$
$$= \int_0^1 x_2^3\big[\frac{x_1^2}{2}\big]_0^1dx_2 = \frac{1}{2}\int_0^1 x_2^3dx_2 = \frac{1}{2}\big[\frac{x_2^4}{4}\big]_0^1 = \frac{1}{8}.$$

224.4 Generalization to an Arbitrary Rectangle

The double integral defined on the unit square generalizes directly to integrals over arbitrary rectangles $Q = [a_1, b_1] \times [a_2, b_2]$ with sides parallel to the axis. If $f : Q \to \mathbb{R}$ is Lipschitz continuous, then

$$\int_Q f(x)\,dx = \int_Q f(x_1, x_2)\,dx_1 dx_2 = \int_{a_1}^{b_1} \big(\int_{a_2}^{b_2} f(x_1, x_2)\,dx_2\big)\,dx_1$$
$$= \int_{a_2}^{b_2} \big(\int_{a_1}^{b_1} f(x_1, x_2)\,dx_1\big)\,dx_2.$$

224.5 Interpreting the Double Integral as a Volume

The sum

$$\sum_{i=1}^{N}\sum_{j=1}^{N} f(x_{1,i}^n, x_{2,j}^n)h_n h_n \tag{224.9}$$

represents the sum of the volumes

$$f(x_{1,i}^n, x_{2,j}^n)h_n h_n \tag{224.10}$$

of thin boxes with cross-section of area $h_n h_n$ and height $f(x_{1,i}^n, x_{2,j}^n)$. Intuitively, this is an approximation of the volume under the graph of $f(x_1, x_2)$ with (x_1, x_2) varying over Q. It is thus natural to define the volume $V(f, Q)$ under the graph of $f(x_1, x_2)$ over Q to be

$$V(f, Q) = \int_0^1 \int_0^1 f(x_1, x_2)\,dx_1\,dx_2 \tag{224.11}$$

EXAMPLE 224.2. We compute the volume of a pyramid of height 1 with base $[0, 2] \times [0, 2]$, see Fig. 224.4. One quarter of the volume is equal to the integral $\int_Q f(x)\,dx$, where $Q = [0, 1] \times [0, 1]$, $f(x) = x_2$ for $x \in Q$ such that $x_2 \le x_1$ and $f(x) = x_1$ for $x \in Q$ such that $x_1 \le x_2$. We have

$$V(f, Q) = \int_Q f(x)\,dx = \int_0^1 \big(\int_0^{x_1} x_2\,dx_2 + \int_{x_1}^1 x_1\,dx_2\big)\,dx_1$$
$$= \int_0^1 \big(\frac{x_1^2}{2} + x_1(1 - x_1)\,dx_1 = \big[\frac{x_1^2}{2} - \frac{x_1^3}{6}\big]_0^1 = \frac{1}{2} - \frac{1}{6} = \frac{1}{3}.$$

We conclude that the volume of the pyramid is equal to $\frac{4}{3}$. This agrees with the standard formula stating that the volume of a pyramid is equal to $\frac{1}{3}Bh$, where B is the area of the base and h is the height.

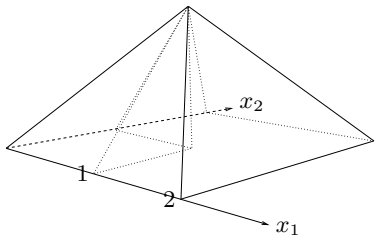


FIGURE 224.4. Volume of pyramid

224.6 Extension to General Domains

We next define the double integral of a function $f(x)$ over a more general domain Ω in the plane. We start by assuming that the boundary Γ of Ω is described by two curves $x_2 = \gamma_1(x_1)$ and $x_2 = \gamma_2(x_1)$ for $0 \le x_1 \le 1$, as shown in Fig. 224.5, so that $\Omega = \{x \in [0,1] \times \mathbb{R} : \gamma_1(x_1) \le x_2 \le \gamma_2(x_1)\}$ We assume that the functions $\gamma_i : [0,1] \to \mathbb{R}$ are Lipschitz continuous with Lipschitz constant L_γ. We further assume that $f : \Omega \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L_f.

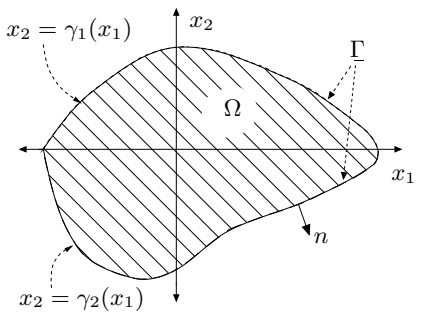


FIGURE 224.5. The domain Ω in the plane

We assume that Ω is contained in the unit square Q. We partition Q as above into squares $I_i^n \times J_j^n$ of area $h_n h_n$, where $I_i^n = (x_{1,i-1}^n, x_{1,i}^n]$ $J_j^n = (x_{2,j-1}^n, x_{2,j}^n]$. We denote by ω_n the set of indices (i,j) such that the square $I_i^n \times J_j^n$ intersects Ω, and we let Ω_n be the union of the squares $I_i^n \times J_j^n$ with indices $(i,j) \in \omega_n$. In other words, Ω_n is an approximation of Ω consisting of all the squares $I_i^n \times J_j^n$ in Q that intersect Ω. We consider the Riemann sum

$$S_n = \sum_{(i,j)\in\omega_n} f(x_{1,i}^n, x_{2,j}^n) h_n h_n. \tag{224.12}$$

We shall prove that $\lim_{n\to\infty} S_n$ exists and then naturally define

$$\int_\Omega f(x)\,dx = \lim_{n\to\infty} \sum_{(i,j)\in\omega_n} f(x_{1,i}^n, x_{2,j}^n) h_n h_n. \tag{224.13}$$

To this end, we estimate the difference $S_n - S_{n+1}$, which now has contributions from two sources; (i) from the variation of $f(x)$ over each sub-square $I_i^n \times J_j^n$, and (ii) from the difference between Ω_n and Ω_{n+1}.

The first contribution can be shown to be bounded by $L_f h_n$ by arguing just as for integration over a square. The second contribution is bounded by $2A(1+L_\gamma)h_n$, where A is a bound for $|f(x)|$, that is $|f(x)| \le A$ for $x \in \Omega$. This follows from the observation that if a square $I_i^n \times J_j^n$ of Ω_n, is entirely outside or inside Ω, then so are all the four squares of Ω_{n+1} within $I_i^n \times J_j^n$. The difference between Ω_n and Ω_{n+1} arises from the squares $I_i^n \times J_j^n$ which are partly inside and partly outside Ω. The area of these squares is bounded by $2L_\gamma h_n$, where the factor 2 arises from the fact that there are two curves γ_1 and γ_2, see Fig. 224.6. The difference in area between Ω_n and Ω_{n+1} is thus bounded by $2L_\gamma h_n$.

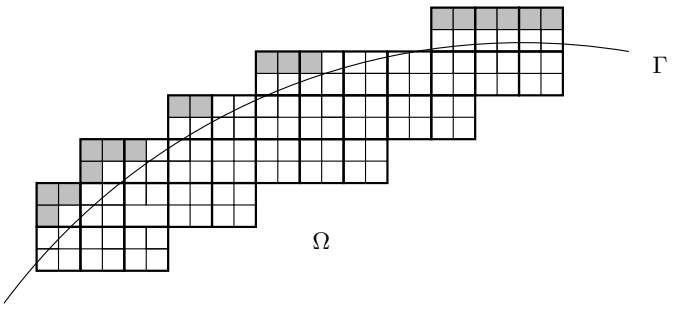


FIGURE 224.6. Approximation of integral over general domain

Together, this shows that

$$|S_n - S_{n+1}| \le (L_f + 2AL_\gamma)h_n, \tag{224.14}$$

which as proves that $\lim_{n\to\infty} S_n$ exists. We summarize as follows:

Theorem 224.4 *Let $\Omega = \{x \in [0,1]\times\mathbb{R} : \gamma_2(x_1) \le x_2 \le \gamma_1(x_1)\}$, where $\gamma_i : [0,1] \to \mathbb{R}$ are Lipschitz continuous, and let $f : \Omega \to \mathbb{R}$ be Lipschitz continuous. Then $\lim_{n\to\infty} S_n$ exists, where S_n is the Riemann sum defined by (224.12), and we define*

$$\int_\Omega f(x)\,dx = \int_\Omega f(x_1,x_2)\,dx_1dx_2 = \lim_{n\to\infty} S_n. \tag{224.15}$$

224.7 Iterated Integrals over General Domains

The integral of a function $f(x)$ over a domain $\Omega = \{x \in [0,1]\times\mathbb{R} : \gamma_2(x_1) \le x_2 \le \gamma_1(x_1)\}$ may be computed by iterated integration in one dimension as follows

$$\int_\Omega f(x)\,dx = \int_\Omega f(x_1,x_2)\,dx_1\,dx_2 = \int_0^1 \left(\int_{\gamma_2(x_1)}^{\gamma_1(x_1)} f(x_1,x_2)\,dx_2\right) dx_1. \tag{224.16}$$

This is another way of expressing the fact that,

$$\begin{aligned}\int_\Omega f(x)\,dx &= \lim_{n\to\infty} \sum_{(i,j)\in\omega_n} f(x^n_{1,i},x^n_{2,j})h_n h_n \\ &= \lim_{n\to\infty} \sum_{i=1}^N \Big(\sum_{j:(i,j)\in\omega_n} f(x^n_{1,i},x^n_{2,j})h_n\Big)h_n\end{aligned}$$

The role of x_1 and x_2 may be interchanged and the integral is independent of the particular representation of Γ. To handle a more general domain Ω, we split Ω into appropriate sub-domains Ω_j and define $\int_\Omega f\,dx = \sum_j \int_{\Omega_j} f\,dx$. Again the integral of f over Ω represents the volume of the domain under the graph of f over Ω.

Evaluation of an integral over a two-dimensional domain by repeated integration was used by Euler in 1738, when he computed the gravitational attraction of an elliptic lamina.

EXAMPLE 224.3. We compute the double integral

$$I = \int_\Omega (x_1^2 + x_2)\,dx,$$

over the domain $\Omega = \{x \in \mathbb{R}^2 : x_1^2 \le x_2 \le x_1,\ 0 \le x_1 \le 1\}$. We have

$$\begin{aligned} I &= \int_0^1 \Big(\int_{x_1^2}^{x_1} (x_1^2 + x_2)\,dx_2\Big)dx_1 = \int_0^1 \big[x_1^2 x_2 + \frac{x_2^2}{2}\big]_{x_1^2}^{x_1} dx_1 \\ &= \int_0^1 (x_1^3 + \frac{x_1^2}{2} - x_1^4 - \frac{x_1^4}{2})dx_1 = \frac{1}{4} + \frac{1}{6} - \frac{1}{5} - \frac{1}{10} = \frac{7}{60}.\end{aligned}$$

EXAMPLE 224.4. We compute the double integral

$$I = \int_\Omega \frac{1}{x_2}\,dx$$

over the domain $\Omega = \{x \in \mathbb{R}^2 : 1 \le x_2 \le \exp(x_1),\ 0 \le x_1 \le 1\}$. We have

$$I = \int_0^1 \Big(\int_1^{\exp(x_1)} \frac{1}{x_2}\, dx_2\Big) dx_1 = \int_0^1 \big[\log(x_2)\big]_1^{\exp(x_1)} dx_1 = \int_0^1 x_1 dx_1 = \frac{1}{2}.$$

224.8 The Area of a Two-Dimensional Domain

We define the *area* $A(\Omega)$ of a domain Ω in $\mathbb{R}^2$ by

$$A(\Omega) = \int_\Omega dx, \tag{224.17}$$

i.e. by integration of the constant function $f(x) = 1$ over Ω. If $\Omega = \{x \in [0,1] \times \mathbb{R} : \gamma_1(x_1) \le x_2 \le \gamma_2(x_1)\}$, then

$$A(\Omega) = \int_0^1 \left(\int_{\gamma_2(x_1)}^{\gamma_1(x_1)} dx_2 \right) dx_1 = \int_0^1 \big(\gamma_2(x_1) - \gamma_1(x_1)\big)\, dx_1,$$

which conforms with the previous formula of the area between the curves $\gamma_1(x_1)$ and $\gamma_1(x_1)$ as the integral of the difference $\gamma_2(x_1) - \gamma_1(x_1)$.

EXAMPLE 224.5. The area of the triangle Ω with corners at $(0,0)$, $(1,0)$ and $(1,1)$, can be computed as follows

$$A(\Omega) = \int_\Omega dx = \int_0^1 \Big(\int_0^{x_1} dx_2\Big)\, dx_1 = \int_0^1 \frac{1}{2}\, dx_1 = \frac{1}{2}.$$

224.9 The Integral as the Limit of a General Riemann Sum

We defined the integral using uniform subdivisions in x_1 and x_2, resulting in approximate subdivisions of a given domain Ω in $\mathbb{R}^2$ into squares or rectangles. We can however use more general subdivisions of Ω. Suppose that $f : \Omega \to \mathbb{R}$ is a Lipschitz continuous function and that the boundary of a domain Ω can be made up of pieces of Lipschitz curves $x_2 = \gamma(x_1)$ or $x_1 = \gamma(x_2)$. For $N = 1, 2, ...$, we divide Ω into a collection $\{\Omega_i\}_{i=1}^N$ of pairwise disjoint sets Ω_i such that the union of the Ω_i is equal to Ω. Let $d\Omega_i$ be the area of Ω_i and let d_N be the maximal diameter of Ω_i for $i = 1, ..., N$, see Fig. 224.7. We assume that d_N tends to zero as N tends to infinity.

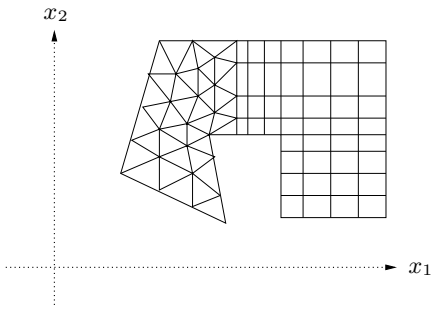


FIGURE 224.7. Subdivison of general domain

The arguments used above show that

$$\int_\Omega f(x)\,dx = \lim_{N\to\infty} \sum_{i=1}^{N} f(x_i) d\Omega_i, \tag{224.18}$$

where x_i is a point in Ω_i for $i = 1, .., N$. The first step in proving this result is to use the estimate

$$|f(x) - f(y)| \leq L_f d_N \quad \text{if } x, y \in \Omega_i, \tag{224.19}$$

which implies that the variation of $f(x)$ with x ranging over Ω_i is small if the diameter of Ω_i is small. The second step involves the Lipschitz continuity of the boundary of Ω and the boundedness of $f(x)$. By the way, a byproduct of the proof of this result is the estimate

$$|\int_\Omega f(x)\,dx - \sum_{i=1}^{N} f(x_i) d\Omega_i| \leq L_f d_N A(\Omega), \tag{224.20}$$

where $A(\Omega)$ is the area of Ω.

224.10 Change of Variables in a Double Integral

We next extend the idea of changing variables in a one-dimensional integral to a two dimensional integral. More precisely, we want to make a change of variables in an integral

$$\int_\Omega f(x)\,dx = \int_\Omega f(x_1, x_2)\,dx_1 dx_2, \tag{224.21}$$

where Ω is a given domain in $\mathbb{R}^2$ and the integration variable x runs over Ω. We assume that $g : \tilde{\Omega} \to \Omega$ is a one-to-one mapping of $y \in \tilde{\Omega}$ onto $x = g(y)$ in Ω that represents the change of variables. We shall prove that (224.21)

with respect to x can be rewritten as an integral with respect to y in the form

$$\int_\Omega f(x)\,dx = \int_{\tilde{\Omega}} f(g(y))\,G(y)\,dy, \tag{224.22}$$

where $G(y)$ is defined as

$$G(y) = |\det g'(y)|.$$

That is $G(y)$ is the absolute value of the determinant of the Jacobian $g'(y)$ of $g(y)$. Formally, this gives $dx = |\det g'(y)|\,dy$ or $|\det g'(y)| = |\det \frac{dx}{dy}|$, and $|\det g'(y)|$ is the local change of area measure as we go from y-coordinates to x-coordinates. The change of variable formula can therefore be written

$$\int_\Omega f(x)\,dx = \int_{\tilde{\Omega}} f(g(y))|\det g'(y)|\,dy, \tag{224.23}$$

To prove this let $\tilde{\Omega}_i$ be a small subdomain of $\tilde{\Omega}$ and let $\Omega_i = g(\tilde{\Omega}_i)$ be the image of $\tilde{\Omega}_i$ under the mapping $x = g(y)$. If $g'(y)$ were constant over $\tilde{\Omega}_i$, and so $g(y)$ were linear on $\tilde{\Omega}_i$, then

$$d\Omega_i = |\det g'(y_i)|d\tilde{\Omega}_i,$$

where y_i is a point in $\tilde{\Omega}_i$, $d\Omega_i$ is the area of Ω_i, and $d\tilde{\Omega}_i$ is the area of $\tilde{\Omega}_i$. If $\{\tilde{\Omega}_i\}_{i=1}^n$ is a subdivision of $\tilde{\Omega}$ into subdomains $\tilde{\Omega}_i$ of maximal diameter d_n, we have

$$\begin{aligned}\int_\Omega f(x)\,dx &\approx \sum_i f(x_i)d\Omega_i \\ &\approx \sum_i f(g(y_i)|\det g'(y_i)|d\tilde{\Omega}_i \approx \int_{\tilde{\Omega}} f(g(y))|\det g'(y)|\,dy,\end{aligned}$$

where $x_i = g(y_i)$ and the approximations are bounded by d_n times Lipschitz constants of the functions $f(x)$, $f(g(y))$ and $|\det g'(y)|$. The change of variables formula (224.23) follows by passing to the limit as n tends to infinity and d_n tends to 0.

We summarize:

Theorem 224.5 (Change of variables) *Assume $y \to x = g(y)$ maps a domain $\tilde{\Omega}$ in $\mathbb{R}^2$ onto a domain Ω in $\mathbb{R}^2$, where the Jacobian of g is Lipschitz continuous and let $f : \Omega \to \mathbb{R}$ be Lipschitz continuous. Then*

$$\int_\Omega f(x)\,dx = \int_{\tilde{\Omega}} f(g(y))|\det g'(y)|\,dy, \tag{224.24}$$

EXAMPLE 224.6. Consider the mapping $x = g(y) = (2y_1 + y_2, y_1 - 2y_2)$ mapping the unit square $\tilde{\Omega} = [0,1] \times [0,1]$ onto the parallelogram Ω

spanned by the vectors $(2,1)$ and $(1,-2)$. We have $\det g'(y) = -5$, and thus

$$\int_\Omega f(x)\,dx = \int_{\tilde\Omega} f(2y_1 + y_2, y_1 - 2y_2)\,|-5|\,dy$$
$$= 5\int_0^1 \int_0^1 f(2y_1 + y_2, y_1 - 2y_2)\,dy.$$

If $f(x) = x_2$ then

$$\int_\Omega f(x)\,dx = 5\int_0^1 \int_0^1 (y_1 - 2y_2)\,dy = 5(\frac{1}{2} - 1) = -\frac{5}{2}.$$

Polar Coordinates

A particularly important change of variables is from rectangular coordinates to polar coordinates,

$$(x_1, x_2) = (r\cos(\theta), r\sin(\theta))$$

where $x = (x_1, x_2) \in \mathbb{R}^2$ and $r \geq 0$, $0 \leq \theta < 2\pi$, see Fig. 224.8.

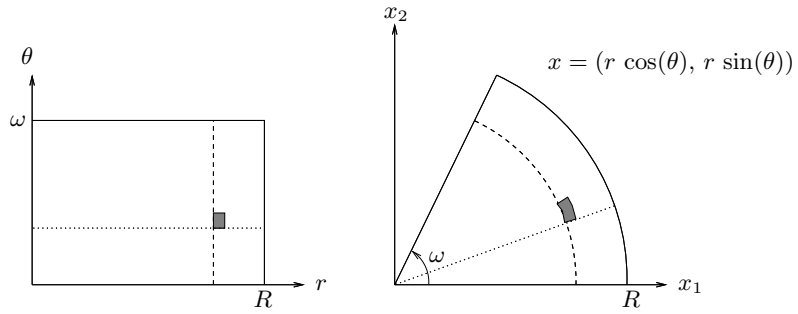


FIGURE 224.8. Polar coordinates

The Jacobian of the mapping $(r,\theta) \to (x_1, x_2)$ is given by

$$\frac{d(x_1,x_2)}{d(r,\theta)} = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix},$$

and

$$\det \frac{d(x_1,x_2)}{d(r,\theta)} = r(\cos^2(\theta) + \sin^2(\theta)) = r.$$

EXAMPLE 224.7. If $\Omega = \{x \in \mathbb{R}^2 : |x| \leq 1, x_1 \geq 0, x_2 \geq 0\}$ is the part of the unit circle in the positive quadrant, then the corresponding domain

in polar coordinates takes the from $\tilde{\Omega} = \{(r,\theta) : 0 \le r \le 1, 0 \le \theta \le \frac{\pi}{2}\}$, and

$$\int_\Omega f(x_1,x_2)\,dx_1dx_2 = \int_{\tilde{\Omega}} f(r\cos(\theta, r\sin(\theta))\,rdr\,d\theta.$$

In particular with $f(x) = 1$, we have

$$\begin{aligned} A(\Omega) &= \int_\Omega dx_1dx_2 = \int_{\tilde{\Omega}} rdr\,d\theta \\ &= \int_0^{\frac{\pi}{2}}\int_0^1 r\,dr\,d\theta = \int_0^{\frac{\pi}{2}} \frac{1}{2}\,d\theta = \frac{\pi}{4}. \end{aligned}$$

We have now computed the area of a quarter of a unit disc to be equal to $\frac{\pi}{4}$, so the area of a unit disc is π. A basic result of mathematics!

EXAMPLE 224.8. Using polar coordinates, we have

$$\int_{\mathbb{R}^2} e^{-x_1^2-x_2^2}\,dx = \int_0^{2\pi}\int_0^\infty e^{-r^2} r\,dr\,d\theta = 2\pi\big[-\frac{1}{2}e^{-r^2}\big]_0^\infty = \pi.$$

Since

$$\int_{\mathbb{R}^2} e^{-x_1^2-x_2^2}\,dx = \int_{-\infty}^\infty e^{-x_1^2}\,dx_1 \int_{-\infty}^\infty e^{-x_2^2}\,dx_2,$$

we conclude that

$$\int_{-\infty}^\infty e^{-x^2}\,dx = \sqrt{\pi}. \tag{224.25}$$

Evidently, we did something magical: although we do not know a primitive function to e^{-x^2} we are able to obtain an analytic expression for $\int_{-\infty}^\infty e^{-x^2}\,dx$.

Chapter 224 Problems

224.1. Compute with $\Omega = [0,1]\times[0,1]$ the unit square the integrals (a) $\int_\Omega (x_1 + x_2)\,dx$ (b) $\int_\Omega x_1x_2\,dx$ (c) $\int_\Omega \frac{dx}{x_1+x_2}$ (d) $\int_\Omega \exp(-x_1x_2)\,dx$

224.2. Compute with $\Omega = \{(x_1,x_2) : 0 \le x_1 \le x_2 \le 1\}$ the integrals (a) $\int_\Omega \frac{x_1}{x_2}\,dx$ (b) $\int_\Omega \exp^{2x_2}\,dx$ (c) $\int_\Omega \exp^{x_2^2}\,dx$

224.3. Change the order of integration in the following integrals

1. $\int_{1/2}^1 \int_0^{1-x_1} f(x_1,x_2)\,dx_2dx_1$
2. $\int_0^1 \int_0^{\sqrt{1-x_1^2}} f(x_1,x_2)\,dx_2dx_1$
3. $\int_0^1 \int_{x_2-1}^0 f(x_1,x_2)\,dx_1dx_2$

4. $\int_0^1 \int_{1-x_1}^{1+x_1} f(x_1, x_2)\, dx_2 dx_1$

224.4. Evaluate the following integrals:

1. $\int_\Omega (x_1^2 + 2x_2^3)\, dx$, with Ω a triangle with vertices $(0,0)$, $(1,0)$, $(0,1)$.
2. $\int_\Omega x_1^2 x_2\, dx$, with $\Omega = \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1,\ 0 \leq x_2$.
3. $\int_\Omega (x_1 + x_2)\, dx$, with Ω the tetrahedron with vertices $(0,0)$, $(1,0)$, $(2,1)$, $(2,2)$.
4. $\int_\Omega |1 - x_1 - x_2|\, dx$, with Ω the unit square.

224.5. Find the volume under the graph of the following functions

1. $f(x) = e^{x_1} \cos(x_2)$, $0 \leq x_1 \leq 1$, $0 \leq x_2 \leq \frac{\pi}{2}$.
2. $f(x) = x_1^2 e^{-x_1 - x_2}$, $0 \leq x_1 \leq 1$, $0 \leq x_2 \leq 2$.
3. $f(x) = x_1^2 x_2$, $0 \leq x_1 \leq 1$, $x_1 + 1 \leq x_2 \leq x_1 + 2$.
4. $f(x) = \sqrt{x_1^2 - x_2^2}$, $x_1^2 - x_2^2 \geq 0$, $0 \leq x_1 \leq 1$.

224.6. A cylindrical hole of radius b is drilled symmetrically through a metal sphere of radius $a > b$. Find the volume of metal removed.

224.7. Evaluate

$$\int_\Omega (1 - \frac{x_1^2}{a_1^2} - \frac{x_2^2}{a_2^2})^{3/2}\, dx$$

where Ω is the ellipse $\{x \in \mathbb{R}^2 : \frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} \leq 1\}$.

224.8. Evaluate

$$\int_\Omega \frac{x_1 + x_2}{x_1^2} e^{x_1 + x_2}\, dx,$$

where $\Omega = \{x \in \mathbb{R}^2 : x_2 \leq x_1 \leq 2 - x_2,\ 0 \leq x_2 \leq 1\}$. Hint: Use the substitution $y_1 = x_1 + x_2$, $y_2 = \frac{x_2}{x_1}$.

224.9. Compute the area of one petal of the rose $0 \leq r \leq 3\sin(\theta)$ (polar coordinates).

224.10. Compute the area within the cardoid $r = 1 + \cos(\theta)$.

224.11. Compute the following double integrals:

1. $\int_\Omega x_1 \exp(x_1 x_2)\ dx$, for $\Omega = \{x : 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 2\}$,
2. $\int_\Omega x_1 x_2 \exp(x_1 + x_2)\ dx$, for $\Omega = \{x : 1 \leq x_1 \leq 2 \leq x_2 \leq 3\}$,
3. $\int_\Omega x\ dx$, for $\Omega = \{x : 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1\}$.

224.12. Compute the following double integrals:

1. $\int_\Omega \exp(-x_1)\ dx$, for $\Omega = \{x : 0 \leq x_1 \leq 1, |x_2| \leq x_1\}$,
2. $\int_\Omega x_1 x_2 \|x\|\ dx$, for $\Omega = \{x : 0 \leq x_1 \leq 1, 1 \leq x_2 \leq 2\}$,
3. $\int_\Omega \frac{x_1}{1+x_2}\ dx$, for $\Omega = \{x : 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1 - x_1\}$.

224.13. Compute the following double integrals by changing variables:

1. $\int_\Omega \|x\|^2\ dx$, for $\Omega = \{x : x_1^2 + x_2^2 - 2x_1 - 2x_2 \leq 0\}$,
2. $\int_\Omega x_1 x_2\ dx$, for $\Omega = \{x : 3x_1^2 + x_2^2 - 2x_1 \leq 0\}$,
3. $\int_\Omega \exp(-\|x\|^2)\ dx$, for $\Omega = \mathbb{R}^2$.

225
Surface Integrals

King Karl XII of Sweden (1682-1717) had an extraordinary talent for mathematics. He was by Swedenborg (the great Swedish Universal Genius, 1688-1772) considered equal if not better than Leibniz himself. King Karl XII could easily multiply large numbers without pen and paper, and proposed 64 as the right choice of basis of the natural numbers. Over night he constructed symbols and gave names to all the digits $0, 1, ..., 62, 63$. (from *The History of Sweden*, by Herman Lindquist).

225.1 Introduction

Previously, in Chapter *Curve integrals* we defined the notion of an integral computed over a curve or a curve integral. In this chapter, we use the same ideas to define an integral over a surface or a *surface integral*. We start with the surface integral representing *surface area*.

225.2 Surface Area

Let S be a surface in $\mathbb{R}^3$ parameterized by the mapping $s : \Omega \to \mathbb{R}^3$, where Ω is a domain in $\mathbb{R}^2$ with coordinates $y = (y_1, y_2) \in \mathbb{R}^2$, so that $s = s(y) = (s_1(y), s_2(y), s_3(y))$. We define the *area* $A(S)$ of the surface S

as the following integral over the parameter domain Ω,

$$A(S) = \int_\Omega \|s'_{,1} \times s'_{,2}\|\, dy, \tag{225.1}$$

where

$$s'_{,1} = \begin{pmatrix} \frac{\partial s_1}{\partial y_1} \\ \frac{\partial s_2}{\partial y_1} \\ \frac{\partial s_3}{\partial y_1} \end{pmatrix}, \qquad s'_{,2} = \begin{pmatrix} \frac{\partial s_1}{\partial y_2} \\ \frac{\partial s_2}{\partial y_2} \\ \frac{\partial s_3}{\partial y_2} \end{pmatrix},$$

are the columns of the Jacobian

$$s' = \begin{pmatrix} \frac{\partial s_1}{\partial y_1} & \frac{\partial s_1}{\partial y_2} \\ \frac{\partial s_2}{\partial y_1} & \frac{\partial s_2}{\partial y_2} \\ \frac{\partial s_3}{\partial y_1} & \frac{\partial s_3}{\partial y_2} \end{pmatrix}.$$

Note all the coefficients are functions of $y \in \Omega$.

To motivate this definition, recall that the linearization of the mapping $s : \Omega \to \mathbb{R}^3$ at $\bar{y}$ is given by

$$y \to \hat{s}(y) = s(\bar{y}) + (y_1 - \bar{y}_1)s'_{,1}(\bar{y}) + (y_2 - \bar{y}_2)s'_{,2}(\bar{y}).$$

Consider a small square $R(\bar{y}, h) = [\bar{y}_1, \bar{y}_1 + h] \times [\bar{y}_2, \bar{y}_2 + h]$ in Ω of side length h and area h^2 with lower left-hand corner at the point $\bar{y} \in \Omega$. Here, we think of h as small. The linearization $\hat{s}(y)$ maps the square $R(\bar{y}, h)$ into a small parallelogram $P(s(\bar{y}), h)$ in the tangent plane of S through $s(\bar{y})$ spanned by the two vectors $s'_{,1}(\bar{y})$ and $s'_{,2}(\bar{y})$, with one of the corners of the parallelogram at $s(\bar{y})$. Recall now from Chapter *Analytic Geometry in* $\mathbb{R}^2$ that the area of a parallelogram spanned by two vectors a and b in $\mathbb{R}^2$ is equal to $\|a \times b\|$. So, the area of $P(s(\bar{y}), h)$ is equal to

$$\|s'_{,1}(\bar{y}) \times s'_{,2}(\bar{y})\|h^2.$$

The change of scale of area is thus $\|s'_{,1}(\bar{y}) \times s'_{,1}(\bar{y})\|$. A small piece (square) of area h^2 at $\bar{y} \in \Omega$ in the parameter domain, thus corresponds to a small piece of the surface S at $s(\bar{y})$ of area approximately $\|s'_{,1}(\bar{y}) \times s'_{,2}(\bar{y})\|h^2$, where the approximation improves as h gets smaller.

Summing over all little pieces and letting h tend to zero, we are led to define the area $A(S)$ of the surface S by (225.1), which we write as

$$A(S) = \int_\Omega \|s'_{,1}(y) \times s'_{,2}(y)\|\, dy = \int_\Omega \|s'_{,1} \times s'_{,2}\|\, dy = \int_S ds.$$

We thus write $ds = \|s'_{,1} \times s'_{,2}\|\, dy$, which expresses the change of scale. Of course, we assume that $\|s'_{,1} \times s'_{,2}\|$ is Lipschitz continuous to guarantee that the integral exists.

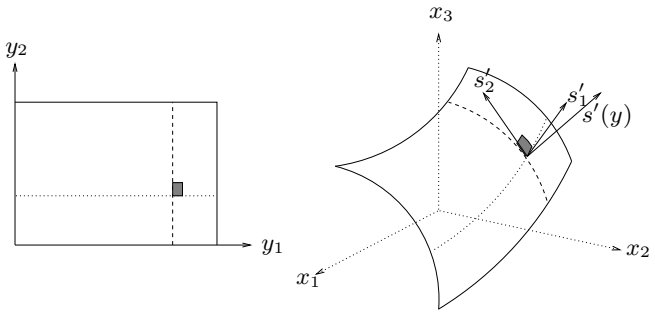


FIGURE 225.1. The surface area scale

EXAMPLE 225.1. Consider the surface S of a sphere of radius one centered at the origin. We describe this using spherical coordinates,

$$x = s(y_1, y_2) = (\sin(y_2)\cos(y_1), \sin(y_2)\sin(y_1), \cos(y_2))^\top,$$

where $0 \leq y_1 < 2\pi$, $0 \leq y_2 < \pi$, see Fig. 226.3. We have

$$\begin{aligned} s'_{,1} &= (-\sin(y_2)\sin(y_1), \sin(y_2)\cos(y_1), 0)^\top, \\ s'_{,2} &= (\cos(y_2)\cos(y_1), \cos(y_2)\sin(y_1), -\sin(y_2))^\top, \end{aligned} \tag{225.2}$$

and thus by a direct computation $\|s'\| = \sin(y_2)$. We compute

$$A(S) = \int_0^{2\pi}\int_0^{\pi} \sin(y_2)\, dy_2\, dy_1 = \int_0^{2\pi} 2\, dy_1 = 4\pi,$$

and thus conclude that the surface area of a sphere of radius 1 is equal to 4π.

EXAMPLE 225.2. We compute the area $A(S)$ of the surface S given by $s(y_1, y_2) = (2y_1y_2, y_1^2, 2y_2^2)$ with $0 \leq y_1, y_2 \leq 1$. We have

$$s'(y) = (2y_2, 2y_1, 0) \times (2y_1, 0, 4y_2) = 4(2y_1y_2, -2y_2^2, -y_1^2)$$

so that $\|s'(y)\| = 4(y_1^2 + 2y_2^2)$, and thus

$$A(S) = \int_0^1\int_0^1 4(y_1^2 + 2y_2^2)\, dy_1 dy_2 = 4(\frac{1}{3} + \frac{2}{3}) = 4.$$

225.3 The Surface Area of a the Graph of a Function of Two Variables

In the case S is given as the graph of a function $f : \Omega \to \mathbb{R}$, so that $s(y_1, y_2) = (y_1, y_2, f(y_1, y_2))$, then

$$A(S) = \int_S ds = \int_\Omega \|s'_{,1} \times s'_{,2}\|\, dy = \int_\Omega \sqrt{1 + f_{,1}^2 + f_{,2}^2}\, dy_1 dy_2, \tag{225.3}$$

where $f_{,i}$ denotes the partial derivative of f with respect to y_i. This follows from

$$s'_{,1} \times s'_{,2} = (1, 0, f_{,1}) \times (0, 1, f_{,2}) = (-f_{,1}, -f_{,2,}, 1).$$

EXAMPLE 225.3. The surface S of a hemisphere of radius 1 and centered at the origin is given by $s(y_1, y_2) = (y_1, y_2, \sqrt{1 - y_1^2 - y_2^2})$ with $y \in \Omega = \{y \in \mathbb{R}^2 : y_1^2 + y_2^2 \leq 1\}$. We have

$$\begin{aligned} A(S) &= \int_\Omega \sqrt{1 + f_{,1}^2 + f_{,2}^2}\, dy_1 dy_2 = \int_\Omega \frac{1}{\sqrt{1 - y_1^2 - y_2^2}}\, dy \\ &= \int_0^{2\pi} \int_0^1 \frac{1}{\sqrt{1 - r^2}}\, r\, dr\, d\theta = 2\pi \big[-\sqrt{1 - r^2}\big]_0^1 = 2\pi. \end{aligned} \tag{225.4}$$

We retrieve the above result that the surface area of a sphere of radius 1 is equal to 4π.

225.4 Surfaces of Revolution

Surfaces of revolution occur in many practical applications. To generate a surface of revolution, we let $f : [a, b] \to \mathbb{R}$ be a given positive function and consider the surface S represented by

$$s(x_1, x_2) = (x_1, f(x_1)\cos(x_2), f(x_1)\sin(x_2)),$$

with $a \leq x_1 \leq b$ and $0 \leq x_2 < 2\pi$, see Fig. 225.2. We use (x_1, x_2) as reference coordinates instead of (y_1, y_2). We have

$$s'_{,1} \times s'_{,2} = (1, f'(x_1)\cos(x_2), f'(x_1)\sin(x_2)) \times (0, -f(x_1)\sin(\theta), f(x_1)\cos(\theta))$$

and thus by a direct computation

$$\|s'_{,1} \times s'_{,2}\| = f(x_1)\sqrt{1 + (f'(x_1))^2}. \tag{225.5}$$

The area $A(S)$ of S is given by:

$$\begin{aligned} A(S) &= \int_0^{2\pi} \int_a^b f(x_1)\sqrt{1 + (f'(x_1))^2}\, dx_1 d\theta \\ &= 2\pi \int_a^b f(x_1)\sqrt{1 + (f'(x_1))^2}\, dx_1. \end{aligned} \tag{225.6}$$

EXAMPLE 225.4. Consider the surface S of a parabolic reflector obtained by rotating the curve $f(x_1) = \sqrt{x_1}$ around the x_1-axis between $x_1 = 0$ and $x_1 = 1$. We have

$$A(S) = 2\pi \int_0^1 \sqrt{x_1}\sqrt{1 + \frac{1}{4x_1}}\, dx_1 = \pi \int_0^1 \sqrt{4x_1 + 1}\, dx_1 = \frac{\pi}{6}(5^{3/2} - 1).$$

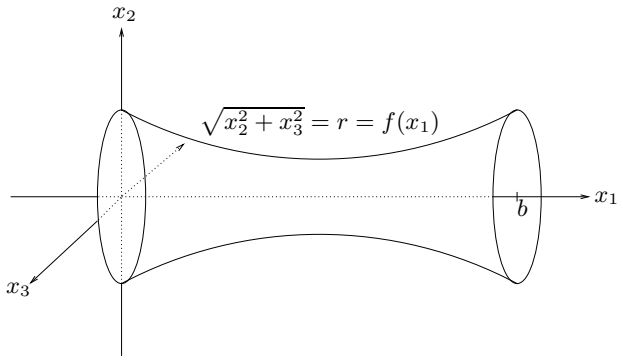


FIGURE 225.2. A surface of revolution

225.5 Independence of Parameterization

We shall prove that if $t : \tilde{\Omega} \to \Omega$ is a one-to-one mapping of $\eta \in \tilde{\Omega} \subset \mathbb{R}^2$ onto $y = t(\eta) \in \Omega$, and $r(\eta) = s(t(\eta))$ maps $\tilde{\Omega}$ onto S, then

$$\int_S ds = \int_{\tilde{\Omega}} \|r'_{,1} \times r'_{,2}\| \, d\eta = \int_{\Omega} \|s'_{,1} \times s'_{,2}\| \, dy. \tag{225.7}$$

This shows that the surface area of the surface S is independent of the parametrization of S.

We need to show that with $y = t(\eta)$, we have

$$\|r'_{,1}(\eta) \times r'_{,2}(\eta)\| = \|s'_{,1}(y) \times s'_{,2}(y)\| \, |\det t'(\eta)|, \tag{225.8}$$

where $|\det t'|$ is the determinant of the Jacobian $t'(\eta)$ of $t(\eta)$. This follows after a lengthy computation that starts with differentiating $r(\eta) = s(t(\eta))$ using the Chain rule.

225.6 Surface Integrals

Let $S = s(\Omega)$ be a surface in $\mathbb{R}^3$ parameterized by the mapping $s : \Omega \to \mathbb{R}^3$, where Ω is a domain in $\mathbb{R}^2$, and let $u : S \to \mathbb{R}$ be a real-valued function defined on S. We assume that u, s and $\|s'_{,1} \times s'_{,2}\|$ are Lipschitz continuous. We define the integral of u over S to be

$$\int_S u \, ds = \int_{\Omega} u(s(y)) \|s'_{,1}(y) \times s'_{,2}(y)\| \, dy. \tag{225.9}$$

EXAMPLE 225.5. Let $S = s(\Omega)$ be the "dome" given by $s(y_1, y_2) = (y_1, y_2, 1 - y_1^2 - y_2^2)$ and $\Omega = \{y \in \mathbb{R}^2 : y_1^2 + y_2^2 \leq 1\}$, and $u(x) =$

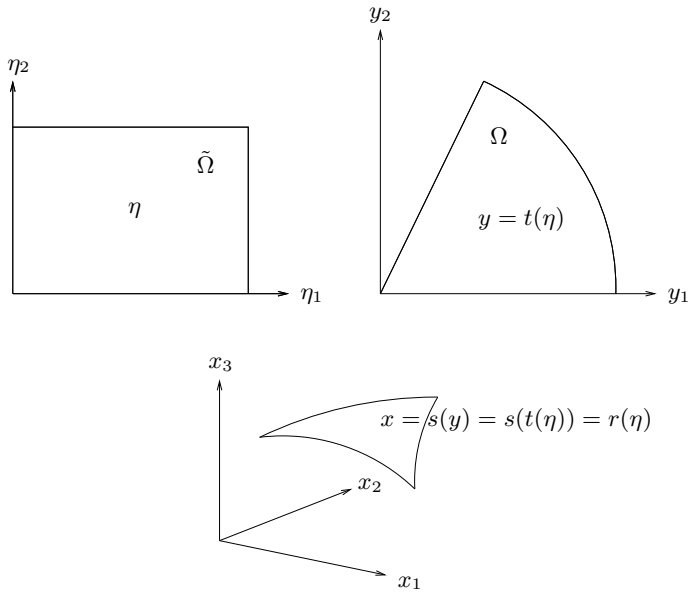


FIGURE 225.3. Reparametrization $r(\eta) = s(t(\eta))$ of a surface given by $s(y)$.

$(5x_1^2 + 5x_2^2 + x_3)^{1/2}$, so that $u(s(y)) = (5y_1^2 + 5y_2^2 + 1 - y_1^2 - y_2^2)^{1/2} = (1 + 4y_1^2 + 4y_2^2)^{1/2}$. We compute

$$\|s'_{,1}(y) \times s'_{,2}(y)\| = \|(1, 0, -2y_1) \times (0, 1, -2y_2)\| = (1 + 4y_1^2 + 4y_2^2)^{1/2},$$

and get using polar coordinates:

$$\begin{aligned}\int_S u\,ds &= \int_\Omega u(s(y))\|s'_{,1}(y) \times s'_{,2}(y)\|\,dy \\ &= \int_\Omega (1 + 4y_1^2 + 4y_2^2)^{1/2}(1 + 4y_1^2 + 4y_2^2)^{1/2}\,dy \\ &= 2\pi \int_0^1 (1 + 4r^2)r\,dr = \frac{3}{2}.\end{aligned}$$

225.7 Moment of Inertia of a Thin Spherical Shell

The *moment of inertia* of a thin sphere $S = \{\|x\| = 1\}$ of (uniformly distributed) total mass m about the x_3-axis, is equal to

$$I = \frac{m}{4\pi} \int_S (x_1^2 + x_2^2)\,ds. \tag{225.10}$$

If the sphere rotates with angular speed ω around the x_3 axis, then the total *kinetic energy* is equal to

$$E = \frac{1}{2}\frac{m}{4\pi}\int_S \omega^2(x_1^2 + x_2^2)\,ds = \frac{1}{2}\omega^2 I. \tag{225.11}$$

Using spherical coordinates to compute gives

$$I = \frac{2m}{3}. \tag{225.12}$$

Chapter 225 Problems

225.1. (a) Verify that $\|s'_{,1}(y) \times s'_{,2}(y)\| = \sin(y_2)$ in (225.2). (b) Verify (225.5). (c) Prove (225.8).

225.2. Determine which famous building is defined by the *MATLAB*© code given below, and compute the surface area of its roof.

```
r=0:.1:1;
v=0:pi/20:2*pi;
[R,V]=meshgrid(r,v);
surf(10*cos(V),10*sin(V),R.*(5+cos(V).^2-sin(V).^2))
hold on
surf(10*R.*cos(V),10*R.*sin(V),5+(R.*cos(V)).^2-(R.*sin(V)).^2)
hold off
axis('equal')
```

225.3. Another famous building. What does it take to repaint it?

```
w=0:pi/20:3*pi/4;
v=0:pi/20:2*pi;
[W,V]=meshgrid(w,v);
h=surf(sin(W).*cos(V),sin(W).*sin(V),cos(W));
set(h,'FaceColor',[1 1 1])
axis('equal')
```

225.4. Motivate (225.11), and prove (225.12).

225.5. (a) Consider the surface $S = \{x : x = y_1 a + y_2 b + (1 - y_1 - y_2)c, y \in T\}$, where $a, b, c \in \mathbb{R}^3$ and $T = \{y \in \mathbb{R}^2 : y_1 + y_2 \leq 1, y_i \geq 0, i = 1, 2\}$. Give a geometric description of S and compute its area.
(b) Find a parametrization of the form $x = My + b$ of the (flat) triangular surface S with corners in $(1, 0, 0)$, $(0, 0, 3)$ and $(0, 3, -9)$, with parameter domain T as in (a), where b is a 3-vector and M a 3-by-2 matrix.
(b) Compute the area of S. Does the area depend on b? Interpret!
(c) Compute $\int_S (x_1 + 2x_2)\,dS$

225.6. Compute (a) $\int_S dS$ (b) $\int_S f(x)\,dS$ where $S = \{x : x = My, y \in Q\}$, Q is the unit square in $\mathbb{R}^2$ and M is the 3-by-2 matrix with columns $(1,0,1)^\top$ and $(0,1,2)^\top$, and $f(x) = x_3$. Also, plot the surface S and interpret (a) as the area of S. Compare the computation of (a) with the method for computing the area of a parallelogram using the cross product in Linear algebra.

225.7. Compute (a) $\int_S dS$ (b) $\int_S x_2\,dS$ where $S = \{x : x = y_1(1-y_2)(1,0,0) + (1-y_1)(1-y_2)(1,2,0) + (1-y_1)y_2(0,1,1) + y_1y_2(0,0,3), 0 \le y_i \le 1, i = 1,2\}$. Plot the surface and describe its geometry.

225.8. Compute (a) $\int_S dS$ (b) $\int_S x_1x_2\,dS$ where $S = \{(y_1, y_2, y_1y_2) : 0 \le y_i \le 1, i = 1,2\}$.

225.9. Consider for given $r > 0$ and $h > 0$ the surface

$$S = \{x : x = (r\cos(v), r\sin(v), z),\ 0 \le v \le 2\pi,\ 0 \le z \le h\}.$$

(a) Give a geometrical description of S, and give corresponding parameterizations of the surfaces (b) $S = \{x \in \mathbb{R}^3 : x_2^2 + x_3^2 = 4,\ |x_1| \le 5\}$ (c) $S = \{x \in \mathbb{R}^3 : x_2^2 + 4x_3^2 = 4,\ 0 \le x_1 \le x_2^2 + x_3^2\}$.

225.10. Compute $\int_S (x_1+x_2+x_3)\,dS$ where $S = \{(x_1,x_2,x_3) : x_1 = y_1\cos(y_2), x_2 = y_1\sin(y_2), x_3 = y_1(\cos(y_2) + \sin(y_2))\}$.

225.11. Compute $\int_S (x_1,x_2,x_3)\cdot n\,ds$ if S is the boundary of $\Omega = \{x : x_1 + x_2 + x_3 \le 1, x_i \ge 0, i = 1,2,3\}$.

225.12. Compute $\int_S \frac{(x_1,x_2,x_3)}{\|x\|^2}\cdot n\,dS$ for the cylindrical shell $S = \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 = 1, -a \le x_3 \le a\}$, and the corresponding limit as $a \to \infty$.

225.13. Compute the moment of inertia of the cylindrical shell $S = \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 = 1, -1 \le x_3 \le 1\}$ with respect to the x_1-axis.

225.14. Compute $\int_S (x_1, 0, x_3)\cdot n\,dS$ where $S = \{(y_1 + y_2, y_1^2 - y_2^2, y_1y_2) : 0 \le y_1 \le 1, 0 \le y_2 \le 1\}$, and n is the normal to S with $n_3 < 0$.

225.15. Compute the area of the torus (donut) in $\mathbb{R}^3$ given by

$$s(y_1,y_2) = \big((a + b\cos(y_2))\cos(y_1), (a + b\cos(y_2))\sin(y_1), b\sin(y_2)\big)$$

with $a > b$ constants and $0 \le y_1, y_2 < 2\pi$.

225.16. Plot and compute the area of the surface $S = \{(r\cos(v), r\sin(v), v) : 1 \le r \le 2,\ 0 \le v \le 4\pi\}$. In what type of buildings can one find constructions like this?

225.17. Describe/plot the surfaces (of rotation, if you wish) (a) $x_1^2 + x_2^2 = x_3^2$, $x_3 > 0$ (b) $5 + x_1^2 + x_2^2 = x_3^2 \le 9$, $x_3 > 0$ and compute its area.

226
Multiple Integrals

> We met weekly, (sometimes at Dr Goddard's lodgings, sometimes at the Mitre in Wood Street near-by) at a certain hour, under a certain penalty, and a weekly contribution for the charge of experiments, with certain rules agreed among us. There, to avoid being diverted to other discourses and for some other reasons, we barred all discussion of Divinity, of State Affairs, and of news (other than what concerned our business of philosophy) confining ourselves to philosophical inquiries, and related topics; as medicine, anatomy, geometry, astronomy, navigation, statics, mechanics, and natural experiments. (Wallis about the formation of the Royal Society)

226.1 Introduction

We now consider *triple integrals* over domains in $\mathbb{R}^3$ and more generally *multiple integrals* over domains in $\mathbb{R}^n$ with $n > 3$.

226.2 Triple Integrals over the Unit Cube

A triple integral over the unit cube $Q = \{x \in \mathbb{R}^3 : 0 \le x_i \le 1,\ i = 1, 2, 3\}$ of a Lipschitz continuous function $f : Q \to \mathbb{R}$ takes the form

$$\int_Q f(x)\,dx = \int_0^1 \int_0^1 \int_0^1 f(x_1, x_2, x_3)\,dx_1\,dx_2\,dx_3.$$

This can be computed by iterated integration in any order, for example,

$$\int_Q f(x)\,dx = \int_0^1 \Big(\int_0^1 \big(\int_0^1 f(x_1,x_2,x_3)\,dx_3 \Big)\,dx_2 \Big)\,dx_1.$$

The definition of the integral and the verification of the iterated integration formula is a direct generalization of the corresponding steps in the case of a double integral over the unit square.

EXAMPLE 226.1. We compute the integral of $x_1^2 x_2 e^{x_1x_2x_3}$ over the unit cube Q,

$$\int_Q x_1^2 x_2 e^{x_1x_2x_3}\,dx = \int_0^1\int_0^1 \Big(\int_0^1 x_1^2 x_2 e^{x_1x_2x_3}\,dx_3 \Big) dx_1dx_2$$
$$= \int_0^1\int_0^1 \Big[x_1 e^{x_1x_2x_3} \Big]_{x_3=0}^{x_3=1} dx_1dx_2 = \int_0^1\int_0^1 x_1(e^{x_1x_2}-1)\,dx_1dx_2.$$

which leaves a double integral that we know how to handle.

226.3 Triple Integrals over General Domains in $\mathbb{R}^3$

Let $\Omega = \{x \in \mathbb{R}^3 : \gamma_2(x_1,x_2) \le x_3 \le \gamma_1(x_1,x_2),\ (x_1,x_2) \in \omega\}$, where ω is a domain in $\mathbb{R}^2$ and $\gamma_1 : \omega \to \mathbb{R}$ and $\gamma_2 : \omega \to \mathbb{R}$ are given functions of (x_1,x_2), see Fig. 226.1. Let $f : \Omega \to \mathbb{R}$ Lipschitz continuous. We define the triple integral of $f(x)$ over Ω by

$$\int_\Omega f(x)\,dx = \int_\omega \Big(\int_{\gamma_2(x_1,x_2)}^{\gamma_1(x_1,x_2)} f(x_1,x_2,x_3)\,dx_3 \Big) dx_1dx_2$$

via iterated integration first with respect to x_3 and then with respect to $(x_1,x_2) \in \omega$.

Expanding the double integral over ω into two one-dimensional integrals, assuming $\omega = \{(x_1,x_2,x_3) : \alpha_2 \le x_1 \le \alpha_1,\ \beta_2(x_1) \le x_2 \le \beta_1(x_1)\}$, we have

$$\int_\Omega f(x)\,dx = \int_{\alpha_2}^{\alpha_1} \Big(\int_{\beta_2(x_1)}^{\beta_1(x_1)} \Big(\int_{\gamma_2(x_1,x_2)}^{\gamma_1(x_1,x_2)} f(x)\,dx_3 \Big)\,dx_2 \Big)\,dx_1$$
$$= \int_{\alpha_2}^{\alpha_1} \Big(\int_{\omega(x_1)} f(x_1,x_2,x_3)\,dx_2dx_3 \Big)\,dx_1,$$

where $\omega(x_1) = \{(x_2,x_3) : \beta_2(x_1) \le x_2 \le \beta_1(x_1),\ \gamma_2(x_1,x_2) \le x_3 \le \gamma_1(x_1,x_2)\}$ is the cross-section of the domain Ω with a plane with fixed x_1-coordinate. This way of splitting a triple integral into an one-dimensional integral of double integrals over domain cross-sections corresponds to cutting a piece of bread or ham into slices.

We may define triple integrals similarly for more general domains by dividing the domain suitably into pieces.

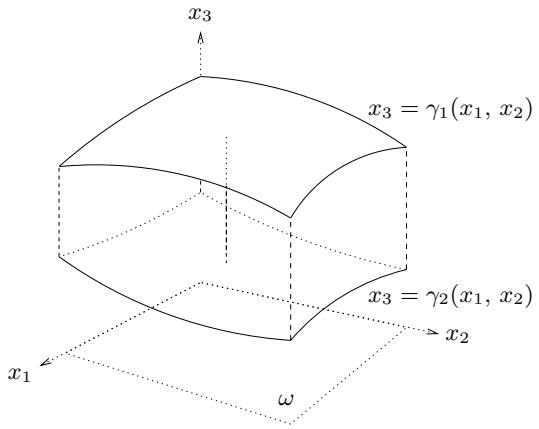


FIGURE 226.1. Integration over a volume by first integrating in the x_3-direction.

226.4 The Volume of a Three-Dimensional Domain

We define the volume $V(\Omega)$ of a domain Ω in $\mathbb{R}^3$ as

$$V(\Omega) = \int_\Omega dx,$$

i.e. by integrating $f(x) = 1$ over $x \in \Omega$. If $\Omega = \{x \in [0,1] \times [0,1] \times \mathbb{R} : \gamma_2(x_1, x_2) \le x_3 \le \gamma_1(x_1, x_2)\}$, then

$$\begin{aligned} V(\Omega) &= \int_0^1 \int_0^1 \Big(\int_{\gamma_2(x_1,x_2)}^{\gamma_1(x_1,x_2)} dx_3 \Big)\, dx_1 dx_2 \\ &= \int_0^1 \int_0^1 \Big(\gamma_1(x_1) - \gamma_2(x_1) \Big)\, dx_1\, dx_2, \end{aligned}$$

which conforms to the previous formula of the volume between the surfaces $\gamma_1(x_1, x_2)$ and $\gamma_2(x_1, x_2)$ as the integral of the difference $\gamma_1(x_1, x_2) - \gamma_2(x_1, x_2)$.

EXAMPLE 226.2. The volume of the pyramid Ω with corners at $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ described as $\{x \in \mathbb{R}^3 : 0 \le x_1 + x_2 + x_3 \le 1,\ x_1 x_2, x_3 \ge 0\}$, can be computed with $\omega = \{(x_1, x_2) : 0 \le x_1 \le 1, 0 \le x_2 \le 1 - x_1\}$ as follows

$$\begin{aligned} V(\Omega) &= \int_\omega \Big(\int_0^{1-x_1-x_2} dx_3 \Big) dx_1 dx_2 = \int_\Omega dx \\ &= \int_0^1 \Big(\int_0^{1-x_1} \Big(\int_0^{1-x_1-x_2} dx_3 \Big)\, dx_2 \Big) dx_1 \\ &= \int_0^1 \Big(\int_0^{1-x_1} (1 - x_1 - x_2)\, dx_2 \Big) dx_1 = \int_0^1 (1-x_1)^2/2\, dx_1 = \frac{1}{6}, \end{aligned}$$

which agrees with the earlier computation giving the volume of a pyramid as $\frac{1}{3}Bh$, where B is the area of the base and h the height, see Fig. 226.2.

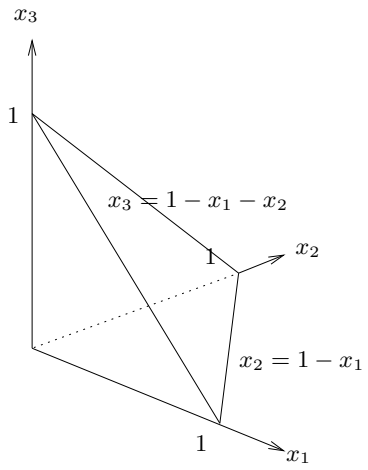


FIGURE 226.2. Integration over a pyramid.

226.5 Triple Integrals as Limits of Riemann Sums

We may also define integrals over domains in $\mathbb{R}^3$ as limits of Riemann sums

$$\int_\Omega f(x)\,dx = \lim_{N\to\infty} \sum_{i=1}^{N} f(x_i) d\Omega_i, \tag{226.1}$$

where $\{\Omega_i\}_{i=1}^N$ is a subdivision of the given domain Ω into pieces Ω_i with volume $V(\Omega_i) \le d_N$ and quadrature points $x_i \in \Omega_i$, where d_N tends to 0 as N tends to infinity. The error estimate (224.20) for double integrals generalizes directly to three dimensions.

226.6 Change of Variables in a Triple Integral

We next prove an analog of the change of variable formula for two dimensional integrals. We thus want to make a change of variables in an integral

$$\int_\Omega f(x)\,dx = \int_\Omega f(x_1,x_2,x_3)\,dx_1dx_2dx_3, \tag{226.2}$$

where Ω is a given domain in $\mathbb{R}^3$ and the integration variable x runs over Ω. If $g : \tilde{\Omega} \to \Omega$ is a one-to-one mapping of $y \in \tilde{\Omega}$ onto $x = g(y)$ in Ω, we have the following change of variables formula

$$\int_{\Omega} f(x)\,dx = \int_{\tilde{\Omega}} f(g(y))|\det g'(y)|\,dy, \tag{226.3}$$

where $|\det g'(y)|$ is the modulus of the determinant of the Jacobian $g'(y)$ of $g(y)$. Formally, we write $dx = |\det g'(y)|dy$ or $|\det g'(y)| = |\det \frac{dx}{dy}|$, and $|\det g'(y)|$ is the local change of volume measure as we go from y-coordinates to x-coordinates.

To prove this, let $\tilde{\Omega}_i$ be a small subdomain of $\tilde{\Omega}$ and let $\Omega_i = g(\tilde{\Omega}_i)$ be the image of $\tilde{\Omega}_i$ under the mapping $x = g(y)$. If $g'(y)$ were constant and$g(y)$ were linear over $\tilde{\Omega}_i$, then

$$d\Omega_i = |g'(y_i)|d\tilde{\Omega}_i, \tag{226.4}$$

where y_i is a point in $\tilde{\Omega}_i$, $d\Omega_i$ is the area of Ω_i, and $d\tilde{\Omega}_i$ is the area of $\tilde{\Omega}_i$. Thus,

$$\begin{aligned}\int_{\Omega} f(x)\,dx &\approx \sum_i f(x_i)d\Omega_i \\ &\approx \sum_i f(g(y_i)|\det g'(y_i)|d\tilde{\Omega}_i \approx \int_{\tilde{\Omega}} f(g(y))|\det g'(y)|\,dy,\end{aligned}$$

where $x_i = g(y_i)$ and $\{\tilde{\Omega}_i\}_{i=1}^N$ is a subdivision of $\tilde{\Omega}$ of maximal diameter d_N. Assuming now that $f(x)$, $f(g(y))$ and $|\det g'(y)|$ are Lipschitz continuous, the formula (226.3) follows by passing to the limit as d_N tends to 0.

Spherical Coordinates

As a particular important change of variables, we consider spherical coordinates,

$$(x_1, x_2, x_3) = (r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta), r\cos(\varphi)),$$

where $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ and $r \geq 0$, $0 \leq \theta < 2\pi$, $0 \leq \varphi < \pi$, see Fig. 226.3.

The Jacobian of the mapping $(r, \theta, \varphi) \to (x_1, x_2, x_3)$ is equal to

$$\frac{d(x_1, x_2, x_3)}{d(r, \theta, \varphi)} = \begin{pmatrix} \sin(\varphi)\cos(\theta) & -r\sin(\varphi)\sin(\theta) & r\cos(\varphi)\cos(\theta) \\ \sin(\varphi)\sin(\theta) & r\sin(\varphi)\cos(\theta) & r\cos(\varphi)\sin(\theta) \\ \cos(\varphi) & 0 & -r\sin(\varphi) \end{pmatrix}$$

and by a direct computation

$$|\det \frac{d(x_1, x_2, x_3)}{d(r, \theta, \varphi)}| = r^2\sin(\varphi). \tag{226.5}$$

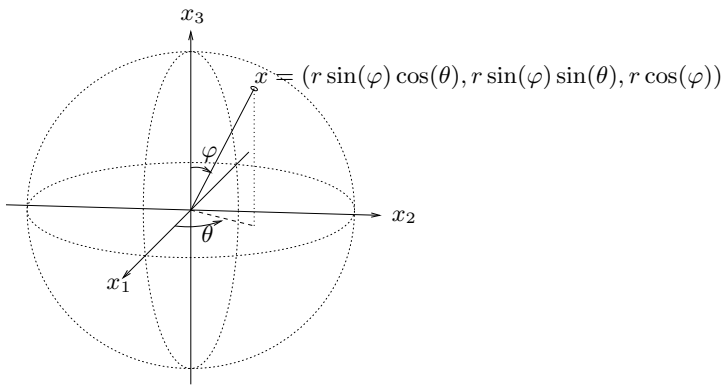


FIGURE 226.3. Spherical coordinates

The change of variables formula from Cartesian x-coordinates to spherical coordinates takes the form

$$\int_\Omega f(x)\,dx = \int_{\tilde\Omega} f\big(r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta), r\cos(\varphi)\big) r^2\sin(\varphi)\,dr\,d\theta\,d\varphi,$$

where it is understood that $\tilde\Omega$ is a subdomain of $\{(r,\theta,\varphi) : 0 \le r,\ 0 \le \theta \le 2\pi,\ 0 \le \varphi \le \pi\}$, so that $(r,\theta,\varphi) \to x$ is a one-to-one mapping of $\tilde\Omega$ onto Ω.

EXAMPLE 226.3. The unit ball $B = \{x \in \mathbb{R}^3 : |x| \le 1,\}$ is described in spherical coordinates as $\tilde B = \{(r,\theta,\varphi) : 0 \le r \le 1,\ 0 \le \theta < 2\pi,\ 0 \le \varphi < \pi\}$. The volume $V(B)$ of B is given by

$$\begin{aligned} B = \int_B dx = \int_{\tilde B} dr\,d\theta\,d\varphi &= \int_0^\pi \int_0^{2\pi} \int_0^1 r^2\sin(\varphi)\,dr\,d\theta\,d\varphi \\ &= \int_0^\pi \sin(\varphi)\,d\varphi \int_0^{2\pi} d\theta \int_0^1 r^2\,dr = 2\cdot 2\pi\frac{1}{3} = \frac{4\pi}{3}. \end{aligned}$$

Note the way the triple integral splits into a product of three one-dimensional integrals because the limits of integration are fixed numbers in all the coordinate directions and the function to be integrated is a product of functions of the individual variables.

We have shown that the volume of the unit ball in $\mathbb{R}^3$ to be equal to $\frac{4\pi}{3}$. Another basic result of Calculus!

226.7 Solids of Revolution

To generate a *solid of revolution*, we let $f : [a, b] \to \mathbb{R}$ be a given (positive) function and consider the body B in $\mathbb{R}^3$ represented by $s(x_1, r, \theta) = (x_1, r\cos(\theta), r\sin(\theta))$ with $a \le x_1 \le b$, $0 \le \theta < 2\pi$ and $0 \le r \le f(x_1)$, see Fig. 226.4. We have

$$\frac{d(x_1, x_2, x_3)}{d(x_1, r, \theta)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & \sin(\theta) \\ 0 & -r\sin(\theta) & r\cos(\theta) \end{pmatrix}$$

and thus by a direct computation

$$|\det \frac{d(x_1, x_2, x_3)}{d(x_1, r, \theta)}| = r.$$

The coordinate system (x_1, r, θ) is an example of so called cylindrical coordinates suitable for data with rotational symmetry.

The volume $V(B)$ of B is given by:

$$V(B) = \int_0^{2\pi} \int_a^b \int_0^{f(x_1)} r\, dr dx_1 d\theta = \pi \int_a^b f^2(x_1)\, dx_1. \tag{226.6}$$

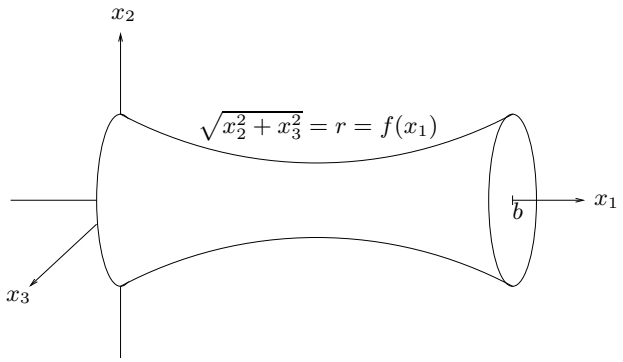


FIGURE 226.4. A solid of revolution

EXAMPLE 226.4. Consider the body B obtained by rotating the parabola $f(x_1) = \sqrt{x_1}$ around the x_1-axis between $x_1 = 0$ and $x_1 = 1$. We have

$$V(B) = \pi \int_0^1 x_1\, dx_1 = \frac{\pi}{2}.$$

EXAMPLE 226.5. The center of mass $\bar{x}$ of a body B of revolution obtained rotating a curve $f(x_1)$ around the x_1-axis from $x_1 = a$ to

$x_1 = b$ is given by $\bar{x}_2 = \bar{x}_3 = 0$ (rotational symmetry) and

$$\bar{x}_1 = \frac{\pi}{V(B)} \int_a^b x_1 f^2(x_1)\, dx_1.$$

226.8 Moment of Inertia of a Ball

The *moment of inertia* about the x_3-axis of the ball $B = \{\|x\| = 1\}$ of (uniformly distributed) total mass m, is equal to

$$I = \frac{m}{V(B)} \int_B (x_1^2 + x_2^2)\, dx. \tag{226.7}$$

If the ball rotates with angular speed ω around the x_3 axis, then the total *kinetic energy* is equal to

$$E = \frac{1}{2}\frac{m}{V(B)} \int_B \omega^2 (x_1^2 + x_2^2)\, dx = \frac{1}{2}\omega^2 I. \tag{226.8}$$

Using spherical coordinates gives

$$I = \frac{2m}{5}. \tag{226.9}$$

Chapter 226 Problems

226.1. Motivate (226.8) and prove (226.9).

226.2. Verify (226.5).

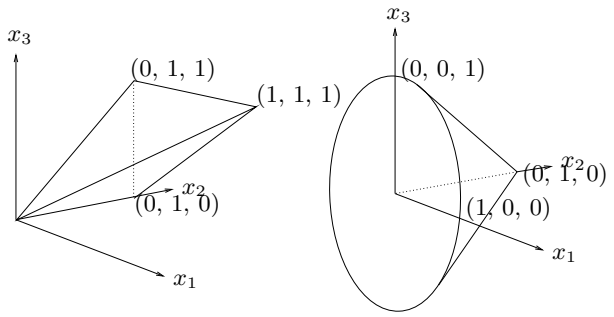
226.3. Compute the following triple integrals:

1. $\int_\Omega \|x\|^2\, dx$, for $\Omega = \{x \in \mathbb{R}^3 : 0 \le x_i \le 1, i = 1, 2, 3\}$,
2. $\int_\Omega \exp(x_1 + x_2 + x_3)\, dx$, for $\Omega = \{x \in \mathbb{R}^3 : 0 \le x_i \le 1, i = 1, 2, x_3 \le x_1 + x_2\}$,
3. $\int_\Omega 1/\|x\|^2\, dx$, for $\Omega = \{x \in \mathbb{R}^3 : 1 \le \|x\| \le 2\}$.

226.4. Compute with domains Ω as in Fig. 226.4 $\int_\Omega (1 - x_2)\, dx$.

226.5. Compute for $\Omega = \{(x_1, x_2, x_3) : x_1^2 + x_2^2 \le 1, |x_3| \le 1\}$

1. $\int_\Omega \frac{dx}{\|x\|^2}$
2. The moment of inertia of Ω with respect to the x_3-axis.
3. The moment of inertia of Ω with respect to the x_2-axis.



226.6. Compute the following multiple integrals:

1. $\int_\Omega \frac{\exp(-\|x\|)}{\|x\|}\ dx$, for $\Omega = \{x \in \mathbb{R}^3 : \|x\| > 1\}$,
2. $\int_\Omega x_1 + x_2 + x_3 + x_4\ dx$, for $\Omega = \{x \in \mathbb{R}^4 : 0 \le x_i \le 1, i = 1, 2, 3, 4\}$,
3. $\int_\Omega x_1 + \ldots + x_n\ dx$, for $\Omega = \{x \in \mathbb{R}^n : 0 \le x_i \le 1, i = 1, \ldots, n\}$.

226.7. Compute the following multiple integrals:

1. $\int_\Omega x\ dx$,
2. $\int_\Omega \|x\|\ dx$,
3. $\int_\Omega \|x\|^2\ dx$,

where $\Omega = \{x \in \mathbb{R}^3 : \|x\| \le 1\}$.

226.8. Try to generalize the result in the previous exercise to $\mathbb{R}^n$, denoting the area of the unit sphere, $\{x \in \mathbb{R}^n : \|x\| = 1\}$, by S_n.

226.9. Compute the integral $\int_{\mathbb{R}^2} \exp(-\|x\|^2)\ dx$ and use the result to compute $\int_{\mathbb{R}^n} \exp(-\|x\|^2)\ dx$.

226.10. Find the moment of inertia of a unit cube with respect to its diagonal.

226.11. Let E_y be the domain in $\mathbb{R}^n$ where the absolute value of $f : \mathbb{R}^n \to \mathbb{R}$ is larger than y, i.e. $E_y = \{x \in \mathbb{R}^n : |f(x)| > y\}$, and let $g(y)$ be the volume (size, measure) of this domain, i.e. $g(y) = \int_{E_y} dx$. Show, by changing the order of integration, that

$$\int_{\mathbb{R}^n} |f(x)|\ dx = \int_0^\infty g(y)\ dy.$$

227

Gauss' Theorem and Green's Formula in $\mathbb{R}^2$

Mathematics at its best: it looks impressive (incomprehensible), but is trivial for anyone who understands the notation. (R. Reagan)

227.1 Introduction

We now turn to two of the corner stones of calculus in several dimensions, namely *Gauss' theorem* and *Green's formula* , beginning with two dimensions. We shall see that these famous (and useful) results are direct consequences of the fundamental formula,

$$\int_\Omega \frac{\partial u}{\partial x_2}\,dx = \int_\Gamma u\,n_2\,ds, \tag{227.1}$$

where Ω is a domain in $\mathbb{R}^2$ with boundary Γ and $n(x) = (n_1(x), n_2(x))$ is the *outward unit normal* to Γ at $x \in \Gamma$, that is $n(x)$ is orthogonal to the tangent to Γ at x and points out of Ω and $\|n(x)\| = 1$, see Fig. 227.2. We shall see that this formula is an analog of the Fundamental Theorem

$$\int_a^b \frac{du}{dx}\,dx = u(b) - u(a), \tag{227.2}$$

stating that the integral over an interval $[a, b]$ of the derivative $\frac{du}{dx}$ of a function u is equal to the difference between the end-point values $u(b)$ and $u(a)$.

227.2 The Special Case of a Square

To see the connection between (227.1) and (227.2), we first assume that Ω is the unit square $[0,1]\times[0,1]$. In this case $n_2 = 1$ on the top Γ_1 of the square and $n_2 = -1$ on the bottom Γ_3, and $n_2 = 0$ on the vertical sides Γ_2 and Γ_4, see Fig. 227.2. Therefore,

$$\int_\Gamma u\, n_2\, ds = \int_0^1 u(x_1,1)\, dx_1 - \int_0^1 u(x_1,0)\, dx_1$$

if we parameterize Γ_1 by $s(x_1) = (x_1,1)$ and Γ_3 by $s(x_1) = (x_1,0)$. On the other hand, integrating first with respect to x_2 and then with respect to x_1 and using (227.2), we have

$$\begin{aligned}\int_\Omega \frac{\partial u}{\partial x_2}\, dx &= \int_0^1 \Big(\int_0^1 \frac{\partial u}{\partial x_2}(x_1,x_2)\, dx_2\Big)\, dx_1 = \int_0^1 \big(u(x_1,1) - u(x_1,0)\big)\, dx_1 \\ &= \int_0^1 u(x_1,1)\, dx_1 - \int_0^1 u(x_1,0)\, dx_1 = \int_\Gamma u\, n_2\, ds,\end{aligned}$$

which proves (227.1) when Ω is a square. We see that (227.1) results from using (227.2) with $\frac{du}{dx}\, dx$ replaced by $\frac{\partial u}{\partial x_2} dx_2$, followed by an integration with respect to x_1. The net result is that the integral of $\frac{\partial u}{\partial x_2} dx_2$ over Ω is replaced by a curve integral of un_2 over the boundary Γ of Ω.

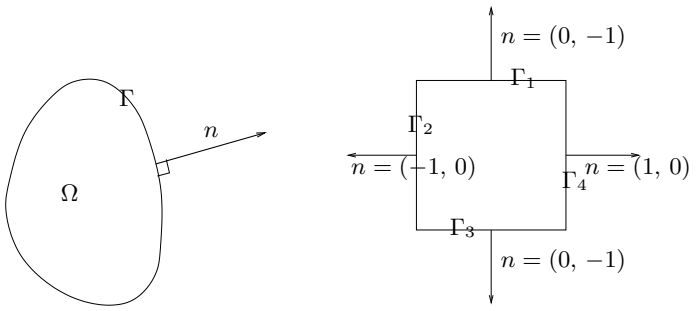


FIGURE 227.1. To the left: A domain Ω with boundary Γ and normal n. To the right: A special case.

227.3 The General Case

We now consider a domain Ω bounded by two curves Γ_1 parameterized by $s_1(x_1) = (x_1, \gamma_1(x_1))$ and Γ_2 parameterized by $s_2(x_1) = (x_1, \gamma_2(x_1))$ with $a \le x_1 \le b$, and $n = (n_1, n_2)$ is the outward normal to Γ, see Fig. 227.2.

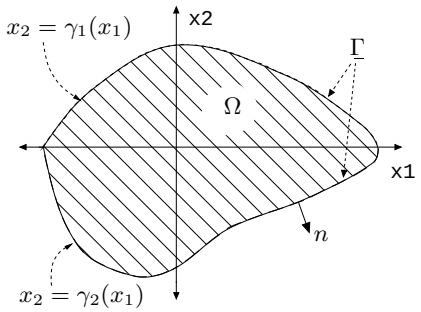


FIGURE 227.2. A domain Ω with two curves defining the boundary Γ.

The proof of (227.1) depends on the key observation that

$$\|\frac{ds_1}{dx_1}\| = \sqrt{1+(\gamma_1')^2}, \quad n_2 = \frac{1}{\sqrt{1+(\gamma_1')^2}},$$
$$\|\frac{ds_2}{dx_1}\| = \sqrt{1+(\gamma_2')^2}, \quad n_2 = -\frac{1}{\sqrt{1+(\gamma_2')^2}}.$$

Formally, $n_2 ds_1 = dx_1$ and $n_2 ds_2 = -dx_1$, see Fig. 227.3.

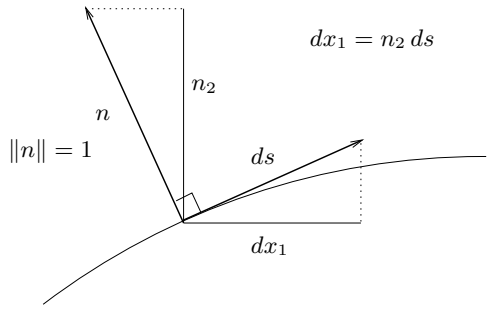


FIGURE 227.3. The key observation that $\frac{dx_1}{ds} = \frac{n_2}{1}$ by similarity.

Note that n_2 is positive on the upper boundary curve s_1 and negative on the lower boundary curve s_2. We thus have

$$\int_{\Gamma_1} u n_2\, ds_1 = \int_a^b u(x_1, \gamma_1(x_1)) n_2 \|\frac{ds_1}{dx_1}\|\, dx_1 = \int_a^b u(x_1, \gamma_1(x_1))\, dx_1,$$
$$\int_{\Gamma_2} u n_2\, ds_2 = \int_a^b u(x_1, \gamma_2(x_1)) n_2 \|\frac{ds_2}{dx_1}\|\, dx_1 = -\int_a^b u(x_1, \gamma_1(x_1))\, dx_1.$$

Secondly, integrating first with respect to x_2 and then with respect to x_1 and using the Fundamental Theorem, we see that

$$\int_\Omega \frac{\partial u}{\partial x_2}\,dx = \int_\Omega \frac{\partial u}{\partial x_2}\,dx_2\,dx_1 = \int_a^b \Big(\int_{\gamma_2(x_1)}^{\gamma_1(x_1)} \frac{\partial u}{\partial x_2}\,dx_2\Big)\,dx_1$$
$$= \int_a^b u(x_1,\gamma_1(x_1))\,dx_1 - \int_a^b u(x_1,\gamma_2(x_1))\,dx_1.$$

Since

$$\int_\Gamma u\,n_2\,ds = \int_{\Gamma_1} un_2\,ds_1 + \int_{\Gamma_2} un_2\,ds_2,$$

the desired formula (227.1) now follows. The proof generalizes to arbitrary domains bounded by smooth curves with Lipschitz continuous tangents. We summarize in the following basic theorem:

Theorem 227.1 *If Ω is a domain in $\mathbb{R}^2$ with boundary Γ with outward unit normal (n_1,n_2) and $u:\Omega\to\mathbb{R}$ is differentiable, then*

$$\int_\Omega \frac{\partial u}{\partial x_i}\,dx = \int_\Gamma u\,n_i\,ds, \quad i=1,2. \tag{227.3}$$

Applying (227.3) to the product vw of two functions v and w, we obtain the following analog of integration by parts in two dimensions:

Theorem 227.2 (Integration by parts in 2d) *If Ω is a domain in $\mathbb{R}^2$ with boundary Γ with outward unit normal (n_1,n_2) and $v,w:\Omega\to\mathbb{R}$, then*

$$\int_\Omega \frac{\partial v}{\partial x_i} w\,dx = \int_\Gamma vw\,n_i\,ds - \int_\Omega v\frac{\partial w}{\partial x_i}\,dx, \quad i=1,2. \tag{227.4}$$

Applying (227.3) to the components u_i of a vector valued function $u=(u_1,u_2)$ and summing over $i=1,2$, we obtain the *Divergence theorem*, or *Gauss' theorem*

$$\int_\Omega \nabla\cdot u\,dx = \int_\Gamma u\cdot n\,ds, \tag{227.5}$$

where $u\cdot n = u_1n_1+u_2n_2$ and

$$\nabla\cdot u = (\frac{\partial}{\partial x_1},\frac{\partial}{\partial x_2})\cdot(u_1,u_2) = \frac{\partial u_1}{\partial x_1}+\frac{\partial u_2}{\partial x_2}.$$

Applying (227.4) with w replaced by $\frac{\partial w}{\partial x_i}$ and summing over $i=1,2$, we obtain *Green's formula*:

$$\int_\Omega \nabla v\cdot\nabla w\,dx = \int_\Gamma v\partial_n w\,ds - \int_\Omega v\Delta w\,dx, \tag{227.6}$$

where

$$\partial_n w = \nabla w\cdot n = \frac{\partial w}{\partial x_1}n_1 + \frac{\partial w}{\partial x_2}n_2, \tag{227.7}$$

is the *outward normal derivative* of w on Γ. We often use Green's formula in the form

$$\int_\Omega v\Delta w\,dx - \int_\Omega \Delta v\,w\,dx = \int_\Gamma v\,\partial_n w\,ds - \int_\Gamma \partial_n v\,w\,ds, \tag{227.8}$$

which results after applying (227.6) twice and using

$$\begin{aligned}\Delta w = \text{div grad} w = \nabla\cdot\nabla w \\ = (\frac{\partial}{\partial x_1},\frac{\partial}{\partial x_2})\cdot(\frac{\partial w}{\partial x_1},\frac{\partial w}{\partial x_2}) = \frac{\partial}{\partial x_1}(\frac{\partial w}{\partial x_1})+\frac{\partial}{\partial x_2}(\frac{\partial w}{\partial x_2}),\end{aligned}$$

which can be written succinctly as $\Delta w = \frac{\partial^2 w}{\partial x_1^2}+\frac{\partial^2 w}{\partial x_2^2}$.

We also note the following analog of the Divergence theorem:

$$\int_\Omega \nabla\times u\,dx = \int_\Gamma n\times u\,ds, \tag{227.9}$$

where $u:\Omega\to\mathbb{R}^2$ and $\nabla\times u = \frac{\partial u_2}{\partial x_1}-\frac{\partial u_1}{\partial x_2}$, and $n\times u = u_2n_1-u_1n_2$. This is just a restatement of

$$\int_\Omega (\frac{\partial u_2}{\partial x_1}-\frac{\partial u_1}{\partial x_2})\,dx = \int_\Gamma \big(u_2n_1-u_1n_2\big)\,ds \tag{227.10}$$

and therefore follows from (227.3). We further note that $\tau=(-n_2,n_1)$ is a unit tangent to Γ, since $n=(n_1,n_2)$ is a unit normal and $(-n_2,n_1)\cdot(n_1,n_2)=0$, and $\tau=(-n_2,n_1)$ is directed in the counter-clockwise direction of Γ, see Fig. 227.4.

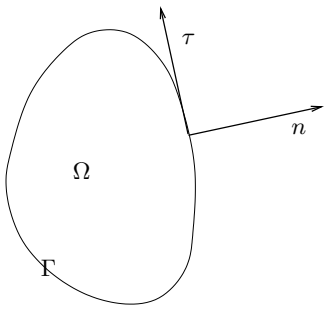


FIGURE 227.4. The unit tangent $\tau=(-n_2,n_1)$ to Γ expressed in terms of the normal $n=(n_1,n_2)$.

We often write

$$\int_\Gamma u_2n_1-u_1n_2\,ds = \int_\Gamma u\cdot\tau\,ds = \int_\Gamma u\cdot ds,$$

interpreting ds in the last integral as the *vector* τds with the old use of ds as the element of curve length. This is consistent with the notation

$$\int_\Gamma u \cdot ds = \int_a^b u(s(t)) \cdot s'(t)\, dt,$$

where $s : [a, b] \to \mathbb{R}^2$ represents Γ, which was introduced in Chapter *Curve Integrals*. Caution: we here use "ds" with two different interpretations: as the element of curve length (a scalar), and as an element of the tangent vector (a vector).

We summarize the basic results derived in this chapter as follows:

Theorem 227.3 *If Ω is a domain in $\mathbb{R}^2$ with boundary Γ with outward unit normal (n_1, n_2), and $u : \Omega \to \mathbb{R}^2$ and $v, w : \Omega \to \mathbb{R}$, then*

$$\int_\Omega \frac{\partial v}{\partial x_i}\, dx = \int_\Gamma v\, n_i\, ds, \quad i = 1, 2, \tag{227.11}$$

$$\int_\Omega \frac{\partial v}{\partial x_i} w\, dx = \int_\Gamma vw\, n_i\, ds - \int_\Omega v \frac{\partial w}{\partial x_i}\, dx, \quad i = 1, 2, \tag{227.12}$$

$$\int_\Omega \nabla \cdot u\, dx = \int_\Gamma u \cdot n\, ds, \tag{227.13}$$

$$\int_\Omega \nabla \times u\, dx = \int_\Gamma n \times u\, ds = \int_\Gamma u \cdot ds, \tag{227.14}$$

$$\int_\Omega \nabla v \cdot \nabla w\, dx = \int_\Gamma v \partial_n w\, ds - \int_\Omega v \Delta w\, dx, \tag{227.15}$$

$$\int_\Omega v \Delta w\, dx - \int_\Omega \Delta v\, w\, dx = \int_\Gamma v\, \partial_n w\, ds - \int_\Gamma \partial_n v\, w\, ds. \tag{227.16}$$

EXAMPLE 227.1. For $u(x_1, x_2) = x_1$ and $i = 1$ in (227.11), we obtain $\int_\Omega dx = \int_\Gamma x_1 n_1\, ds = \int_\Gamma x_1\, dx_2$. An interesting observation from this is that you may compute the area $\int_\Omega dx$, for example of a piece of land, simply by walking its boundary and computing $\int_\Gamma x_1\, dx_2$. The *planetometer* is a mechanical devise for computing the area of plane domains built on this principle, which has been used extensively by Surveyors.

EXAMPLE 227.2. If $\nabla \times u = 0$ in the domain Ω between two curves Γ_1 and Γ_2 that both start at the point a and end at the point b, then $\int_{\Gamma_1} u \cdot ds = \int_{\Gamma_2} u \cdot ds$, where ds is the vector tangential to the curves in the direction from a to b of length equal to the element of curve length. This follows from the fact that $\int_{\Gamma_1 \cup \Gamma_2^-} u \cdot ds = \int_\Gamma u \cdot ds = 0$, by (227.14), where Γ_2^- denotes the curve Γ_2 with the direction of ds reversed. We conclude that curve integrals of a field $u = (u_1, u_2)$ with $\nabla \times u = 0$,

that is of an *irrotational* field, is independent of the particular "path" of the curve from a to b. The integral of $u \cdot ds$ only depends on the two end-points a and b of the integration. Fields $u = (u_1, u_2)$ of this type are called *conservative*. As we shall see below, such fields are given by a *potential*, that is, they are the gradient field of some scalar potential $\varphi = \varphi(x)$ so $u = \nabla\varphi$.

Furthermore, $\int_\gamma u \cdot ds = \varphi(b) - \varphi(a)$ for a curve γ from a to b. For example, the field $u = (x_2, x_1)$ has $u_1(x_1, x_2) = x_2$ and $u_2(x_1, x_2) = x_1$ and thus $\nabla u = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} = 1 - 1 = 0$. We find easily that $u = \nabla\varphi$ for $\varphi(x) = x_1 x_2$, and the integral of $u \cdot ds$ from a point $a = (a_1, a_2)$ to $b = (b_1, b_2)$ is given by $b_1 b_2 - a_1 a_2$.

Chapter 227 Problems

227.1. Derive (227.4), (227.5), (227.6) and (227.8) from (227.3).

227.2. (a) Explain why (227.1) is valid also for a domain like $\{(x_1, x_2) : x_1 \leq |x_2|,\ x_1^2 + x_2^2 \leq 1\}$. (b) Verify by direct computation of $\int_\Omega \frac{\partial u}{\partial x_2}\, dx$ and $\int_\Gamma u\, n_2\, ds$ that (227.1) is valid for $u = r^{1/4}\sin(v/4)$ and $\Omega = \{(r\cos(v), r\sin(v)) : 0 < r < 1,\ 0 < v < 2\pi\}$, where $r = \sqrt{x_1^2 + x_2^2}$ and $v = \operatorname{arccot}(x_1/x_2)$ for $x_2 > 0$, $v = \operatorname{arccot}(x_1/x_2) + \pi$ for $x_2 < 0$ are the usual polar coordinates. Recall that by the chain rule you may express $\frac{\partial u}{\partial x_2}$ in terms of $\frac{\partial u}{\partial r}$ and $\frac{\partial u}{\partial v}$ if you like.

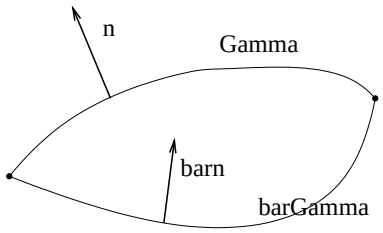
227.3. Assume $u = (u_1, u_2)$ is divergence free in Ω with boundary Γ. What can be said about (a) $\int_\Gamma u \cdot n\, ds$, (b) $u(x) \cdot n(x)$ for points x on Γ.

227.4. Assume $\int_\Gamma u \cdot n\, ds = 0$, where Γ is the boundary of a domain Ω with exterior unit normal n. What can be said about $\nabla \cdot u$ in Ω? (Before you give a too definite answer you may want to consider for example the case $u = (x_1^2, x_2^2)$ with Ω the unit disc.) Assume $\int_\gamma u \cdot n\, ds = 0$ for all closed curves γ in Ω, and the derivatives of u_i are Lipschitz. What can then be said about u in Ω?

227.5. Consider a "deformation" of $\mathbb{R}^2$ where the points $x = (x_1, x_2)$ are displaced to new positions $x + u(x)$, $u = (u_1, u_2)$, $u_i = u_i(x)$. We call $u(x)$ the displacement field and the Jacobian $u'(x)$ of $u(x)$ the deformation tensor (matrix). Consider for simplicity the case $u_i(x) = a_i x_i, i = 1, 2$, and assume the displacement is "area preserving", corresponding to "incompressibility" of the deformed material. Show that for small deformations, one has div $u \approx 0$. Hint: Consider $x \to x + u(x)$ as a change of variables and use an established fact about the Jacobian of area preserving maps.

227.6. Consider the vector field $u(x) = x/\|x\|^2$. Let Ω be the disc $\{x \in \mathbb{R}^2 : \|x - a\| \leq 1\}$, and Γ its boundary with exterior unit normal n. Compute $\int_\Gamma u \cdot n\, ds$ for $a = (2, 0)$ and $a = 0$. Do the results conform with the Divergence theorem? Make an "arrow plot" of u in the (x_1, x_2)-plane. Can you see a connection the eruption of a volcano? Does the Divergence theorem apply in the case $a = (0, 0)$?

227.7. Show that if $\nabla \cdot u = 0$ and Γ and $\bar{\Gamma}$ are curves with normals n and $\bar{n}$ as in Fig. 227.7, then $\int_\Gamma u \cdot n \, ds = \int_{\bar{\Gamma}} u \cdot \bar{n} \, ds$.



227.8. For $u = (\frac{2x_1x_2}{1+x_2^2}, -\log(1+x_2^2))$ and Γ the curve (a) $\{(x_1,x_2) : x_1^2 + x_2^2 = 1, x_i \geq 0, i = 1,2\}$ (b) $\{(x_1,x_2) : x_1 = 2-(x_2-1)^2, x_1 \geq 1\}$, compute $\int_\Gamma u \cdot n \, ds$. Hint: Close the curves and use the Divergence theorem.

227.9. Show that the field $u = e^{x_1x_2}(1 + x_1x_2, x_1^2)$ is irrotational, and find a potential φ such that $u = \nabla\varphi$.

227.10. Evaluate the integrals in (227.16) for w a solution to the *differential equation* $-\Delta w = f$ in $\Omega = \mathbb{R}^2$ and $v = -\frac{1}{2\pi}\log(x - \bar{x})$, assuming w and $\partial_n w$ vanish for $\|x\|$ large. Show that this gives a formula for $w(\bar{x})$ in terms of f and v. Hint: Take $\Omega = \{x \in \mathbb{R}^2 : \|x - \bar{x}\| > \epsilon\}$ and let ϵ tend to zero.

227.11. Let w be the solution to $-\Delta w = f$ in the upper half plane $x_2 > 0$, $-\frac{\partial w}{\partial x_2} = g$ for $x_2 = 0$, and assume w and ∇w vanish for $\|x\|$ large. Show that for $\bar{x} = (\bar{x}_1, 0)$ on $\Gamma = \{(x_1,x_2) : x_2 = 0\}$, one has $\frac{1}{2}w(\bar{x}) = \int_{\{x:x_2>0\}} vf \, dx + \int_{\{x:x_2=0\}} vg \, ds$, where $v = -\frac{1}{2\pi}\log(x - \bar{x})$. Hint: Take $\Omega = \{x \in \mathbb{R}^2 : x_2 > 0, \|x - \bar{x}\| > \epsilon\}$ in (227.16), and let ϵ tend to zero.

227.12. Show that for *harmonic functions* v and w, that is with $\Delta v = 0$ and $\Delta w = 0$, one has $\int_\Gamma \partial_n v w \, ds = \int_\Gamma v \partial_n w \, ds$ for a closed curve Γ.

227.13. Find the area of the domain enclosed by the curve

$$\Gamma = \{(r\cos(v), r\sin(v)) : r = 2 + \sin(v),\ 0 \leq v < 2\pi\}.$$

Hint: Integrals of the form $\int \sin^4(v)\,dv$ and $\int \cos^4(v)\,dv$ may be computed using integration by parts, as follows:

$$\begin{aligned} I &= \int \cos^4(v)\,dv = \int (1-\sin^2(v))\cos(v)\cdot\cos(v)\,dv = \{\text{int. by parts}\} \\ &= (\sin(v) - \frac{1}{3}\sin^3(v))\cdot\cos(v) - \int (\sin(v) - \frac{1}{3}\sin^3(v))(-\sin(v))\,dv \\ &= (\sin(v) - \frac{1}{3}\sin^3(v))\cdot\cos(v) + \int \sin^2(v)\,dv - \frac{1}{3}\int (1-\cos^2(v))^2\,dv \\ &= (\sin(v) - \frac{1}{3}\sin^3(v))\cos(v) + \int \sin^2(v)\,dv - \frac{1}{3}\int (1-2\cos^2(v))\,dv - \frac{1}{3}I, \end{aligned}$$

from which I can be computed.

228

Gauss' Theorem and Green's Formula in $\mathbb{R}^3$

> Of those who with me have written something about these matters, either I alone am mad, or I alone am not mad. No third option can be maintained, unless (as perchance it may seem to some) we are all mad. (Hobbes to Wallis)
>
> If he is mad, he is not likely to be convinced by reason; on the other hand, if we be mad, we are in no position to attempt it. (Wallis to Hobbes)

We now extend the results of the previous chapter to three dimensions. The basic result is the following analog of (227.1): If Ω is a domain in $\mathbb{R}^3$ with boundary Γ, then

$$\int_\Omega \frac{\partial u}{\partial x_3}\,dx = \int_\Gamma u\,n_3\,ds, \tag{228.1}$$

where (n_1, n_2, n_3) is the outward normal to Γ. To prove this, we assume that Γ is composed of the two surfaces Γ_1 given by $s_1(x_1, x_2) = (x_1, x_2, \gamma_1(x_1, x_2))$ and Γ_2 given by $s_2(x_1, x_2) = (x_1, x_2, \gamma_2(x_1, x_2))$, where $(x_1, x_2) \in \omega$ and the parameter domain ω is a domain in $\mathbb{R}^2$, and we assume that $\Omega = \{x \in \mathbb{R}^3 : (x_1, x_2) \in \omega, \gamma_2(x_1, x_2) < x_3 < \gamma_1(x_1, x_2)\}$, see Fig. 228.1. We have $s'_{i,1} \times s'_{i,2} = (1, 0, \gamma_{i,1}) \times (0, 1, \gamma_{i,2})$ for $i = 1, 2$, where $\gamma_{i,j} = \frac{\partial \gamma_i}{\partial x_j}$, and thus on Γ_1

$$\|s'_{1,1} \times s'_{1,2}\| = \sqrt{1 + (\gamma'_{1,1})^2 + (\gamma'_{1,2})^2}, \quad n_3 = \frac{1}{\sqrt{1 + (\gamma'_{1,1})^2 + (\gamma'_{1,2})^2}}$$

and on Γ_2

$$\|s'_{2,1} \times s'_{2,2}\| = \sqrt{1 + (\gamma'_{2,1})^2 + (\gamma'_{2,2})^2}, \quad n_3 = -\frac{1}{\sqrt{1 + (\gamma'_{2,1})^2 + (\gamma'_{2,2})^2}}.$$

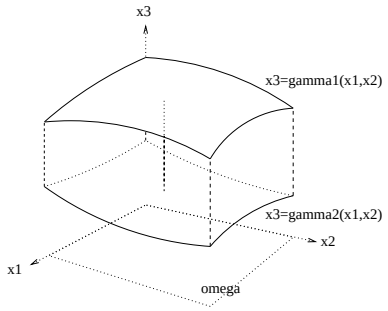


FIGURE 228.1. A domain Ω bounded by two graphs Γ_1 and Γ_2.

Integrating first with respect to x_3 and using the Fundamental Theorem we get

$$\begin{aligned}
&\int_\Omega \frac{\partial u}{\partial x_3}\,dx = \int_\Omega \frac{\partial u}{\partial x_3}\,dx_3dx_1dx_2 \\
&= \int_\omega u(x_1,x_2,\gamma_1(x_1,x_2))\,dx_1dx_2 - \int_\omega u(x_1,x_2,\gamma_2(x_1,x_2))\,dx_1dx_2 \\
&= \int_\omega u\,n_3\|s'_{1,1}\times s'_{1,2}\|dx_1dx_2 + \int_\omega u\,n_3\|s'_{2,1}\times s'_{2,2}\|dx_1dx_2 \\
&= \int_{\Gamma_1} u\,n_3\,ds + \int_{\Gamma_2} u\,n_3\,ds = \int_\Gamma u\,n_3\,ds,
\end{aligned}$$

which proves (228.1). Note that $n_3 = 0$ on the "vertical" parts of Γ! This result generalizes to

$$\int_\Omega \frac{\partial u}{\partial x_i}\,dx = \int_\Gamma u\,n_i\,ds, \quad i = 1,2,3, \tag{228.2}$$

for a general domain Ω in $\mathbb{R}^3$ with boundary Γ with outward unit normal (n_1,n_2,n_3).

Applying (228.2) to the product vw of two functions v and w, we obtain the analog of integration by parts in three dimensions,

$$\int_\Omega \frac{\partial v}{\partial x_i}w\,dx = \int_\Gamma vw\,n_i\,ds - \int_\Omega v\frac{\partial w}{\partial x_i}\,dx, \quad i = 1,2,3. \tag{228.3}$$

Applying (228.3) to the components u_i of a vector valued function $u = (u_1,u_2,u_3)$ with $w = 1$ and summing over i, we obtain the *Divergence*

theorem, or *Gauss' theorem* in three dimensions

$$\int_\Omega \nabla \cdot u \, dx = \int_\Gamma u \cdot n \, ds, \tag{228.4}$$

where $\nabla \cdot u = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}) \cdot (u_1, u_2, u_3) = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3} = \sum_{i=1}^3 \frac{\partial u_i}{\partial x_i}$, and $u \cdot n = u_1 n_1 + u_2 n_2 + u_3 n_3$ is the component of u in the direction of the normal n. If u represents a *flux* of some quantity, like heat flux or water flux, then $u(x) \cdot n(x)$ at a point $x \in \Gamma$ represents the flux through Γ (out of Ω), or *normal flux*, and thus

$$\int_\Gamma u \cdot n \, ds$$

represents the *total flux* through Γ.

We also directly obtain the following analog of Gauss' theorem for a function $u : \Omega \to \mathbb{R}^3$:

$$\int_\Omega \nabla \times u \, dx = \int_\Gamma n \times u \, ds, \tag{228.5}$$

which is now a vector equation!!

Another consequence of (228.3) is *Green's formula*:

$$\int_\Omega \nabla v \cdot \nabla w \, dx = \int_\Gamma v \partial_n w \, ds - \int_\Omega v \Delta w \, dx, \tag{228.6}$$

where $\partial_n v = \nabla v \cdot n = \frac{\partial v}{\partial x_1} n_1 + \frac{\partial v}{\partial x_2} n_2 + \frac{\partial v}{\partial x_3} n_3$ is the outward normal derivative of v on Γ, and now $\Delta w = \frac{\partial^2 w}{\partial x_1^2} + \frac{\partial^2 w}{\partial x_2^2} + \frac{\partial^2 w}{\partial x_3^2}$. We often use Green's formula in the form

$$\int_\Omega v \Delta w \, dx - \int_\Omega \Delta v \, w \, dx = \int_\Gamma v \, \partial_n w \, ds - \int_\Gamma \partial_n v \, w \, ds, \tag{228.7}$$

which results after applying (228.6) twice.

We summarize the basic results derived in this chapter as follows:

Theorem 228.1 *If Ω is a domain in $\mathbb{R}^3$ with boundary Γ with outward unit normal $n = (n_1, n_2, n_3)$, and $u : \Omega \to \mathbb{R}^3$ and $v, w : \Omega \to \mathbb{R}$, then*

$$\int_\Omega \frac{\partial v}{\partial x_i} \, dx = \int_\Gamma v \, n_i \, ds, \quad i = 1, 2. \tag{228.8}$$

$$\int_\Omega \frac{\partial v}{\partial x_i} w \, dx = \int_\Gamma v w \, n_i \, ds - \int_\Omega v \frac{\partial w}{\partial x_i} \, dx, \quad i = 1, 2. \tag{228.9}$$

$$\int_\Omega \nabla \cdot u \, dx = \int_\Gamma u \cdot n \, ds \tag{228.10}$$

$$\int_\Omega \nabla \times u\,dx = \int_\Gamma n \times u\,ds, \tag{228.11}$$

$$\int_\Omega \nabla v \cdot \nabla w\,dx = \int_\Gamma v\partial_n w\,ds - \int_\Omega v\Delta w\,dx, \tag{228.12}$$

$$\int_\Omega v\Delta w\,dx - \int_\Omega \Delta v\,w\,dx = \int_\Gamma v\,\partial_n w\,ds - \int_\Gamma \partial_n v\,w\,ds. \tag{228.13}$$

EXAMPLE 228.1. We compute the total flow of the vector field $u(x) = (x_1 + x_2^5, x_2 + x_3x_1, x_3 + x_1x_2)$ out of the boundary S of the unit ball $B = \{x \in \mathbb{R}^3 : \|x\| = 1\}$, that is the integral,

$$\int_S u\cdot n\,ds = \int_S ((x_1+x_2^5)x_1+(x_2+x_3x_1)x_2+(x_3+x_1x_2)x_3)\,ds, \tag{228.14}$$

where we used that the outward unit normal n to S at $x \in S$ is given by $n(x) = x$. Since div $u(x) = 3$ for $x \in \mathbb{R}^3$, we have by Gauss's theorem

$$\int_S u \cdot n\,ds = \int_B 3\,dx = 3V(B) = 4\pi,$$

which gives a quick way of computing the quite difficult integral (228.14).

228.1 George Green (1793-1841)

George Green, a millers son and self-taught mathematician (he left school at age 9 after two years of study), published 1827 on his own "An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism" introducing in particular so-called *Green's functions* forming the basis of the modern theory of partial differential equations. His importance in mathematics was only recognized after his death in the work by in particular Maxwell on electromagnetics.

Chapter 228 Problems

228.1. Write out and verify (228.5) from (228.2).

228.2. (a) Prove Green's formula (228.6) using (228.3). (b) Prove (228.13).

228.3. Compute the integral $\int_\Gamma \frac{x}{\|x\|^3}\cdot n\,ds$, where $\Gamma = \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 + (x_3 - ja)^2 = a^2\}$ for $a > 0$ and $j = 0, 1, 2$, respectively, where n is the exterior unit normal to Γ. Interpret the results.

228.4. Compute the integral $\int_\Gamma \frac{1}{x_1^2+x_2^2}\frac{(-x_2,x_1)}{x_1^2+x_2^2} \times ds$, where $\Gamma = \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 + x_3 = 1\}$.

228.5. Let Γ be the unit sphere in $\mathbb{R}^3$ with exterior unit normal n and compute the following integrals:

1. $\int_\Gamma x \cdot n \, ds$,
2. $\int_\Gamma x \times n \, ds$,
3. $\int_\Gamma \frac{1}{\|x\|^2} \frac{x}{\|x\|} \cdot n \, ds$,
4. $\int_\Gamma \frac{1}{\|x\|^2} \frac{x}{\|x\|} \times n \, ds$.

228.6. Verify that for a radial field $F(x) = \|x\|^\alpha \frac{x}{\|x\|}$ one has div $F = (\alpha+2)\|x\|$.

228.7. What is the smallest possible value of the integral $\int_\Gamma F \cdot n \, ds$, where $F(x) = (x_1x_2^2 - 4x_1x_2, 4x_2x_3^2 + 8x_2x_3 + 5x_2, x_1^2x_3 - 2x_1x_3)$ and Γ is a closed surface in $\mathbb{R}^3$, and n its exterior unit normal? Hint: Enclose all the "sinks" of F, that is, consider the domain where div $F \leq 0$.

228.8. Compute the surface integral $\int_\Gamma F \cdot n \, ds$, where

$$F(x) = (x_2^2, x_1x_2(\cos(x_1))^2 + x_1x_2^3 + \exp(\cos(x_1x_3^2)), x_1x_3(\sin(x_1))^2 - 3x_1x_2^2x_3),$$

and Γ is the part of the sphere $\|x\| = 2$ with positive x_3-coordinate, and n its normal with also positive x_3-component. Hint: The function F is chosen *seemingly* difficult only to confuse you.

228.9. Let $\{x_1, x_2, ..., x_N\}$ be a set of points in $\mathbb{R}^n$, and let

$$F(x) = \sum_{j=1}^{N} \frac{1}{4\pi\|x - x_j\|^2} \frac{x - x_j}{\|x - x_j\|}.$$

Compute the surface integral $\int_\Gamma F \cdot n \, ds$ for any closed surface Γ containing $k \leq N$ of the points $\{x_1, x_2, ..., x_N\}$, with n the exterior unit normal as usual.

228.10. Show, as you did in Chapter Newton's Nightmare, that the gravitation from a sphere is the same as if all the mass of the sphere was concentrated to its center, but now using Gauss' theorem to make things easier. Use as starting point that the divergence of the gravitational field is (proportional to) the density, i.e.

$$\nabla \cdot F = \rho/c$$

for some constant c, and assume spherical symmetry, i.e. the direction of the gravitational field is in the radial direction from the center of the sphere.

228.11. Show that if $-\Delta u = f$ in Ω, then for any function v that is zero on Γ, the boundary of Ω, one has

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv.$$

Also prove that if $\partial_n u = g$ on Γ, then for all functions v,

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv + \int_\Gamma gv \, ds.$$

FIGURE 228.2. Stokes at age 22: "After taking my degree I continued to reside in College and took private pupils. I thought I would try my hand at original research...."

229
Stokes' Theorem

> I too feel that I have been thinking too much of late, but in a different way, my head running on divergent series, the discontinuity of arbitrary constants, ... I often thought that you would do me good by keeping me from being too engrossed by those things. (Stokes asking Mary Susanna Robinson to marry him 1857)

229.1 Introduction

Stokes' theorem states that if $u : \mathbb{R}^3 \to \mathbb{R}^3$ is differentiable, then

$$\int_S (\nabla \times u) \cdot n \, ds = \int_\Gamma u \cdot ds \,, \tag{229.1}$$

where S is a surface in $\mathbb{R}^3$ bounded by a closed curve Γ, n is a unit normal to S, and Γ is oriented in a clockwise direction following the positive direction of the normal n, see Fig. 229.1. The integral

$$\int_\Gamma u \cdot ds$$

is called the *circulation* of u around Γ. The integral

$$\int_S (\nabla \times u) \cdot n \, ds$$

is the *total flow of the rotation* $\nabla \times u$ *across the surface* S. Stokes' theorem states that the total flow of $\nabla \times u$ across S is equal to the circulation of u around the boundary Γ of S.

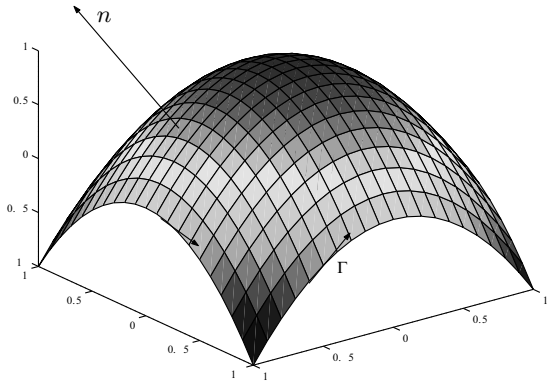


FIGURE 229.1. A Stokes surface S with boundary curve Γ.

Stokes (1819-1903), an Irish mathematician/physicist and professor in Cambridge 1849, gave basic contributions to the theory of viscous fluid flow modeled by the Navier-Stokes equations, see Fig. 228.2.

229.2 The Special Case of a Surface in a Plane

We start by considering the special case of a plane surface $\bar{S}$ in the plane $\{x \in \mathbb{R}^3 : x_3 = 0\}$ with normal $\bar{n} = (0, 0, 1)$ and with boundary Γ, see Fig. **??**. In this case, Stokes' theorem takes the form

$$\int_{\bar{S}} (\nabla \times u) \cdot \bar{n}\, ds = \int_{\bar{S}} \left(\frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} \right) dx_1\, dx_2$$
$$= \int_\Gamma u \cdot ds = \int_\Gamma (u_2 n_1 - u_1 n_2)\, ds. \quad (229.2)$$

By identifying the plane $\{x_3 = 0\}$ with $\mathbb{R}^2$, this is (227.10) and is a direct consequence of (227.3). This result is often referred to as *Green's formula in two dimensions*. We have thus proved Stokes' theorem in the case of a plane surface S in the plane $\{x_3 = 0\}$.

Note that the unit tangent direction is given by $\tau = (-\tilde{n}_2, \tilde{n}_1)$, where $\tilde{n} = (\tilde{n}_1, \tilde{n}_2)$ is the outward normal direction to Γ in the plane $\{x : x_3 = 0\}$ with a counter clockwise orientation when viewed from the top of the normal $\bar{n} = (0, 0, 1)$ of $\bar{S}$. The orientation is consistent with the specification that τ should be oriented clockwise when following the direction of the normal to $\bar{S}$.

EXAMPLE 229.1. Let $S = \{x \in \mathbb{R}^3 : \|x\| \leq 1, x_3 = 0\}$ be the unit disc in the plane $\{x_3 = 0\}$ bounded by the curve Γ parameterized

by $s(t) = (\cos(t), \sin(t), 0)$, $0 \le t \le 2\pi$. Choose $n = (0, 0, 1)$ and let $u(x) = (-x_2, x_1, 0)$ so that $\nabla \times u(x) = (0, 0, 2)$. We compute

$$\int_S (\nabla \times u)\, ds = 2\pi, \quad \int_\Gamma u \cdot ds = \int_0^{2\pi} (\cos^2(t) + \sin^2(t))dt = 2\pi,$$

in accordance with Stokes' theorem.

EXAMPLE 229.2. Ampere's law states the $\nabla \times H = J$, where H is the magnetic field and J the electric current. Stokes' theorem states that the circulation of H around a closed curve Γ bounding a surface S is equal to the total current through the surface S. Stokes' theorem is thus one of the corner-stones of electromagnetic field theory.

229.3 Generalization to an Arbitrary Plane Surface

We shall now verify that both the left and right hand side of Stokes' equality

$$\int_S (\nabla \times u) \cdot n\, ds = \int_\Gamma u \cdot ds\,,$$

are invariant under orthogonal coordinate transformations. We thus obtain a proof of Stokes' theorem for a given plane surface S through the origin, by choosing coordinates so that S lies in the plane $\{x_3 = 0\}$, and using the proof of the previous section. The case of a surface S not passing through the origin is reduced to the previous case by a simple translation of the origin of the coordinate system.

To prove the invariance, let $x = Q\bar{x}$ be an orthogonal coordinate transformation with Q an orthogonal 3×3 matrix from a set of coordinates $\bar{x}$ to x. The dependent vector variable u also transforms as $u = Q\bar{u}$, where u are the components in x-coordinates and $\bar{u}$ the coordinates of the same quantity in $\bar{x}$-coordinates. We have a similar relation between the elements of integration $ds = s'(t)dt$ and $d\bar{s} = \bar{s}'(t)dt$ in the different coordinates since $s'(t) = Q\bar{s}'(t)$, that is $ds = Qd\bar{s}$. Therefore,

$$\int_\Gamma u \cdot ds = \int_\Gamma Q\bar{u} \cdot Qd\bar{s} = \int_\Gamma Q^\top Q\bar{u} \cdot d\bar{s} = \int_\Gamma \bar{u} \cdot d\bar{s},$$

and the invariance of the right hand side of (241.23) follows.

To prove the invariance of the left hand side of (241.23), we use the Chain rule to obtain the following relation between the gradient ∇ with respect to x and the gradient $\bar{\nabla}$ with respect to $\bar{x}$,

$$\nabla = Q\bar{\nabla}.$$

A direct computation shows that

$$(\nabla \times u) \cdot n = (Q\bar{\nabla} \times Q\bar{u}) \cdot Q\bar{n} = (\bar{\nabla} \times \bar{u}) \cdot \bar{n}, \tag{229.3}$$

which proves the invariance since $d\bar{x} = dx$. Note that (229.3) is analogous to the the relation

$$(Qa \times Qb) \cdot Qc = (a \times b) \cdot c$$

for $a, b, c \in \mathbb{R}^3$. This expresses the invariance of the volume spanned by three vectors a, b and c under orthogonal coordinate transformations.

229.4 Generalization to a Surface Bounded by a Plane Curve

Suppose that S is a surface bounded by a curve Γ contained in the plane $\{x_3 = 0\}$, see Fig. **??**. We do not assume that S is contained in $\{x_3 = 0\}$. Let $\bar{S}$ be the surface in the plane $\{x_3 = 0\}$ with the boundary Γ and let Ω be the volume bounded by the surface S and the plane surface $\bar{S}$. Since $\nabla \cdot (\nabla \times u) = 0$, the Divergence theorem implies

$$0 = \int_\Omega \nabla \cdot (\nabla \times u)\, dx = \int_S \nabla \times u \cdot n\, ds + \int_{\bar{S}} \nabla \times u \cdot n\, ds, \qquad (229.4)$$

where n is the outward unit normal to the boundary $\partial\Omega$ of Ω. If n is a normal to S and $\bar{n} = -n$ is a normal to $\bar{S}$, then (229.4) implies

$$\int_S \nabla \times u \cdot n\, ds = \int_{\bar{S}} \nabla \times u \cdot \bar{n}\, ds.$$

Applying Stokes' theorem to $\bar{S}$, we obtain

$$\int_S \nabla \times u \cdot n\, ds = \int_{\bar{S}} \nabla \times u \cdot \bar{n}\, ds = \int_\Gamma u \cdot ds,$$

which proves Stokes' theorem for the surface S bounded by the plane curve Γ.

A proof of Stokes' theorem for the case of a general curve is outlined in Problem 229.1. We now summarize:

Theorem 229.1 (Stokes' theorem). *If S is a surface in $\mathbb{R}^3$ with unit normal n, and Γ is the boundary of S oriented clockwise following the direction of n, then*

$$\int_S (\nabla \times u) \cdot n\, ds = \int_\Gamma u \cdot ds.$$

We state the following important direct consequence of Stokes' theorem:

Theorem 229.2 *If $u : \Omega \to \mathbb{R}^3$ with Ω a domain in $\mathbb{R}^3$ is a differentiable vector field such that*

$$\int_\Gamma u \cdot ds = 0$$

for all closed curves Γ in Ω, then $\nabla \times u = 0$ in Ω.

Chapter 229 Problems

229.1. Prove Stokes theorem for a curve Γ given by $s(t) = (x_1(t), x_2(t), f(x_1(t), x_2(t)))$, $t \in [a,b]$, where $f : \mathbb{R}^2 \to \mathbb{R}$, bounding a surface $\Omega = \{x \in \mathbb{R}^3 : x_3 - f(x_1,x_2) = 0\}$ in $\mathbb{R}^3$. Hint: The projection of Γ on the x_1x_2-plane is the curve $\tilde{\Gamma}$ represented by $\tilde{s}(t) = (x_1(t), x_2(t), 0)$ which bounds the domain $\tilde{\Omega}$ in the x_1x_2-plane. Show that, writing $u_i = u_i(x_1, x_2, f(x_1,x_2))$,

$$\begin{aligned}\int_\Gamma u\cdot ds &= \int_a^b \big(u_1x_1' + u_2x_2' + u_3(\frac{\partial f}{\partial x_1}x_1' + \frac{\partial f}{\partial x_2}x_2')\big)\,dt\\ &= \int_a^b \big((u_1 + u_3\frac{\partial f}{\partial x_1})x_1' + (u_2 + u_3\frac{\partial f}{\partial x_2})x_2'\big)\,dt\\ &= \int_{\tilde{\Gamma}} \big(u_1 + u_3\frac{\partial f}{\partial x_1}, u_2 + u_3\frac{\partial f}{\partial x_2}\big)\cdot ds = I.\end{aligned}$$

Then use the Stokes theorem for a plane curve established above, to show that

$$I = \int_{\tilde{\Omega}} \big(\frac{\partial}{\partial x_1}(u_2 + u_3\frac{\partial f}{\partial x_2}) - \frac{\partial}{\partial x_2}(u_1 + u_3\frac{\partial f}{\partial x_1})\big)\,dx,$$

and prove by performing the differentiations and direct computation that

$$I = \int_\Omega (\nabla\times u)\cdot n\,ds,$$

where $n\,ds = (-\frac{\partial f}{\partial x_1}, -\frac{\partial f}{\partial x_2}, 1)\,dx$.

229.2. Give a proof of the equality $\int_\Omega \nabla\times u\,dx = \int_\Gamma n\times u\,ds$, where Ω is a subset of $\mathbb{R}^3$ with boundary Γ with outward unit normal n, by applying the divergence theorem to $u\times a$ with a an arbitrary constant vector.

229.3. Study the relation between Green's formula (227.9), in the form (227.10), and the divergence theorem applied to the two-dimensional domain S with boundary Γ:

$$\int_S \left(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2}\right)\,dx_1\,dx_2 = \int_\Gamma (v_1n_1 + v_1n_2)\,ds\,,$$

with the identification $(u_1,u_2) = (-v_2, v_1)$ corresponding to counter clockwise rotation of the vector (v_1,v_2) by $\pi/2$. Explain how the clockwise direction in Stokes' theorem becomes a counter clockwise direction in (227.9).

229.4. Compute the integral

$$\int_\Gamma (x_1^2x_2, -x_1^3)/\|x\|^4\cdot ds,$$

where Γ is the curve in the x_1x_2-plane from $(1,0)$ to $(0,1)$ defined by $(x_1(t), x_2(t)) = (\cos(t)^{15}, \sin(t)^{17})$, $0 \le t \le \pi/2$.

229.5. Compute the integral

$$\int_\Gamma \frac{1}{x_1^2+x_2^2}\frac{(-x_2,x_1,x_3)}{\|x\|}\cdot ds,$$

where Γ is a curve traversing the unit circle in the x_1x_2-plane five times counterclockwise, then two times clockwise, and then again four times counterclockwise, as viewed from the positive x_3-axis.

229.6. Use Stokes' theorem to prove that

$$\int_\Gamma v\,ds = \int_S n\times\nabla v\,ds,$$

where S is a surface in $\mathbb{R}^3$ bounded by the closed curve Γ. Hint: Use Stokes' theorem with $u = va$ and a is a arbitrary vector in $\mathbb{R}^3$.

229.7. Verify by direct computation Stokes' theorem for (a) S the hemisphere $\{x\in\mathbb{R}^3 : \|x\| = 1,\, x_3\geq 0\}$ and $u = (x_2, 2x_3, 3x_1)$, (b) $S = \{x\in\mathbb{R}^3 : x_3 = 1-x_1^2-x_2^2,\, x_3\geq 0\}$.

229.8. (a) Let Ω be a domain in $\mathbb{R}^2$ with boundary Γ. Show that the area $A(\Omega)$ is given by the formula

$$A(\Omega) = \frac{1}{2}\int_\Gamma u\cdot ds,$$

where $u(x) = (-x_2, x_1)$ and Γ is oriented counter-clockwise. Use this result to show that the area bounded by the ellipse $x = (a\cos(t), b\sin(t))$, $0\leq t\leq 2\pi$, with half-axes a and b, is equal to πab. (b) Try to design a mechanical instrument for measuring the area of a domain in $\mathbb{R}^2$ (planimeter).

230
Potential Fields

He is a rather tall, lanky-looking man, with moustache and beard about to turn grey with a somewhat harsh voice and rather deaf. He was unwashed, with his cup of coffee and cigar. One of his failings is forgetting time, he pulls his watch out, finds it past three, and runs out without even finishing the sentence. (Thomas Hirst about Dirichlet 1850)

230.1 Introduction

We know from Chapter Curve integrals that potential force fields play an important role in mechanics. Let $u : \Omega \to \mathbb{R}^3$ be a given vector function, where Ω is a domain in $\mathbb{R}^3$. How can we check if $u(x)$ is a *potential field*, that is, if there is a scalar function or scalar *potential*, φ such that

$$u(x) = \nabla\varphi(x) \quad \text{for } x \in \Omega? \tag{230.1}$$

We recall that if $u = \nabla\varphi$ is a potential field and Γ is a curve parameterized by $s : [0,1] \to \mathbb{R}^3$ from $a = s(0)$ to $b = s(1)$, then the work of u along Γ is given by

$$\int_\Gamma u \cdot ds = \int_0^1 \nabla\varphi(s(t)) \cdot s'(t)\,dt = \int_0^1 \frac{d\varphi(s(t))}{dt}\,dt = \varphi(b) - \varphi(a).$$

In particular, the work is the same along all curves from a to b, and if the curve is closed with $\varphi(1) = \varphi(0)$ then the work performed when moving

around the curve is zero. A field with the property that the work along a closed curve is zero is referred to as a *conservative field*. A potential field is thus a conservative field.

A basic example of a gradient field is the gravitational field of a mass m at the origin,

$$u(x) = -m\frac{x}{\|x\|^3} = \nabla(\frac{m}{\|x\|}),$$

normalizing units so the gravitational constant is one. The electrical field of a charge m at the origin has the same form. In that case, the potential $\varphi(x) = m/\|x\|$ represents *potential energy* (gravitational or electrical), and curve integrals $\int_\Gamma u \cdot ds = \varphi(b) - \varphi(a)$ represents the work performed by a unit mass or charge when moved from a to b along Γ.

230.2 An Irrotational Field Is a Potential Field

We saw earlier that a potential field $u = \nabla\varphi$ is *irrotational*, that is $\nabla \times u = \nabla \times (\nabla\varphi) = 0$. This follows from a direct computation using $\frac{\partial^2\varphi}{\partial x_i \partial x_j} = \frac{\partial^2\varphi}{\partial x_j \partial x_i}$. In other words, $\nabla \times u = 0$ in Ω is a *necessary condition* for u to be a potential field in Ω.

We shall now prove that, the condition $\nabla \times u = 0$ in Ω is a *sufficient condition* for u to be a potential field in Ω, under the assumption that Ω is convex. We recall that Ω is *convex* if for any two points x and $\bar{x}$ in Ω, the entire line segment $\bar{x} + t(x - \bar{x}), 0 \le t \le 1$ between $\bar{x}$ and x, is also in Ω, see Fig. 230.1. Convexity implies in particular that Ω has "no holes". We

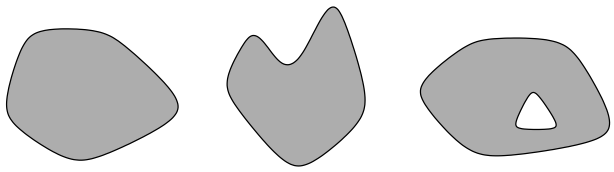


FIGURE 230.1. One convex and two non-convex domains.

thus conclude that u is a potential field in a convex domain Ω if and only u is irrotational in Ω. In other words, $u = \nabla\varphi$ in Ω for some potential φ if and only if $\nabla \times u = 0$ in Ω.

We carry out the proof by *constructing* a potential φ such that $\nabla\varphi = u$ for a given irrotational field $u(x)$ in the convex domain Ω. For the construction, we choose a fixed point $\bar{x}$ in Ω. For each point x, we let Γ_x be a curve in Ω connecting $\bar{x}$ to x and we define

$$\varphi(x) = \int_{\Gamma_x} u \cdot ds. \tag{230.2}$$

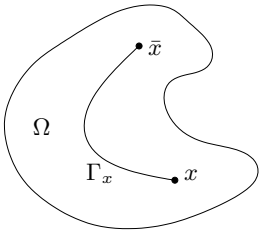


FIGURE 230.2. A curve Γ_x in Ω joining $\bar{x}$ and x.

We first prove that $\varphi(x)$ is independent of the choice of curve Γ_x from $\bar{x}$ to x. Assume that Γ_x and $\tilde{\Gamma}_x$ are two curves from $\bar{x}$ to x. Together they form a closed curve Γ bounding a surface S so Stokes theorem implies

$$\int_\Gamma u \cdot ds = \pm \int_S (\nabla \times u) \cdot n \, ds = 0,$$

since $\nabla \times u = 0$ on S. Now

$$\int_\Gamma u \cdot ds = \int_{\Gamma_x} u \cdot ds - \int_{\tilde{\Gamma}_x} u \cdot ds$$

if we orient Γ in the same direction as Γ_x and thus in the opposite direction as $\tilde{\Gamma}_x$. We conclude that

$$\int_{\tilde{\Gamma}_x} u \cdot ds = \int_{\Gamma_x} u \cdot ds,$$

and the independence of the choice of curve connecting $\bar{x}$ with x follows.

Next, we prove that the function $\varphi(x)$ defined by (230.2) satisfies $\nabla\varphi(x) = u(x)$ for $x \in \Omega$. We do this by choosing a curve Γ_x to connect to x along the x_1-axis, the x_2-axis, or the x_3-axis. Letting Γ_x connect along the x_1-axis according to Fig. 230.3, for $\hat{x}$ close to x we have

$$\varphi(x) - \varphi(\hat{x}) = \int_{\hat{x}_1}^{x_1} u_1(t, x_2, x_3) \, dt \,,$$

and the Fundamental Theorem implies

$$\frac{\partial \varphi}{\partial x_1}(x) = u_1(x).$$

Similarly, we obtain $\frac{\partial \varphi}{\partial x_i}(x) = u_i(x)$ for $i = 2, 3$. We summarize:

Theorem 230.1 *If $u : \Omega \to \mathbb{R}^d$, with Ω being a convex domain in $\mathbb{R}^d$ for $d = 2, 3$, satisfies $\nabla \times u(x) = 0$ for all $x \in \Omega$, then there is a function $\varphi : \Omega \to \mathbb{R}$ such that $u(x) = \nabla\varphi(x)$ for $x \in \Omega$.*

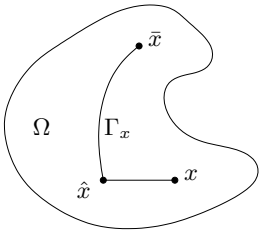


FIGURE 230.3. A curve Γ_x in Ω connecting to x along the x_1-axis.

230.3 A Counter-Example for a Non-Convex Ω

Consider the function $u : \Omega \to \mathbb{R}^2$, defined by $u(x) = (-x_2, x_1)/\|x\|^2$ with $\Omega = \{x \in \mathbb{R}^2 : x \neq 0\}$. This function satisfies

$$\nabla \times u(x) = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} = \frac{-2x_1x_2}{\|x\|^4} - \frac{-2x_1x_2}{\|x\|^4} = 0 \quad \text{for } x \in \Omega.$$

Nevertheless, $u(x)$ cannot be written in the form $u(x) = \nabla\varphi(x)$ for $x \in \Omega$. This follows by noting that if, for example, Γ is the closed circle given by $s(t) = r(\cos(t), \sin(t))$, $0 \leq t < 2\pi$, then

$$\int_\Gamma u \cdot ds = \int_0^{2\pi} \frac{1}{r^2} r^2 \, dt = 2\pi,$$

while if $u(x) = \nabla\varphi(x)$, $\int_\Gamma u \cdot ds = 0$ since Γ is closed. The reason is that in this case Ω *is not convex*. The point $x = 0$ does not belong to Ω and thus Ω has a "hole". We cannot extend Ω to include $x = 0$ since the function $u(x)$ is singular at $x = 0$ and in particular not Lipschitz continuous at $x = 0$.

Chapter 230 Problems

230.1. If possible, find a potential φ for (a) $u(x) = (x_1, x_2, x_3)$ (b) $u(x) = (x_3, x_1, x_2)$ (c) $u(x) = (x_2^2 - x_3, 2x_1x_2, 3x_3^2 - x_1)$.

230.2. We recall from above that $\nabla \times u = 0$ if and only if $u = \nabla\varphi$ for some φ.

We now ask the question if $\nabla \cdot u = 0$ if and only if $u = \nabla \times \psi$ for some (vector) potential ψ. Recall that we already know that the "if-part" of this is true, namely that $\nabla \cdot u = 0$ if $u = \nabla \times \psi$ for some ψ for some ψ.

It turns out that also the "only if-part" is true, that is, if $\nabla \cdot u = 0$ we may construct a (vector) potential ψ such that $u = \nabla \times |psi$. Verify this, using the construction $\psi(x) = \int_0^1 u(tx) \times tx \, dt$, and assuming $\nabla \cdot u$ in all of $\mathbb{R}^3$ for simplicity.

230.3. Extend the above counterexample to the function $u : \mathbb{R}^3 \to \mathbb{R}^3$ given by $u(x) = (-x_2, x_1, 0)/\|x\|^2$ representing the magnetic field around a current along the x_3-axis.

231

Center of Mass and Archimedes' Principle*

The simplest schoolboy is now familiar with facts for which Archimedes would have sacrificed his life. (Ernest Renan)

.

231.1 Introduction

We now turn to a study of the stability of floating bodies, including the question of how to design a big ship or a sailing boat so that it does not tip over. An example of an unfortunate design is given by the warship Vasa, which tipped over on its maiden voyage on August 10 1628 in the harbor of Stockholm and sank along with 50 of the crew of 150 people. In the resulting trial, it was decided that the ship was "well built, but badly proportioned" and no-one was held guilty for the disaster. The ship can now be studied at the Vasa museum in Stockholm.

Evidently, the stability properties of Vasa came as a surprise. Vasa had a new design with two gundecks with heavy artillery instead of one and the planned ballast of stone was not sufficient as a counterbalance. The old rules of ship design apparently did not apply to the new design and Calculus and scientific computing at that time was too primitive for trustworthy predictions.

Let's see what we can do today with a little bit of Calculus. We start with the concept of *center of mass*, pass on to *Archimedes principle* and the question of stability of floating bodies.

231.2 Center of Mass

Consider a body B occupying the volume V in $\mathbb{R}^3$. Suppose the *density* of the body at x is given by $\rho(x)$. The total mass $m(B)$ of the body is

$$m(B) = \int_V \rho(x)\,dx.$$

The *center of mass* $\bar{x} = (\bar{x}_1, \bar{x}_2, \bar{x}_3) \in \mathbb{R}^3$ of the body B is defined by

$$\bar{x}_i \int_V \rho(x)\,dx = \int_V x_i \rho(x)\,dx\,, \quad i = 1, 2, 3,$$

that is

$$\bar{x}_i = \frac{\int_V x_i \rho(x)\,dx}{\int_V \rho(x)\,dx}\,, \quad i = 1, 2, 3.$$

In vector form, this is

$$\bar{x} = \frac{\int_V x\rho(x)\,dx}{\int_V \rho(x)\,dx}.$$

We now explain the relevance of the concept of center of mass using the concept of *torque*. Assume the body B is acted upon by a vertical gravity force field $-e_3$ of unit strength with the coordinate direction e_3 oriented vertically upward. The torque about a point $\bar{x}$ of a force F acting at x is equal to

$$(x - \bar{x}) \times F = -F \times (x - \bar{x}),$$

see Fig. 231.1. In other words, the torque is a vector that is perpendicular to the plane generated by the direction of the force F and the lever arm $x - \bar{x}$, with modulus equal to the modulus of F times the distance of the point $\bar{x}$ from the line of action of F.

The torque of the gravity field (assuming the acceleration of gravity $g = 1$) acting on an element of mass $\rho(x)\,dx$ at position x about a given point $\bar{x}$ is equal to

$$\rho(x)\,dx\, e_3 \times (x - \bar{x}).$$

The total torque T of the gravity field $-e_3$ on the body B about $\bar{x}$ is thus equal to

$$T = e_3 \times \int_V \rho(x)(x - \bar{x})\,dx = 0,$$

by the definition of the center of mass $\bar{x}$. The torque about $\bar{x}$ thus vanishes which means that body will balance if supported at $\bar{x}$, see Fig. 231.2.

More precisely,

$$T = e_3 \times \Big(\int_V \rho(x)x\,dx - \bar{x}\int_V \rho(x)\,dx\Big) = 0 \tag{231.1}$$

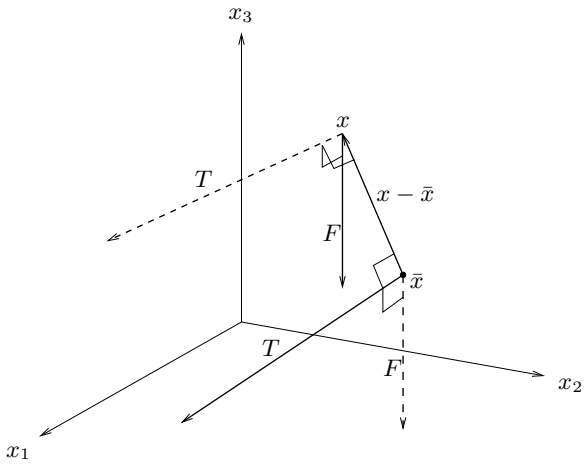


FIGURE 231.1. The torque $T = (x - \bar{x}) \times F$ about the point $\bar{x}$ of F acting at x.

if and only if

$$\bar{x}_i = \frac{\int_V x_i \rho(x)\, dx}{\int_V \rho(x)\, dx},$$

for $i = 1, 2$. This means that the body will balance if supported at a point $x = (x_1, x_2, x_3)$ with $x_1 = \bar{x}_1$ and $x_2 = \bar{x}_2$, while x_3 may be chosen arbitrarily, see Fig. 231.2. Thus, if the body is supported at its center of mass $\bar{x}$ then it will balance independently of its orientation. If the body is supported at a point x different from the center of mass $\bar{x}$, then it will balance only if $\bar{x} - x$ is parallel to the direction of the gravity field.

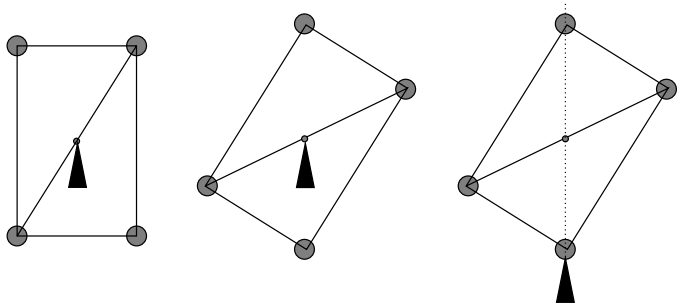


FIGURE 231.2. A body supported at its center of mass, in two stable positions, and a body supported at a boundary point, balanced but unstable.

EXAMPLE 231.1. We compute the center of mass $\bar{x}$ of a thin triangular plate of uniform thickness occupying the region $\Omega = \{x \in \mathbb{R}^2 : 0 \leq$

$x_1, x_2, x_1 + x_2 \leq 1\}$ in the plane. We get

$$\bar{x}_i = \frac{\int_\Omega x_i \, dx}{\int_\Omega dx} = \frac{1/6}{1/2} = \frac{1}{3}.$$

EXAMPLE 231.2. We compute the center of mass of the half-ball $\Omega = \{x \in \mathbb{R}^3 : \|x\| \leq 1, \, x_3 \geq 0\}$. By symmetry $\bar{x}_1 = \bar{x}_2 = 0$. For $\bar{x}_3$ we get using spherical coordinates

$$\begin{aligned}\int_\Omega x_3 \, dx &= \int_0^{2\pi} \int_0^{\pi/2} \int_0^1 r\cos(\varphi)\, r^2 \sin(\varphi)\, dr d\varphi d\theta \\ &= \int_0^{2\pi} \int_0^{\pi/2} \frac{1}{2}\sin(2\varphi)\Big[\frac{1}{4}r^4\Big]_0^1 \, d\varphi d\theta \\ &= \frac{1}{4}\int_0^{2\pi} \Big[\frac{1}{4}\cos(2\varphi)\Big]_0^{\pi/2} d\theta = \frac{1}{4}\int_0^{2\pi} d\theta = \pi/4,\end{aligned}$$

that is, $\bar{x}_3 = \int_\Omega x_3 \, dx / \int_\Omega dx = \frac{\pi/4}{2\pi/3} = \frac{3}{4}$.

231.3 Archimedes' Principle

Archimedes principle states that (i) the *buoyancy force* acting on a body B totally immersed in a liquid is equal to the weight of the displaced liquid and (ii) acts along a vertical line through the center of mass of the displaced fluid, which we refer to as the *center of bouyancy* c_b. We shall now prove this fact using vector Calculus. The force from the fluid acting on an element $ds = ds(x)$ of the surface S of the body B at position x is equal to $-p(x)n(x)\, ds$, where $p(x)$ is the pressure of the liquid and $n(x)$ is the outward (from B) unit normal to S at x. The total pressure force on B is thus

$$F = -\int_S p(x)n(x)\, ds(x) = -\int_S pn \, ds.$$

Since

$$\int_V \frac{\partial p}{\partial x_i}\, dx = \int_S p n_i \, ds, \quad i = 1, 2, 3,$$

where V is the volume occupied by B, we have

$$F = -\int_V \nabla p(x)\, dx$$

The *pressure* $p(x)$ in a fluid at rest, called the *hydro-static pressure*, is given by

$$p(x) = \rho z(x) + p_0,$$

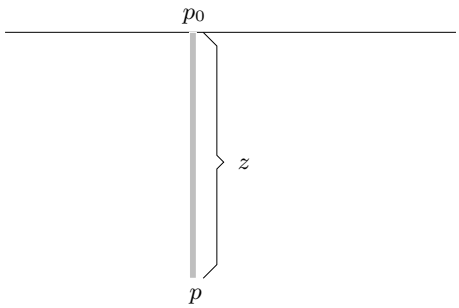


FIGURE 231.3. Hydrostatic pressure $p(x) = \rho z(x) + p_0$.

where $z(x)$ is the depth, ρ is the constant density of the fluid and p_0 is the pressure on the surface of the fluid, see Fig. 231.3.

The pressure force at a point x is equal in all directions and its modulus $p(x)$ equal is to the weight $\rho z(x)$ of the column of fluid above the point x plus the surface pressure p_0 from the atmosphere. We conclude that

$$\nabla p(x) = -\rho e_3,$$

where we assume that the coordinate direction e_3 is vertical and pointed upwards. Therefore,

$$F = \int_V \rho \, dx \, e_3 \equiv W e_3,$$

where $W = \int_V \rho \, dx$ is the total weight of the displaced fluid. This proves the first part of Archimedes principle.

Next, the total torque T from the fluid pressure forces on S about a point $\bar{x}$ is given by

$$T = \int_S (x - \bar{x}) \times (-p(x)n(x)) \, ds(x) = \int_S n(x) \times p(x)(x - \bar{x}) \, ds.$$

Recalling that

$$\int_S n \times F \, ds = \int_V \nabla \times F \, dx,$$

we find that

$$T = \int_V \nabla \times (p(x)(x - \bar{x})) \, dx.$$

But,

$$\nabla \times (p(x)(x - \bar{x})) = \nabla p \times (x - \bar{x}) + p \, \nabla \times (x - \bar{x}).$$

Since $\nabla \times (x - \bar{x}) = 0$, it follows that

$$T = \int_V \nabla p \times (x - \bar{x}) \, dx = -\int_V \rho (x - \bar{x}) \, dx \times e_3$$

and the torque T vanishes if $\bar{x}$ satisfies

$$\bar{x}_i \int_V \rho \, dx = \int_V x_i \rho \, dx \quad \text{for } i = 1, 2.$$

We conclude that the buoyancy force is vertical upward and is acting along a vertical line through the center of mass of the displaced fluid. We have now proved:

Theorem 231.1 (Archimedes' principle) *The buoyancy force acting on a body immersed in a liquid is equal to the weight of the displaced liquid and acts along a vertical line through the center of mass of the displaced fluid.*

We can directly extend Archimedes' principle to a partially immersed body assuming the pressure on the surface of the fluid to be is zero.

231.4 Stability of Floating Bodies

The stability of a floating body B is of central importance in all forms of boating, from canoes to big ships. The question of stability can be reduced to a question of the relative position of (i) the center of mass c_m of the body B and (ii) the center of buoyancy c_b of B according to the following discussion. Consider the body in rest position with the gravity force acting vertically downward from the center of gravity, and the buoyancy force acting vertically upward from the center of buoyancy. We assume the body is in equilibrium with the gravity force and the buoyancy force balancing and acting along the vertical line through the centers of gravity and buoyancy, see Fig. 231.4.

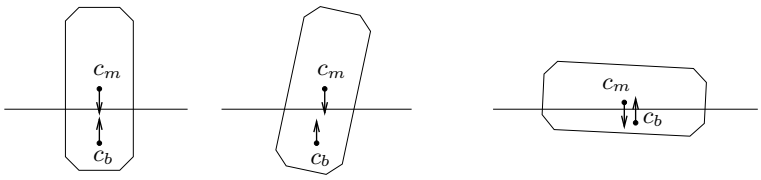


FIGURE 231.4. Floating bodies with centers of gravity and buoyancy.

Assume that the body is tilted a small angle so that the centers of gravity and buoyancy are displaced horizontally, see Fig. 231.4. Let T be the resulting torque from the pair of gravity and buoyancy forces. The sign of T will govern the stability! If T acts in the same direction as the tilting, then the tendency of tilting will be enforced and the body will depart from its equilibrium position and eventually tilt over, see Fig. 231.4. This happens if

the center of gravity is displaced horizontally in the direction of tilting more quickly than the center of buoyancy. Conversely, if the center of gravity is displaced more slowly, then the resulting torque T will be negative and act as a restoring force seeking to bring back the body to the rest position, see Fig. 231.4. We now consider two examples with simple geometry.

EXAMPLE 231.3. Consider a space capsule with hemispherical base and conical top floating in the Pacific and waiting to be recovered. Will the capsule float upright or not? Assuming the capsule is floating upright with a part of the hemispherical base immersed into the water, see Fig. 231.5.

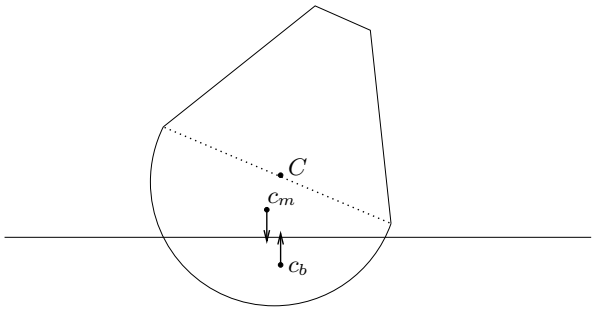


FIGURE 231.5. Space capsule floating upright.

The resultant of the buoyancy forces is directed upward and acts through the center C of the hemisphere, see Fig. 231.5. If the capsule is tilted a little, the resultant of the buoyancy forces is still directed upwards through C and the torque from the gravity force will be de-stabilizing if the center of mass c_m of the capsule is positioned above C, and stabilizing if c_m is below C, for c_m on the symmetry axis of the capsule, see Fig. 231.5.

EXAMPLE 231.4. Consider a rectangular box with square horizontal cross section of width $2w$ and height $2h$ and density $\bar{\rho}$ which is floating in a fluid of density ρ, see Fig. 231.6. Suppose that $\bar{\rho}$ is small compared to ρ so that it penetrates into the fluid only slightly. To test the stability of the box, suppose the box is rotated a small angle θ around the mid-point C at the bottom. The de-stabilizing torque about C resulting from the gravity force through the center of gravity is equal to $g\bar{\rho}(2w)^2 2hh\sin(\theta)$, see Fig. 231.6. The stabilizing torque from the change of buoyancy forces caused by the rotation is equal to

$$2\frac{2}{3}www\sin\theta\frac{1}{2}\rho gw,$$

because the area of the triangle CAB is equal to $ww\sin\theta\frac{1}{2}$, and the center of gravity of CAB is at horizontal distance $2\frac{2}{3}w$ from C. The position is stable if

$$2\frac{2}{3}w^4\sin\theta\frac{1}{2}\rho g > g\bar{\rho}8w^2h^2\sin(\theta),$$

that is, if

$$w^2\rho > 12h^2\bar{\rho}.$$

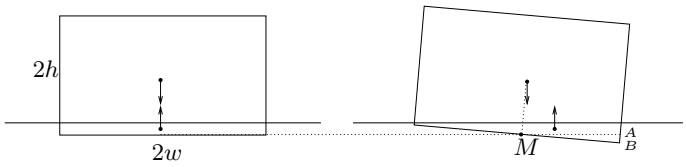


FIGURE 231.6. Floating box.

Chapter 231 Problems

231.1. The density of ice is 0.917 times the density of water (at $-4°$C). How large a part of an iceberg is visible above the water surface?

231.2. How does a log float? Why does it not want to float in an upright position?

231.3. Understand why catamaran ships have good stability properties.

231.4. Find the stable floating position of a "log" with a quadratic cross-section and density $\bar{\rho} = \frac{1}{2}\rho$. Find the stable positions as a function of the ratio $\bar{\rho}/\rho$ (we know from above that for $\bar{\rho}/\rho$ sufficiently small it will float as the box in Fig. 231.6). May there be more than one stable position (disregarding symmetric ones). Discuss! May the conclusion depend on the shape of the cross-section?

231.5. How does a (perfect) ice cube float? How does a barrel (cylinder) float, given hight/diameter ratio and density?

231.6. Study the design of sailing boats from the stability point of view. Study in particular modern designs with good form stability (wide and flat bottom), and classical designs with a narrow deep hull. Connect to the discussion above.

231.7. Extend Archimedes principle to a body immersed into a system of two layers of different fluids on top of each other.

232
Laplacian Models

> ... on aura donc $\Delta u = 0$; cette équation remarquable nous sera de la plus grande utilité .... (Laplace)
>
> If one has to stick to this damned quantum jumping, then I regret ever having been involved in this thing. I don't like it (quantum mechanics), and I'm sorry I ever had anything to do with it. (Schrödinger)

232.1 Introduction

In this chapter, we present some basic models involving the Laplacian, including models for heat conduction, elasticity, electromagnetics, fluid mechanics, and gravitation. In deriving these models, we make use of the basics of Calculus in several dimensions including Gauss' and Stokes' theorems, and we get a quick and easy introduction to some of mysteries of the mechanics and physics of "continuous media". We also make connections to linear algebra when discretizing the Laplacian using the 5-point scheme and variants of "Svensson's formula".

232.2 Heat Conduction

We first model *heat conduction* in a heat-conducting material occupying the volume Ω in $\mathbb{R}^3$ with boundary Γ, over a time interval $I = [0, T]$. We

let $u(x,t)$ denote the *temperature* and $q(x,t)$ the *heat flux* at the point x at time t. The heat flux is a vector $q = (q_1, q_2, q_3)$, where q_i is the heat flux, or rate of heat flowing in the direction x_i. We let $f(x,t)$ denote the rate of heat (per unit of volume) supplied at (x,t) by a *heat source*.

We derive the model using a basic *conservation law* expressing *conservation of heat* in the following form: for any fixed domain V in Ω with boundary S, the rate of the total heat introduced in V by the external source is equal to the rate of the total heat accumulated in V plus the total heat flux through S. This is based on the conviction that the heat introduced in V by the external source can choose between two options only: (i) flow out of V or (ii) be accumulated in V. With S denoting the boundary of V and n denoting the outward unit normal to S, see Fig. 232.1, the conservation law can be expressed as

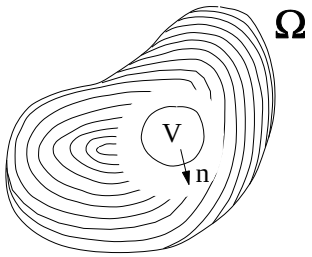


FIGURE 232.1. An arbitrary subset V of a heat conducting body Ω.

$$\int_V f\,dx = \frac{\partial}{\partial t}\int_V \lambda u\,dx + \int_S q\cdot n\,ds, \tag{232.1}$$

where $\lambda(x,t)$ is the *heat capacity coefficient* giving the amount heat per unit of volume needed to raise the temperature one unit, and all functions are evaluated at a specific time $t \in I$. By the Divergence theorem,

$$\int_S q\cdot n\,ds = \int_V \nabla\cdot q\,dx,$$

and combined with (232.1), this implies that

$$\int_V \left(\frac{\partial}{\partial t}(\lambda u) + \nabla\cdot q\right) dx = \int_V f\,dx,$$

where the time derivative could be moved under the integral sign because V does not depend on time t. Since V is arbitrary, assuming the integrands are Lipschitz continuous, it follows that

$$\frac{\partial}{\partial t}(\lambda u)(x,t) + \nabla\cdot q(x,t) = f(x,t) \quad \text{for all } x\in\Omega,\ 0<t\le T, \tag{232.2}$$

which is a differential equation describing *conservation of heat* involving two unknowns: the temperature $u(x,t)$ and the heat flux $q(x,t)$. We thus have one equation and two unknowns and we need yet another equation.

The second equation is a *constitutive equation* that couples the heat flux q to the temperature gradient ∇u. *Fourier's law* states that heat flows from warm to cold regions with the heat flux proportional to the temperature gradient:

$$q(x,t) = -a(x,t)\nabla u(x,t) \quad \text{for } x \in \Omega,\ 0 < t \leq T \tag{232.3}$$

where the factor of proportionality $a(x,t)$ is the coefficient of heat conductivity. Note the minus sign indicating that the heat flows from warm to cold regions, and that the heat conductivity $a(x,t)$ is positive. Combining (232.2) and (232.3), we obtain the basic differential equation describing heat conduction:

$$\frac{\partial}{\partial t}(\lambda u) - \nabla \cdot (a\nabla u) = f \quad \text{in } \Omega \times (0,T], \tag{232.4}$$

where $a(x,t)$ and $\lambda(x,t)$ are given positive coefficients depending on (x,t) and $f(x,t)$ is a given heat source, and the unknown $u(x,t)$ represents the temperature.

To define the solution uniquely, the differential equation is complemented by initial and boundary conditions. The complete model with *Dirichlet boundary conditions* reads

$$\begin{cases} \frac{\partial}{\partial t}(\lambda u) - \nabla \cdot (a\nabla u) = f & \text{in } \Omega \times (0,T], \\ u = u_b & \text{on } \Gamma \times (0,T], \\ u(x,0) = u_0(x) & \text{for } x \in \Omega, \end{cases} \tag{232.5}$$

where u_0 is the initial temperature and u_b is the boundary temperature. The Dirichlet boundary condition corresponds to immersing the body Ω in a large reservoir with a specified temperature u_b and assuming that the boundary acts as a perfect thermal conductor so that the temperature of the body on the boundary is equal to the specified outside reservoir temperature u_b. Note that the given boundary temperature $u_b = u_b(x,t)$ may vary with (x,t).

Other commonly encountered boundary conditions are *Neumann* and *Robin* boundary conditions. A Neumann boundary condition corresponds to prescribing the heat flux $q \cdot n$ across (out of) the boundary:

$$q \cdot n = -a\nabla u \cdot n = -a\frac{\partial u}{\partial n} = -a\partial_n u = g \quad \text{on } \Gamma,$$

with g given. A *homogeneous Neumann boundary condition* with $g = 0$ corresponds to a perfectly insulating boundary with the heat flux across the boundary being zero. A homogenous Robin boundary condition is intermediate with the boundary neither being perfectly conducting nor perfectly

insulated, with the heat flux through the boundary being proportional to the difference of the temperature u inside and a given temperature u_b outside Ω:

$$-a\partial_n u = \kappa(u - u_b)$$

with κ a positive coefficient representing the heat conductivity of the boundary.

Partitioning the boundary Γ into disjoint pieces Γ_1, Γ_2 and Γ_3 with different types of boundary conditions, the *general initial boundary value problem* IBVP for the heat equation has the form,

$$\begin{cases} \frac{\partial}{\partial t}(\lambda u) - \nabla\cdot(a\nabla u) = f & \text{in } \Omega\times(0,T],\\ u = u_b & \text{on } \Gamma_1\times(0,T],\\ -a\partial_n u = g & \text{on } \Gamma_2\times(0,T],\\ a\partial_n u + \kappa(u-u_b) = 0 & \text{on } \Gamma_3\times(0,T],\\ u(x,0) = u_0(x) & \text{for } x\in\Omega, \end{cases} \tag{232.6}$$

where u_b represents a given "exterior" boundary temperature, and g represents a given outward normal heat flux on the boundary.

We note that in a stationary situation with $\frac{\partial}{\partial t}(\lambda u) = 0$ and with the heat source $f = 0$, the equation (232.2) expressing conservation of heat, takes the form

$$\nabla\cdot q = 0. \tag{232.7}$$

If heat is neither produced nor accumulated, then conservation of heat is expressed by the equation $\nabla\cdot q = 0$, that is, the heat flux q is *divergence-free*. Below we shall meet several other examples of divergence-free fields.

232.3 The Heat Equation

We refer to the special case of (232.6) with $\lambda = a = 1$ as the *heat equation*. In the case with homogeneous Dirichlet boundary conditions, we get the model

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & \text{in } \Omega\times(0,T],\\ u = 0 & \text{on } \Gamma\times(0,T],\\ u(x,0) = u_0(x) & \text{for } x\in\Omega, \end{cases}$$

where u_0 is the initial temperature, and $\Delta u = \nabla\cdot(\nabla u)$ is the Laplacian. The heat equation serves as a basic prototype of a *parabolic problem*.

FIGURE 232.2. Poisson (1781-1840): "Life is good for only two things: to study mathematics and to teach it"

232.4 Stationary Heat Conduction: Poisson's Equation

The stationary analog of (232.6) reads

$$\begin{cases} -\nabla \cdot (a\nabla u) = f & \text{in } \Omega, \\ u = u_b & \text{on } \Gamma_1, \\ -a\partial_n u = g & \text{on } \Gamma_2, \\ a\partial_n u + \kappa(u - u_b) = 0 & \text{on } \Gamma_3. \end{cases} \tag{232.8}$$

Choosing $a = 1$ leads to the *Poisson equation*:

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = u_1 & \text{on } \Gamma_1, \\ -a\partial_n u = g_2 & \text{on } \Gamma_2, \\ a\partial_n u + \kappa(u - u_b) = g_3 & \text{on } \Gamma_3. \end{cases} \tag{232.9}$$

In the case of homogeneous Dirichlet boundary conditions on the whole of the boundary, the Poisson equation reads

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma. \end{cases} \tag{232.10}$$

Poisson's equation serves as a basic model of an *elliptic* problem and has numerous applications in physics and mechanics. We present the basic applications below. Poisson's equation $-\Delta u = f$ with $f = 0$ is referred to as *Laplace's equation*: $\Delta u = 0$.

We now give a couple of analytic solutions to the heat equation in simple situations:

EXAMPLE 232.1. The stationary temperature u in a heat conduction unit cube Q with heat production and conduction coefficient equal to one, zero boundary temperature for $x_1 = 0, 1$, and zero heat flux for $x_2, x_3 = 0, 1$, is given by

$$u(x) = \frac{1}{2}(x_1(1 - x_1).$$

We see that the temperature is maximal for $x_1 = 0.5$ and drops off quadratically towards the Dirichlet boundary, see Fig. 232.3 for a plot in the corresponding case in two dimension in the unit square.

EXAMPLE 232.2. Consider the homogenous heat equation in the unit square Q with $f = 0$ and homogenous Dirichlet boundary conditions: the function

$$u(x,t) = e^{-(n^2+m^2)t}\sin(nx_1)\sin(mx_2)$$

with $m, n = 1, 2, 3, ...$ is a solution of the homogenous heat equation $\frac{\partial u}{\partial t} - \Delta u = 0$ with initial value $u_0(x_1, x_2) = \sin(nx_1)\sin(mx_2)$, see Fig. 232.3. We see that the temperature $u(x,t)$ decays exponentially in time very quickly if n and/or m is only moderately large. This corresponds to the fact that a temperature oscillating in space, is quickly levelled out.

EXAMPLE 232.3. The stationary temperature $u(x)$ between the two planes $\{x_3 = 0\}$ and $\{x_3 = 1\}$ bounding a heat conducting layer with heat conductivity coefficient equal to one, zero heat source, and the temperature $u = 1$ on $\{x_3 = 1\}$ and $u = 0$ on $\{x_3 = 0\}$, is given by $u(x) = x_3$ displaying a linear variation of the temperature between the plates. No surprise of course.

232.5 Convection-Diffusion-Reaction

The heat equation models the physical phenomenon of *diffusion*, and we now extend this model to include the phenomena of *convection* and *reaction*. We obtain a scalar *convection-diffusion-reaction* equation, which is another basic model in science. We consider a typical case where u represents the concentration of a certain chemical species subject to convection in a given velocity field $\beta(x,t)$, diffusion with diffusion coefficient $\epsilon(x,t)$ and reaction with reaction rate $\alpha(x,t)$. For example, u may represent the concentration of a contaminant in a volume of water moving with the velocity $\beta(x,t)$.

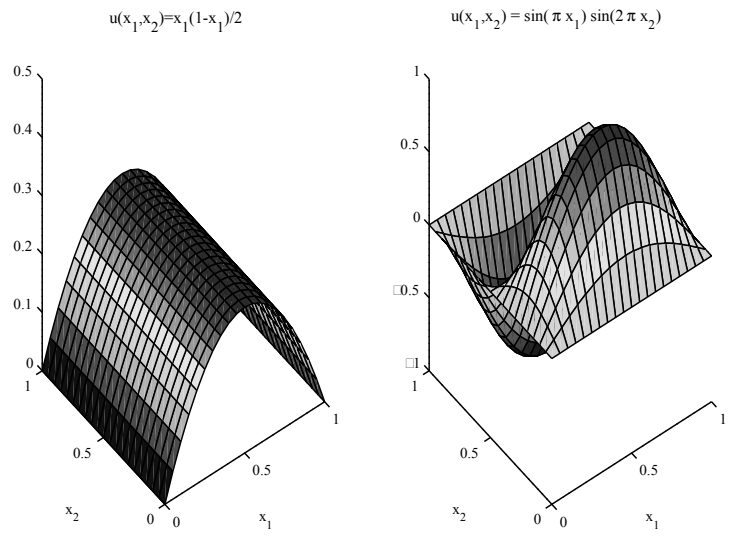


FIGURE 232.3. The functions $\frac{1}{2}(x_1(1-x_1)$ and $sin(\pi x_1)sin(2\pi x_2)$

The model results from a principle of conservation of mass together with a constitutive equation generalizing Fourier's law expressing the *flow rate* q of the chemical species in terms of ∇u and βu. Conservation of mass is expressed by

$$\dot{u} + \nabla \cdot q + \alpha u = f,$$

where f represents a source term, and the constitutive law takes the form

$$q = \beta u - \epsilon \nabla u.$$

which says that the total flow rate q is the sum of a convective rate βu and a diffusive rate $-\epsilon \nabla u$. The model thus takes the form:

$$\dot{u} + \nabla \cdot (\beta u) + \alpha u - \nabla \cdot (\epsilon \nabla u) = f \quad \text{in } \Omega \times (0, T], \tag{232.11}$$

together with initial and boundary conditions, where Ω is domain in space and $[0, T]$ a given time interval. We shall meet this model and generalizations thereof in several different contexts below.

232.6 Elastic Membrane

Consider a horizontal elastic net covering the unit square $Q = \{x \in \mathbb{R}^2 : 0 \leq x_i \leq 1, i = 1, 2\}$ formed by elastic strings tied together at nodes $a_{ij} \in \mathbb{R}^2$ in a uniform quadrilateral mesh with mesh size $h = 1/N$, so that $a_{ij} = (ih, jh)$, $i, j = 0, 1, ..., N$, where N is the number of cells in each

coordinate direction. Assume that the net is stretched so that the tension in each string is equal to h, corresponding to the tension being equal to one per unit of length. Note the normalization introduced says that the tension in each string decreases as the number of strings increases. We refer to the situation in which all the nodes lie in the plane of the square and there is no external load on the net as the unloaded reference configuration of the net.

Suppose the net is subject to a set of downward vertical loads of size $f_{ij}h^2$ at the nodes a_{ij}. The net will deform under the loads and the nodes will be displaced from the initial unloaded reference configuration. Let the vertical displacement of node a_{ij} be denoted by $u_{i,j}$. If the displacements are small, then (recalling the Chapter String theory) the vertical upward force from the net on node a_{ij} is equal to

$$(u_{i,j} - u_{i-1,j}) + (u_{i,j} - u_{i+1,j}) + (u_{i,j} - u_{i,j-1}) + (u_{i,j} - u_{i,j+1}),$$

with contributions from the four pieces of string meeting at a_{ij}. This is because the vertical slope of the line between for example node (i, j) and $(i-1, j)$ is equal to $(u_{i,j} - u_{i-1,j})/h$ and the tension is h. We thus obtain the following vertical equilibrium equation for each node a_{ij}:

$$-\frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} - \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h^2} = f_{ij}.$$

Passing to the limit as h tends to zero, and recalling that Taylor's theorem implies

$$\lim_{h\to 0} \frac{v(x-h) - 2v(x) + v(x+h)}{h^2} = v''(x) = \frac{d^2v}{dx^2}(x)$$

if $v : \mathbb{R} \to \mathbb{R}$ is twice differentiable, we are led to the equation

$$-\Delta u(x) = f(x).$$

This equation expresses the equilibrium of a horizontal membrane made by an elastic fabric and carrying a vertical load of intensity (force per unit area) $f(x)$, where $u(x)$ is the vertical displacement of the membrane at x and we assume that the membrane in its unloaded plane reference configuration is prestressed to uniform tension in all directions.

We can generalize to a horizontal membrane covering a general domain Ω in $\mathbb{R}^2$. Assuming the membrane is fixed at the boundary Γ of Ω, so that the vertical displacement $u(x)$ is zero at Γ, we thus obtain Poisson's equation

$$-\Delta u = f \quad \text{in } \Omega,\ u = 0 \quad \text{on } \Gamma \tag{232.12}$$

as a model for the vertical deflection of a horizontal elastic membrane spanned over the boundary Γ of a domain Ω in $\mathbb{R}^2$, subject to a vertical load of intensity $f(x)$. This is a basic model of elasticity theory.

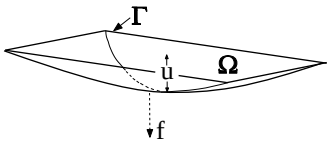


FIGURE 232.4. An elastic membrane under a load $f(x)$ and supported at Γ.

EXAMPLE 232.4. If $\Omega = \{x \in \mathbb{R}^2 : \|x\| < 1\}$ is the unit disc, and the load f is radially symmetric, then the deflection u will also be radially symmetric. Recalling the form of the Laplacian in polar coordinates from the Chapter The divergence, rotation and Laplacian, we can write (232.12) in the form

$$-\Delta u = -\frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial u}{\partial r}) = f(r) \quad \text{for } 0 < r < 1,\ u(1) = 0, \frac{\partial u}{\partial r}(0) = 0.$$

Note the boundary condition $\frac{\partial u}{\partial r}(0) = 0$, which has no counterpart in x-coordinates, says that $u(x)$ is differentiable at $x = 0$. If $\frac{\partial u}{\partial r}(0) \neq 0$, then $u(x)$ has a a conical "to" at $x = 0$ and thus is not differentiable at $x = 0$. If $f(r) = 1$, then the solution is given by

$$u(r) = \frac{1}{4}(1 - r^2) \quad \text{for } 0 \leq r \leq 1.$$

232.7 Solving the Poisson Equation

Suppose we would like to numerically solve the Poisson equation

$$-\Delta u = f \quad \text{in } Q, \quad u = u_b \quad \text{on } \Gamma$$

where Q is the unit square with boundary Γ and $f(x)$ a given function on Q. Recalling the derivation of the model $-\Delta u = f$ from the previous section, we are led to computing approximations $U_{i,j}$ of $u(ih, jh)$ for $i, j = 0, 1, ..., N$, where $h = 1/N$, from the system of equations

$$-\frac{U_{i-1,j} - 2U_{i,j} + U_{i+1,j}}{h^2} - \frac{U_{i,j-1} - 2U_{i,j} + U_{i,j+1}}{h^2} = f(ih, jh), \quad i, j = 1, ..., N-1,$$

that is

$$4U_{i,j} - U_{i-1,j} - U_{i+1,j} - U_{i,j-1} - U_{i,j+1} = h^2 f(ih, jh), \quad i, j = 1, ..., N-1, \quad (232.13)$$

where $U_{i,j} = u_b(ih, jh)$ if i or j is equal to 0 or N. We see that this is an $m \times m$ system of equations with $m = (N-1)\times(N-1)$ in the unknowns $U_{i,j}$ where $i, j = 1, ..., N-1$. This is the famous *5-point scheme* for the Poisson's equation, where the unknown $U_{i,j}$ is coupled to its four neighbors $U_{i-1,j}$, $U_{i+1,j}$, $U_{i,j-1}$, $U_{i,j+1}$.

If $f = 0$, then the 5-point scheme takes the form ("Svensson's formula")

$$U_{i,j} = \frac{1}{4}(U_{i-1,j} + U_{i+1,j} + U_{i,j-1} + U_{i,j+1}),$$

stating that each value $U_{i,j}$ is the mean value of the neighboring values (reflecting a basic feature of the Swedish national character).

Note that (232.13) is a linear system of equations for the values of U that requires some work to solve. For example, we may try to solve (232.13) by fixed point iteration as follows with $k = 0, 1, ...$

$$U_{i,j}^{k+1} = U_{i,j}^k - \alpha\big(4U_{i,j}^k - U_{i-1,j}^k - U_{i+1,j}^k - U_{i,j-1}^k - U_{i,j+1}^k - h^2 f(ih, jh)\big), \quad (232.14)$$

for $i, j = 1, ..., N-1$, with $U_{i,j}^{k+1} = u_b(ih, jh)$ if i or j is equal to 0 or N. Here, $U_{i,j}^k$ is an approximation of $U_{i,j}$ after k steps starting with an initial approximation U_{ij}^0 and α is a positive constant. It turns out that if α is sufficiently small, then the iteration converges, see Problem 232.9, although the convergence gets slower as the step size h decreases.

EXAMPLE 232.5. Assuming x_2 independence, we are led to the model

$$-u''(x) = f(x) \quad \text{for } 0 < x < 1, \quad u(0) = u_0,\ u(1) = u_1,$$

where $u'(x) = \frac{du}{dx}$. The corresponding discrete model takes the form

$$-(U_{i-1} - 2U_i + U_{i+1}) = h^2 f(ih), \quad i = 1, ..., N-1,\ U_0 = u_0,\ U_N = u_1, \quad (232.15)$$

with U_i representing an approximation of $u(ih)$. Assuming for simplicity $u_0 = u_1 = 0$, the discrete model can be written in the form

$$AU = b,$$

with $U = (U_1, ..., U_{N-1})$, $b = (b_1, ..., b_{N-1})$ with $b_i = h^2 f(ih)$, $A = (a_{ij})$ an $(N-1)\times(N-1)$ matrix with $a_{ii} = 2$, $a_{i,i-1} = a_{i-1,i} = -1$ and $a_{ij} = 0$ if $|i-j| > 1$. The fixed point iteration described above can be written

$$U^{k+1} = U^k - \alpha(AU^k - b),$$

and the criterion of convergence is $\|I - \alpha A\| < 1$, which we prove in Problem 232.9 to be valid if $\alpha > 0$ is sufficiently small. Here, $\|I - \alpha A\|$

is the Euclidean norm of the matrix $I - \alpha A$ and the Spectral theorem implies

$$\|I - \alpha A\| = \max_i |1 - \alpha\lambda_i|,$$

where the λ_i are the eigenvalues of the symmetric matrix A.

232.8 The Wave Equation: Vibrating Elastic Membrane

We now model the dynamic motion of the elastic membrane considered above in the static. We complement the given exterior force $f(x,t)$, which now may be dependent on time, by a dynamic force, which according to Newton's law, takes the form $m\ddot{u}$, with m representing mass per unit area and $\ddot{u}$ representing the acceleration of vertical displacement u. This to leads to the *wave equation*, modeling a vibrating membrane subject to an exterior load,

$$\begin{cases} \ddot{u} - \Delta u = f & \text{in } \Omega \times (0,T], \\ u = 0 & \text{on } \Gamma \times (0,T], \\ u(x,0) = u^0(x),\ \dot{u}(x,0) = \dot{u}^0(x) & \text{for } x \in \Omega, \end{cases} \tag{232.16}$$

where Ω denotes a domain in $\mathbb{R}^d$ with boundary Γ, u^0 is a given initial displacement, $\dot{u}^0$ is a given initial displacement velocity, and we assume homogeneous Dirichlet boundary conditions for simplicity. Other boundary conditions, notably periodic boundary conditions, are also relevant for this model.

232.9 Fluid Mechanics

Fluid flow opens a rich field for mathematical modelling. We think of a fluid as a collection of many small "fluid particles" and we seek to describe the fluid flow resulting from the motion of all these fluid particles. We work under the assumption that the particles are so small and there are so many, that we can treat the fluid as a continuum. Usually, we use an *Eulerian* mode of description in which we describe the flow in terms of the *velocity* $u(x,t) \in \mathbb{R}^3$ of the fluid particles at position $x \in \mathbb{R}^3$ at time t, or simply the velocity of the fluid at (x,t). This corresponds to attaching an observer to each fixed point x for the purpose of observing the velocity $u(x,t)$ of the fluid particles that happen to be at position x at time t. The observer thus sits at position x and watches the fluid particles swirl by.

Alternatively, in a *Lagrangian* mode of description, an observer is attached to each fluid particle with the purpose of observing the change of

velocity of the fluid particle with time. In this case, the observer follows the particle. The different modes of description are both useful and may also be used together, see the chapters on convection-diffusion in [**?**].

The Equation of Mass Conservation

We consider the flow of a fluid within a certain volume $\Omega \in \mathbb{R}^3$ using an Eulerian description with $u(x,t)$ representing the velocity of the fluid at x at time t. The velocity u is a vector $u = (u_1, u_2, u_3)$.

Let $\rho(x,t)$ denote the *density* of a fluid at (x,t) measuring the mass of the fluid particles per unit of volume. Let V be a fixed volume with boundary S. The total mass of the fluid in V at time t is given by

$$\int_V \rho(x,t)\,dx.$$

The mass of fluid at time t passing out through the boundary S per unit of time is given by

$$\int_S \rho(x,t)u(x,t)\cdot n(x)\,ds(x) = \int_V \nabla\cdot(\rho u)(x,t)\,dx,$$

where we used the Divergence theorem. The rate of change of mass in V plus the rate of mass flow through the boundary must be zero if we assume that no fluid is added or removed, which leads to the following expression of *mass conservation*,

$$\frac{\partial}{\partial t}\int_V \rho(x,t)\,dx + \int_V \nabla\cdot(\rho u)(x,t)\,dx = 0.$$

If ρ varies smoothly, then $\frac{\partial}{\partial t}$ may be moved under the integral sign and since V was arbitrarily, we are led to the differential equation expressing *mass conservation*,

$$\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho u) = 0, \tag{232.17}$$

Of course this is a basic equation of mathematical modelling. Performing the differentiation with respect to x, we can express mass conservation in the form

$$\frac{\partial \rho}{\partial t} + u\cdot\nabla\rho + \rho\nabla\cdot u = 0. \tag{232.18}$$

Particle Paths and Streamlines

Let the velocity of a fluid be given by the function $u(x,t)$. Consider the IVP

$$\frac{d}{dt}x(t) = u(x(t),t) \quad \text{for } t > 0,\ x(0) = x_0.$$

The solution $x(t)$ represents the curve, or path or trajectory, followed by a fluid particle that starts at position x_0 at time $t = 0$ and moves with velocity $u(x(t), t)$ for $t > 0$. If the velocity $u(x, t) = u(x)$ is independent of time t, then particle paths are also referred to as *streamlines*.

Incompressible Flow

If the fluid velocity $u(x, t)$ satisfies

$$\nabla \cdot u(x,t) = \big(\frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3}\big)(x,t) = 0 \quad \text{for } x \in \Omega\, t > 0,$$

then the flow is said to be *incompressible* in Ω for $t > 0$.

If the flow is incompressible, the equation (232.18) of mass conservation takes the form

$$\frac{\partial \rho}{\partial t} + u \cdot \nabla \rho = 0. \tag{232.19}$$

Since $\frac{dx}{dt} = u$ for a $x(t)$ particle path, the Chain rule implies

$$\frac{\partial}{\partial t}\rho(x(t),t) = \frac{\partial \rho}{\partial t} + u \cdot \nabla \rho = 0.$$

This says that the density is constant along particle paths, or in other words the volume occupied by a certain set of fluid particles is constant. So, the fluid cannot be compressed. It is common to assume that the density of an incompressible fluid is constant.

Water is very nearly incompressible; it is very difficult to change the total volume of a bucket of water. Air is compressible; the air tank of a diver contains a huge volume of air at normal pressure compressed and stored in a small volume at high pressure. But to get the air into the tank consumes energy.

Incompressible Potential Flow

In so-called *stationary flow*, the velocity $u(x, t)$ is independent of time and thus the fluid velocity $u(x)$ is a function of $x \in \Omega$. Note that in a stationary flow the fluid particles at x are moving if $u(x) \neq 0$, but the velocity of the fluid particles at x does not change with time.

The velocity field $u(x)$ of *rotation-free* fluid flow satisfies $\nabla \times u = 0$, which implies $u = \nabla\varphi$ for a scalar *velocity potential* φ under appropriate convexity assumptions. If the fluid is *incompressible*, then $\nabla \cdot u = 0$, and we obtain the Laplace equation $\Delta\varphi = 0$ for the potential of a rotation-free incompressible flow. At a solid boundary, through which the fluid cannot penetrate, the normal velocity of the fluid is zero, which translates into a homogeneous Neumann boundary condition $\partial_n\varphi = 0$ for the potential φ.

We now give some basic examples of incompressible potential flow. For simplicity, we consider situations in which the velocity $u(x)$ is independent of the x_3-coordinate.

EXAMPLE 232.6. The potential

$$\varphi(x_1, x_2) = x_1^2 - x_2^2$$

satisfies $\Delta\varphi = 0$ and the corresponding flow velocity $u = \nabla\varphi$ is given by

$$u(x) = (2x_1, -2x_2).$$

This represents stationary flow in a corner, see Fig. 232.5. A streamline $x(t)$ satisfies $\frac{dx}{dt} = (2x_1, -2x_2)$, which is a separable equation with solutions satisfying

$$x_1(t)x_2(t) = c,$$

where c is a constant, see Fig. 232.5. We check by computing $\frac{d}{dt}x_1x_2 = \dot{x}_1x_2 + x_1\dot{x}_2 = 2x_1x_2 - 2x_12x_2 = 0$.

EXAMPLE 232.7. The potential

$$\varphi(x) = \log(\|x\|)$$

satisfies $\Delta\varphi(x) = 0$ for $x \neq 0$, and the corresponding flow velocity $u = \nabla\varphi$ is given by $u(x) = \frac{x}{\|x\|^2}$, see Fig. 232.5.

EXAMPLE 232.8. We consider incompressible potential flow around an infinite circular cylinder along the x_3-axis with cross-section $\Omega = \{x = (x_1, x_2) \in \mathbb{R}^2 : \|x\| < 1\}$ from left to right, see Fig. 232.5. The potential φ is given in polar coordinates $x = r(\cos(\theta), \sin(\theta))$ by

$$\varphi(x) = \varphi(r, \theta) = (r + \frac{1}{r})\cos(\theta),$$

corresponding to a flow from right to left sweeping around Ω and approaching $u(x) = (1, 0)$ for $\|x_1\|$ large and x_2 bounded. We note that $\Delta\varphi = 0$ for $r \neq 0$, and that $\frac{\partial\varphi}{\partial r} = 1 - 1/r^2 = 0$ for $r = 1$ and thus the flow is tangential to the boundary of Ω.

Note that fluid flow is rarely rotation-free in the whole region occupied by the fluid. In particular, if the fluid is viscous then rotation is generated at solid boundaries.

Incompressible Flow With Rotation

We now consider basic examples of incompressible flow in two dimensions with non-zero rotation. We assume $u(x) = (u_1(x), u_2(x))$ satisfies $\nabla\cdot u = 0$,

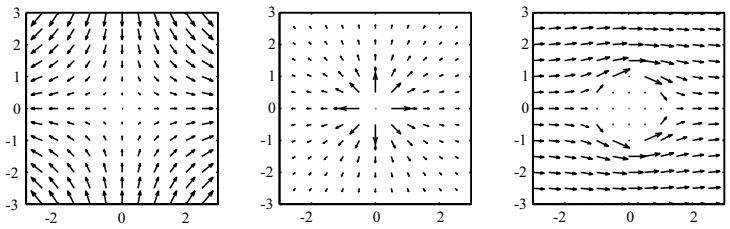


FIGURE 232.5. Examples of incompressible potential flow

where $x = (x_1, x_2)$. Defining $v = (-u_2, u_1)$ this equation reads $\nabla \times v = 0$ and under appropriate convexity assumptions, there is a potential φ with $v = \nabla\varphi$. Thus,

$$u = (v_2, -v_1) = (\frac{\partial\varphi}{\partial x_2}, -\frac{\partial\varphi}{\partial x_1}) = \nabla \times \varphi.$$

With the rotation $\nabla \times u = f$ given, we are led to the Poisson equation for φ,

$$f = \nabla \times u = \nabla \times (\nabla \times \varphi) = -\Delta\varphi.$$

EXAMPLE 232.9. Given $f = 4$ we find the corresponding solution $\varphi(x) = -\|x\|^2$ with $u(x_1, x_2) = (-2x_2, 2x_1)$, see Fig. 232.6. Choosing $\varphi(x) = \log(\|x\|)$ corresponds to $f(x) = 0$ for $x \neq 0$, and the corresponding velocity $u(x_1, x_2) = \|x\|^{-2}(x_2, -x_1)$, see Fig. 232.6.

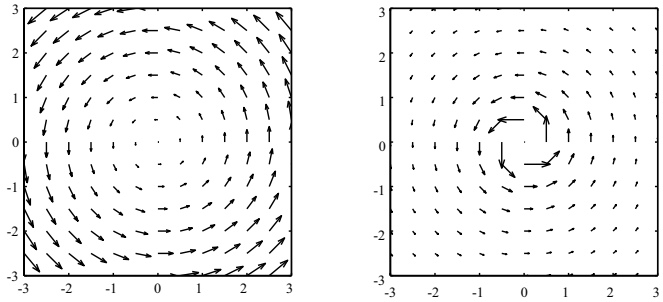


FIGURE 232.6. Incompressible flow with rotation

The Euler and Navier-Stokes Equations

The *Euler equations* for an incompressible *inviscid* fluid with constant density equal to one, take the form

$$\frac{\partial u}{\partial t} + (u \cdot \nabla)u + \nabla p = f, \quad \nabla \cdot u = 0, \tag{232.20}$$

where $u(x,t)$ is the velocity and $p(x,t)$ the *pressure* of the fluid at the point x at time t, and f is an applied volume force like a gravitational force. In an inviscid fluid, the *viscosity* is zero and the only interior force acting between the fluid particles is the pressure force that is equal in all directions and acts normal to any surface. The equation $\nabla \cdot u = 0$ expresses the incompressibility of the flow. The first equation expresses Newton's law stating that the acceleration $\frac{d}{dt}u(x(t),t)$, where $x(t)$ is the trajectory followed by a fluid particle satisfying $\frac{dx}{dt} = u(x(t),t)$, is equal to the force $-\nabla p + f$, consisting of the pressure force $-\nabla p$ and the applied force f. We see this by computing with the the Chain rule and the equation $\frac{dx}{dt} = u(x(t),t)$ to get

$$\frac{d}{dt}u_i(x(t),t) = \frac{\partial u_i}{\partial t} + \frac{dx}{dt} \cdot \nabla u_i = \frac{\partial u_i}{\partial t} + (u \cdot \nabla)u_i,$$

which leads to the vector form (232.20). The *Navier-Stokes equations* are modifications of the Euler equations with an additional viscous force term $-\nu\Delta u$, where ν is the viscosity coefficient. In a fluid with non-zero viscosity, there are also tangential (shear) forces acting on a surface.

232.10 Maxwell's Equations

The interaction between electric and magnetic fields are described by *Maxwell's equations*:

$$\begin{cases} \dfrac{\partial B}{\partial t} + \nabla \times E = 0, \\ -\dfrac{\partial D}{\partial t} + \nabla \times H = J, \\ \nabla \cdot B = 0, \quad \nabla \cdot D = \rho, \\ B = \mu H, \quad D = \epsilon E, \quad J = \sigma E, \end{cases} \tag{232.21}$$

where E is the *electric field*, H is the *magnetic field* , D is the *electric displacement*, B is the *magnetic flux* , J is the *electric current*, ρ is the *charge*, μ is the *magnetic permeability*, ϵ is the *dielectric constant* of *electric permittivity*, and σ is the *electric conductivity*. The first equation is referred to as *Faraday's law*, the second is *Ampère's law*, $\nabla \cdot D = \rho$ is *Coulomb's law*, *Gauss law* $\nabla \cdot B = 0$ expresses the absence of "magnetic

charge", and $J = \sigma E$ is *Ohm's law*. Maxwell, see Fig. 232.7, included the term $\partial D/\partial t$ for purely mathematical reasons and then used Calculus to predict the existence of electromagnetic waves before these had been observed experimentally.

FIGURE 232.7. Maxwell (1831-1879), inventor of the mathematical theory of electromagnetism: "We can scarcely avoid the conclusion that light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena".

Typical boundary conditions include various combinations of $E \cdot n$ (perfect insulator), $E \times n$ (perfect conductor), $H \cdot n$ and $H \times n$.

Maxwell's equations describe the whole world of electromagnetic phenomena with an astounding economy of notation and accuracy of modelling. Our modern information society builds on electromagnetic waves. We shall now pick out a couple of Laplace equation models from Maxwell's equations by considering some basic particular cases.

Electrostatics

A basic problem in *electrostatics* is to describe the stationary electric field $E(x)$ in a volume Ω in $\mathbb{R}^3$ containing *charges* of density $\rho(x)$ and enclosed by a perfectly conducting surface Γ. Faraday's law states that

$$\nabla \times E = 0 \quad \text{in } \Omega,$$

since we assume that $\frac{\partial B}{\partial t} = 0$. Recalling Chapter Potential fields, it follows that the electric field E is the gradient of a scalar *electric potential* φ, i.e. $E = \nabla\varphi$. Coulomb's law says

$$\nabla \cdot E = \rho \quad \text{in } \Omega,$$

so we are led to the Poisson equation for the potential φ,

$$\Delta\varphi = \nabla \cdot \nabla\varphi = \rho \quad \text{in } \Omega.$$

The boundary condition $E \times n = 0$ on the boundary Γ of Ω with n denoting the outward unit normal, says that the tangential component of E vanishes on the boundary. This models a perfectly conducting boundary in which differences in the electric field are leveled out. This means that $E = \nabla\varphi$ is normal to the boundary, so the boundary is a level surface of φ and the potential φ is constant on the boundary. Since φ is undetermined up to a constant, we may assume that $\varphi = 0$ on the boundary and we arrive at Poisson's equation $-\Delta\varphi = f$ with $f = -\rho$ in Ω with homogenous Dirichlet boundary conditions $\varphi = 0$ on Γ.

The potential $\varphi(x)$ of a point charge at the origin is given by

$$\varphi(x) = \frac{c}{\|x\|}$$

with the corresponding electric field

$$E(x) = -\frac{cx}{\|x\|^3}$$

and c a suitable constant. We shall return to this solution below.

EXAMPLE 232.10. Let $\Omega = \{x \in \mathbb{R}^2 : \|x\| < 1, x_1 < 0 \text{ or } x_2 > 0\}$ be a circular disc with a piece cut out and a *reentrant* corner of angle $\omega = \frac{3\pi}{2}$, see Fig. 232.8. By a direct computation we can verify that the function

$$\varphi(x) = r^\alpha \sin(\alpha\theta)$$

expressed in polar coordinates $x = r(\cos(\theta), \sin(\theta))$, where $\alpha = \frac{\pi}{\omega} = \frac{2}{3}$, satisfies the Laplace equation $\Delta\varphi = 0$ in Ω and the boundary condition $\varphi = 0$ in the straight parts of the boundary meeting at the origin. Letting φ represent an electric potential, the corresponding electric field $E(x) = \nabla\varphi(x)$ satisfies

$$\frac{\partial E}{\partial r} = \alpha r^{\alpha-1} \sin(\alpha\theta)$$

and thus since $\alpha < 1$, is singular (infinite) at the corner where $r = 0$. This means that the electric field is very strong close to the corner, and the sharper the corner (α smaller) the stronger is the field. This may support the observation that an electric lightening is more likely to hit the pointed tower of church than a smooth hill, or the design of an electronic scanner where electrons pop out of the pin of a sharp needle.

EXAMPLE 232.11. The potential φ of the electric field between two concentric spheres $S_1 = \{x \in \mathbb{R}^3 : \|x\| < r_1\}$ and $S_2 = \{x \in \mathbb{R}^3 : \|x\| < r_2\}$ with $r_2 > r_1$, is given by

$$\varphi(x) = \frac{1}{\|x\|}$$

if we assume that $\varphi = 1/r_i$ on S_i, $i = 1, 2$.

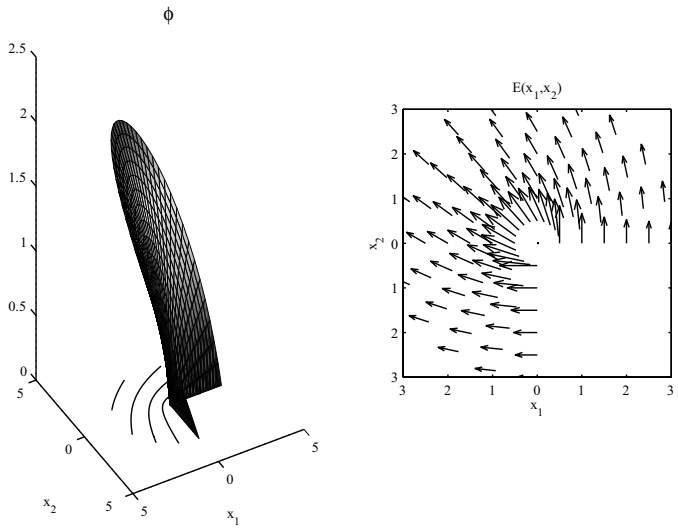


FIGURE 232.8. A potential with singular electric field

EXAMPLE 232.12. The function

$$\varphi(x_1, x_2)) = \arctan(\frac{x_2}{x_1})$$

defined for $x_1 > 0$ satisfies $\Delta\varphi(x) = 0$ for $x_1 > 0$, and is constant $= \arctan(c)$ on rays $x_2 = cx_1$ through the origin of slope c. The corresponding electric field $E(x) = \nabla\varphi(x)$ given by

$$E(x) = \frac{(-x_2, x_1)}{\|x\|^2}.$$

We see that $E(x)$ is singular at $x = 0$.

Magnetostatics

The basic problem in *magnetostatics* arises by combining Gauss' law $\nabla \cdot H = 0$, assuming μ constant and guaranteeing that $H = \nabla \times \varphi$ for some vector potential ψ satisfying $\nabla \cdot \psi = 0$, with Faraday's law $\nabla \times H = J$ to give

$$\nabla \times (\nabla \times \psi) = -\Delta\psi = J,$$

where we use the facts that $\nabla \times (\nabla \times \psi) = -\Delta\psi + \nabla(\nabla \cdot \psi)$ and that $\nabla \cdot \varphi = 0$.

The magnetic field around a unit current J in the x_3-direction is given by

$$H(x) = \frac{1}{2\pi}\frac{(-x_2, x_1, 0)}{\|x\|^2},$$

which can be verified by direct computation showing that $\nabla \cdot H(x) = 0$ and $\nabla \times H(x) = 0$ for $(x_1, x_2) \neq 0$. The presence of the factor $\frac{1}{2\pi}$ makes $\int_\Gamma H \cdot ds = 1$ for any counter-clockwise oriented circle in the x_1x_2-plane, from which by Stokes theorem follows that $\nabla \times H = J$, see the next section on Gravitation and delta functions.

Time-Dependent Magnetics

In low frequency applications, the term $\frac{\partial D}{\partial t}$ so cleverly introduced by Maxwell, plays a minor role and can be discarded. Let's see where this leads. Since $\nabla \cdot B = 0$, we can write B as $B = \nabla \times \psi$, where ψ is a magnetic vector potential. Inserting this into Faraday's law gives

$$\nabla \times \left(\frac{\partial \psi}{\partial t} + E \right) = 0,$$

from which it follows that

$$\frac{\partial \psi}{\partial t} + E = \nabla \varphi,$$

for some scalar potential φ. Multiplying by σ and using the laws of Ohm and Ampère, we obtain a vector equation for the magnetic potential ψ:

$$\sigma \frac{\partial \psi}{\partial t} + \nabla \times \left(\mu^{-1} \nabla \times \psi \right) = \sigma \nabla \varphi.$$

This system reduces to a scalar equation in two variables if we assume that $B = (B_1, B_2, 0)$ is independent of x_3. It then follows that ψ has the form $\psi = (0, 0, u)$ for some scalar function u that depends only on x_1 and x_2, so that $B_1 = \partial u / \partial x_2$ and $B_2 = -\partial u / \partial x_1$. We end up with a scalar equation for the scalar magnetic potential u in the form

$$\sigma \frac{\partial u}{\partial t} - \nabla \cdot \left(\mu^{-1} \nabla u \right) = f, \tag{232.22}$$

for some function $f(x_1, x_2)$. Choosing $\sigma = \mu = 1$ leads to the heat equation,

$$\begin{cases} \frac{\partial}{\partial t} u(x,t) - \Delta u(x,t) = f(x,t) & \text{for } x \in \Omega,\ 0 < t \leq T, \\ u(x,t) = 0 & \text{for } x \in \Gamma,\ 0 < t \leq T, \\ u(x,0) = u_0(x) & \text{for } x \in \Omega, \end{cases} \tag{232.23}$$

where $\Omega \subset \mathbb{R}^2$ with boundary Γ, and we posed homogeneous Dirichlet boundary conditions. In the stationary case, we again obtain Poisson's equation with Dirichlet boundary conditions.

232.11 Gravitation

In his famous treatise *Mécanique Céleste* in five volumes published during 1799-1825, Laplace extended Newton's theory of gravitation and in particular developed a theory for describing gravitational fields based on using gravitational potentials that satisfy Laplace's equation, or more generally Poisson's equation.

We consider a gravitational field in $\mathbb{R}^3$ with gravitational force $F(x)$ at position x, generated by a distribution of mass of density $\rho(x)$. We recall that the work of a unit mass, moving along a curve Γ is given by

$$\int_\Gamma F \cdot ds,$$

If the curve Γ is closed, then the total work performed by the gravitational forces should be zero. Stokes' theorem implies that a gravitational field F should satisfy $\nabla \times F = 0$. Using the basic result of the Chapter Potential fields, we conclude that F is the gradient of a scalar potential φ, i.e.

$$F(x) = \nabla\varphi(x). \tag{232.24}$$

Laplace proposed the following relation between the gravitational field F and the mass distribution ρ,

$$-\nabla \cdot F(x) = \rho(x), \tag{232.25}$$

assuming the gravitational constant is normalized to one. This is analogous to Coulomb's law $\nabla \cdot E = \rho$ in electrostatics and also to the energy balance equation $\nabla \cdot q = f$ for stationary heat conduction, where q is the heat flux and f a heat source. In particular, (239.4) states that $\nabla \cdot F(x) = 0$ at points x where there is no mass so that $\rho(x) = 0$. Combining (239.3) and (239.4), we obtain Poisson's equation $-\Delta\varphi = \rho$ for the gravitational potential φ.

Since the origin and property of gravitation of "acting at a distance" is still lacking a convincing physical explication, the equation $-\nabla \cdot F(x) = \rho(x)$ including $\nabla \cdot F = 0$ in empty space, should be viewed as a basic postulate on the nature of a gravitational field. Of course it seems very difficult to motivate that $\nabla \cdot F$ should be something different from zero in empty space, but a real a "proof" that $\nabla \cdot F$ must be zero in empty space seems to be missing.

Newton considered gravitational fields generated by *point masses*. Mathematically, a unit point mass at a point $z \in \mathbb{R}^3$ is represented by the so-called *delta function* δ_z at z, defined by the property that for any smooth function v,

$$\int_{\mathbb{R}^3} \delta_z \, v \, dx = v(z), \tag{232.26}$$

where the integration is to be interpreted in a generalized sense. We could think of a δ_z as a limit of positive functions $\varphi_h(x)$ such that $\varphi_h(x) = 0$ if

$\|x - z\| > h$ and

$$\int_{\mathbb{R}^3} \varphi_h(x)\,dx = 1,$$

as h tends to zero. For example we may choose

$$\varphi_h(x) = \frac{3}{4\pi h^3} \quad \text{if } \|x - z\| < h$$

and $\varphi_h(x) = 0$ elsewhere. If $v(x)$ is Lipschitz continuous at z, then

$$lim_{h\to 0} \int_{\mathbb{R}^3} \varphi_h(x)v(x)\,dx = v(z),$$

which gives (232.26) its meaning. The function $\varphi_h(x)$ represents a very tall and narrow "hump" around z with volume one.

We expect that the gravitational potential $\Phi(x)$ corresponding to a unit point mass at the origin, to satisfy

$$-\Delta\Phi = \delta_0 \quad \text{in } \mathbb{R}^3, \tag{232.27}$$

assuming the gravitational constant to be equal to one. To give a precise meaning to this equation involving the somewhat mysterious delta function δ_0 at 0, we first formally multiply by a smooth *test function* v vanishing outside a bounded set to get

$$-\int_{\mathbb{R}^3} \Delta\Phi(x)v(x)\,dx = v(0). \tag{232.28}$$

Next, we integrate the left-hand side by parts formally using Green's formula to move the Laplacian from E to v, noting that the boundary terms disappear since v vanishes outside a bounded set. We may thus reformulate (239.5) as seeking a potential $E(x)$ satisfying

$$-\int_{\mathbb{R}^3} \Phi(x)\Delta v(x)\,dx = v(0), \tag{232.29}$$

for all smooth functions $v(x)$ vanishing outside a bounded set. We may view this as the concrete interpretation of (239.5), which is perfectly well defined since now the Laplacian acts on the smooth function $v(x)$ and the potential Φ is assumed to be integrable. We also require the potential $\Phi(x)$ to decay to zero as $\|x\|$ tends to infinity, which corresponds to a "zero Dirichlet boundary condition at infinity".

In the Chapter The divergence, rotation and Laplacian, we showed that the function $1/\|x\|$ satisfies Laplace's equation $\Delta u(x) = 0$ for $0 \neq x \in \mathbb{R}^3$, while it is singular at $x = 0$. We shall prove that the following scaled version of this function satisfies (239.7):

$$\Phi(x) = \frac{1}{4\pi}\frac{1}{\|x\|}. \tag{232.30}$$

We refer to this function as the *fundamental solution* of $-\Delta$ in $\mathbb{R}^3$. We conclude in particular that the gravitational field in $\mathbb{R}^3$ created by a unit point mass at the origin is given by

$$F(x) = \nabla\Phi(x) = -\frac{1}{4\pi}\frac{x}{\|x\|^3},$$

which is precisely Newton's inverse square law of gravitation. Laplace thus gives a motivation why the exponent should be two, which Newton did not (and therefore was criticized by Leibniz). Of course, it still remains to motivate (239.4). In the context of heat conduction, the fundamental solution $E(x)$ represents the stationary temperature in a homogeneous body with heat conductivity equal to one filling the whole of $\mathbb{R}^3$, subject to a concentrated heat source of strength one at the origin and with the temperature tending to zero as $\|x\|$ tends to infinity.

We now prove that the function $\Phi(x)$ defined by (239.8) satisfies (239.7). We first note that since Δv is smooth and vanishes outside a bounded set and $\Phi(x)$ is integrable over bounded domains,

$$\int_{\mathbb{R}^3} \Phi\Delta v\,dx = \lim_{a\to 0^+}\int_{D_a} \Phi\Delta v\,dx, \tag{232.31}$$

where $D_a = \{x \in \mathbb{R}^3 : a < \|x\| < a^{-1}\}$, with a small positive, is a bounded region obtained from $\mathbb{R}^3$ by removing a little sphere of radius a with boundary surface S_a and also points further away from the origin than a^{-1}, see Fig. 239.4. We now use Green's formula on D_a with $w = \Phi$. Since

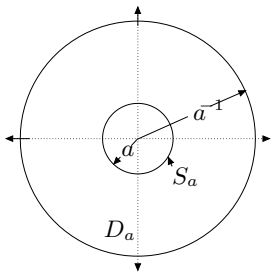


FIGURE 232.9. A cross-section of the domain D_a.

v is zero for $\|x\|$ large, the integrals over the outside boundary vanish when a is sufficiently small. Using the fact that $\Delta\Phi = 0$ in D_a, $\Phi = 1/(4\pi a)$ on S_a and $\partial\Phi/\partial n = 1/(4\pi a^2)$ on S_a with the normal pointing in the direction of the origin, we obtain

$$-\int_{D_a} \Phi\Delta v\,dx = \int_{S_a}\frac{1}{4\pi a^2}\,v\,ds - \int_{S_a}\frac{1}{4\pi a}\frac{\partial v}{\partial n}\,ds = I_1(a) + I_2(a),$$

with the obvious definitions of $I_1(a)$ and $I_2(a)$. Now, $\lim_{a\to 0} I_1(a) = v(0)$ because $v(x)$ is continuous at $x = 0$ and the surface area of S_a is equal to $4\pi a^2$, while $\lim_{a\to 0} I_2(a) = 0$. The desired equality (239.7) now follows recalling (239.9).

The corresponding fundamental solution of $-\Delta$ in $\mathbb{R}^2$ is given by

$$\Phi(x) = \frac{1}{2\pi}\log(\frac{1}{\|x\|}). \tag{232.32}$$

In this case the fundamental solution is not zero at infinity.

Replacing 0 by an arbitrary point $z \in \mathbb{R}^3$, (239.7) becomes

$$-\int_{\mathbb{R}^3} \Phi(z-x)\Delta v(x)\,dx = v(z), \tag{232.33}$$

which leads to a solution formula for Poisson's equation in $\mathbb{R}^3$. For example, if u satisfies the Poisson equation $-\Delta u = f$ in $\mathbb{R}^3$ and $|u(x)| = O(\|x\|^{-1})$ as $\|x\| \to \infty$, then u may be represented in terms of the fundamental solution Φ and the right-hand side f as follows:

$$u(z) = \int_{\mathbb{R}^3} \Phi(z-x)f(x)\,dx = \frac{1}{4\pi}\int_{\mathbb{R}^3} \frac{f(x)}{\|z-x\|}\,dx. \tag{232.34}$$

We see that $u(z)$ is a mean value of f centered around z weighted so that the influence of the values of $f(x)$ is inversely proportional to the distance from z.

Similarly, the potential u resulting from a distribution of mass of density $\rho(x)$ on a (bounded) surface Γ in $\mathbb{R}^3$ is given by

$$u(z) = \frac{1}{4\pi}\int_\Gamma \frac{\rho(x)}{\|z-x\|}\,ds(x). \tag{232.35}$$

Formally, we obtain this formula by simply adding the potentials from all the different pieces of mass on Γ. One can show that the potential u defined by (239.13) is continuous in $\mathbb{R}^3$ if ρ is bounded on Γ, and of course u satisfies Laplace's equation away from Γ. Suppose now that we would like to determine the distribution of mass ρ on Γ so that the corresponding potential u defined by (239.13) is equal to a given potential u_0 on Γ, that is we seek in particular a function u solving the boundary value problem $\Delta u = 0$ in Ω and $u = u_0$ on Γ, where Ω is the volume enclosed by Γ. This leads to the following *integral equation*: given u_0 on Γ find the function ρ on Γ such that

$$\frac{1}{4\pi}\int_\Gamma \frac{\rho(y)}{\|x-y\|}\,ds(y) = u_0(x) \quad \text{for } x \in \Gamma. \tag{232.36}$$

This is a *Fredholm integral equation of the first kind*, named after the Swedish mathematician Ivar Fredholm (1866-1927). In the beginning of

the 20th century, Fredholm and Hilbert were competing to prove the existence of solutions of the basic boundary value problems of mechanics and physics using integral equation methods. The integral equation (239.14) is an alternative way of formulating the boundary value problem of finding u such that $\Delta u = 0$ in Ω, and $u = u_0$ on Γ. Integral equations may also be solved using Galerkin methods.

232.12 The Eigenvalue Problem for the Laplacian

The *eigenvalue problem* for the Laplace operator with Dirichlet boundary conditions on a domain Ω in $\mathbb{R}^d$ with boundary Γ takes the form: Find nonzero *eigen-functions* $\varphi(x)$ with corresponding *eigenvalues* λ such that

$$\begin{cases} -\Delta\varphi = \lambda\varphi & \text{in } \Omega, \\ \varphi = 0 & \text{on } \Gamma. \end{cases} \tag{232.37}$$

In the one-dimensional case with $\Omega = (0, \pi)$, the eigenfunctions are (modulo normalization) $\varphi_n(x) = \sin(nx)$ with corresponding eigenvalues $\lambda_n = n^2$, $n = 1, 2, ...$ For a two-dimensional square $\Omega = (0, \pi) \times (0, \pi)$, the eigenfunctions are $\varphi_{nm}(x_1, x_2) = \sin(nx_1)\sin(mx_2)$, $n, m = 1, 2, ...$, with eigenvalues $\lambda_{nm} = n^2 + m^2$.

It follows by multiplication of (250.2) by φ and integration by parts, that all eigenvalues λ are positive. More precisely, there is an increasing sequence of eigenvalues tending to infinity, and eigenfunctions corresponding to different eigenvalues are orthogonal with respect to the scalar product $(v, w) = \int_\Omega vw\, dx$.

If $\varphi(x)$ is an eigenfunction with corresponding eigenvalue λ, then the (real part of the) function $u(x, t) = \exp(it\sqrt{\lambda})\varphi(x)$ solves the homogeneous wave equation

$$\ddot{u} - \Delta u = 0 \text{ in } \Omega \times \mathbb{R}$$

corresponding to a vibrating elastic membrane (drum head) if $d = 2$ (string if $d = 1$). The smallest eigenvalue corresponds to the basic tone of the drum head.

In Fig. 250.1, we show contour plots for the first four eigenfunctions, corresponding to $\lambda_1 \approx 38.6$, $\lambda_2 \approx 83.2$, $\lambda_3 \approx 111.$, and $\lambda_4 \approx 122.$, in a case where Ω corresponds to the lid of a guitar with Dirichlet boundary conditions on the outer boundary, described as an ellipse, and Neumann boundary conditions at the hole in the lid,

Often the smaller eigenvalues are the most important in considerations of design. This is the case for example in designing suspension bridges, which must be built so that the lower eigenvalues of vibrations in the bridge are not close to possible wind-induced frequencies. This was not well understood in the early days of suspension bridges which caused the famous collapse of the Tacoma bridge in 1940.

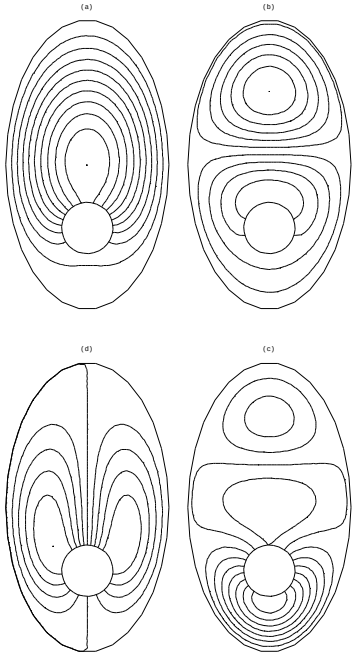


FIGURE 232.10. Contour plots of the first four eigenfunctions of the guitar lid corresponding to (a) $\lambda_1 \approx 38.6$, (b) $\lambda_2 \approx 83.2$, (c) $\lambda_3 \approx 111.$, and (d) $\lambda_4 \approx 122.$. These were computed with Femlab with a fixed mesh size of diameter .02.

The smallest eigenvalue is equal to the minimum value of the *Rayleigh quotient*

$$\frac{(\nabla\psi, \nabla\psi)}{(\psi, \psi)},$$

when varying over functions ψ satisfying the boundary conditions. More, generally, the eigenvalues corresponds to stationary values of the Rayleigh quotient.

232.13 Quantum Mechanics

The two most revolutionary achievements of physics during the 20th century was the development of of the *Theory of General Relativity* for Gravitation on astronomic scales by Einstein, and *Quantum Mechanics* for atomic scales by Schrödinger (1887-1961, Nobel Prize in Physics 1933), see Fig. 250.14. Einstein never fully accepted Quantum Mechanics, and the *Grand Unified Theory* connecting Gravitation and Quantum Mechanics is still missing, with *String Theory* being a recent attempt to fill the gap.

The basic equation of Quantum Mechanics is the *Schrödinger equation*, which for a system of N electrons (with the Born-Oppenheimer approximation) takes the following normalized form:

$$i\frac{\partial\varphi}{\partial t} = H\varphi = \big(-\frac{1}{2}\sum_j \Delta_j + V(r_1,...,r_N)\big)\varphi, \tag{232.38}$$

where $\varphi = \varphi(r_1,...,r_N,t)$ is a *wave function* depending on the set of space coordinates $(r_1,...,r_N)$ with each r_j varying over $\mathbb{R}^3$, together with time t, Δ_j denotes the Laplacian with respect to the coordinate $r_j \in \mathbb{R}^3$, and $V(r_1,...,r_N)$ denotes a *potential* depending on $(r_1,...,r_N)$ representing repulsive Coulomb forces between the electrons and attractive Coulomb forces between the electrons and the (fixed) nuclei of the system, $H = -\frac{1}{2}\sum_j \Delta_j + V$ is the *Hamiltonian* representing a sum of kinetic and potential energies, and i denotes the imaginary unit. The wave function is complex-valued and the square of its modulus represents an electron probability density.

The Schrödinger equation appears to give a very good description of phenomena on atomistic scales, but unfortunately it is not easy to deal with because of the large number of spatial dimensions involved: For a system with 100 electrons, which is still very small, the number of space dimensions is equal to 300, and standard techniques for either analytical or numerical solution fall very short. So, although the Schrödinger equation admittedly is a very beautiful equation which gives a surprisingly concise description of atomistic physics, it is certainly impossible to solve exactly analytically, and approximate solution becomes a key issue. The 1998 Nobel Prize in Chemistry was awarded Robert Kohn for his method for approximate solution of the Schrödinger equation based on using a single *electron density function* with the space dependence restricted to $\mathbb{R}^3$, independent of the number of electrons, and corresponding approximate potentials. Such simplified Schrödinger equations, referred to as *Kohn-Sham equations*, are today used extensively in computational chemistry.

The Hydrogen Atom

The Hydrogen atom consisting of one electron and one neutron is the only case in which analytical solution of the Schrödinger equation is feasible: In this case the Schrödinger equation takes the following (normalized) form assuming the neutron is positioned at the origin: Find the wave function $\varphi(x,t)$ with $x \in \mathbb{R}^3$, such that for $t > 0$

$$i\frac{\partial\varphi}{\partial t} = (-\frac{1}{2}\Delta + V)\varphi \quad \text{in } \mathbb{R}^3, \tag{232.39}$$

FIGURE 232.11. Schrödinger (1887-1961) at age 13: "I was a good student in all subjects, loved mathematics and physics, but also the strict logic of the ancient grammars, hated only memorizing incidental dates and facts. Of the German poets, I loved especially the dramatists, but hated the pedantic dissection of this works"

where Δ is the usual Laplacian with respect to x, $V(x) = -\frac{1}{|x|}$ is the Coulomb potential of the proton, with the normalization that

$$\int_{\mathbb{R}^3} |\varphi(x,t)|^2\,dx = 1 \quad \text{for } t > 0.$$

For a domain $\Omega \in \mathbb{R}^3$, the integral

$$\int_{\Omega} |\varphi(x,t)|^2\,dx$$

represents the probability to find the electron in the domain Ω at time t. Formally, $-\frac{1}{2}\Delta$ corresponds to the kinetic energy $\frac{p^2}{2m}$ with p the momentum and m the mass, replacing p by $-i\nabla$ and setting $m = 1$.

In the time-harmonic case with a time-dependence of the form $exp(-i\omega t)$ with frequency ω, this leads to the eigenvalue problem: Find $\varphi(x) \neq 0$ and

$\omega \in \mathbb{R}$ such that

$$H\varphi = \omega\varphi, \tag{232.40}$$

with $H = -\frac{1}{2}\Delta + V$ the Hamiltonian and the eigenvalue ω representing an energy level. The eigenvalues are real and the (real) eigenfunction corresponding to the smallest eigenvalue (smallest energy) is referred to as the *ground state* and the eigenfunctions corresponding to larger eigenvalues as *bound states*.

Assuming spherical symmetry (232.40) takes the following form in spherical coordinates with r the radius: Find $\varphi(r)$ such that

$$-\frac{1}{2}\frac{d^2\varphi}{dr^2} - \frac{1}{r}\frac{d\varphi}{dr} - \frac{1}{r}\varphi = \omega\varphi \quad \text{for } r > 0,$$

with the side condition that $\varphi(0)$ is finite and $\varphi(x)$ is square integrable over $\mathbb{R}^3$. The ground state is given by the eigenfunction $\varphi(r) = \exp(-r)$ corresponding to the eigenvalue $\omega = -\frac{1}{2}$.

Chapter 232 Problems

232.1. Interpret the fixed point iteration for Poisson's equation as an explicit time stepping scheme for the heat equation $\frac{du}{dt} - \Delta u = f$ with time step αh^2 with the starting value given by the initial approximation U^0. Explain why the convergence is slow if h is small.

232.2. Consider a horizontal elastic membrane spanned over a circular ring with constant tension H in all directions in unloaded configuration. Discuss under what conditions the membrane can carry a non-zero volume of water and try to compute the volume.

232.3. Prove that (239.10) is a fundamental solution of $-\Delta$ in $\mathbb{R}^2$.

232.4. Because the presented mathematical models of heat flow and gravitation, namely Poisson's equation, are the same, it opens the possibility of thinking of a gravitational potential as "temperature" and a gravitational field as "heat flux". Can you "understand" something about gravitation using this analogy?

232.5. Present the integral equation corresponding to (239.14) in the case $d = 2$.

232.6. What equation is obtained if $\partial D/\partial t$ is not neglected in the setting of time-dependent magnetics, but the x_3 independence is kept?

232.7. Derive the heat equation describing the heat conduction in a thin piece of wire of length one whose ends are kept at a fixed temperature (i.e., derive the heat equation in one dimension):

$$\begin{cases} \dot{u} - u'' = f & \text{in } (0,1)\times(0,T], \\ u(0,t) = u(1,t) = 0 & \text{for } t \in (0,T], \\ u(x,0) = u_0(x) & \text{for } x \in (0,1). \end{cases} \tag{232.41}$$

232.8. Let $F(x)$ be the gravitational field generated by a homogeneous ball of mass m occupying the volume $\{x \in \mathbb{R}^3 : \|x\| \leq r\}$, satisfying $\nabla F(x) = \rho$ for $\|x\| < r$ and $\nabla F(x) = 0$ for $\|x\| > r$, where ρ is the density of the sphere. Argue that by symmetry $F(x) = f(\|x\|)\frac{-x}{\|x\|}$ for $\|x\| > r$ for some function $f : (0,\infty) \to \mathbb{R}$. Use the Divergence theorem to see that if $R > r$ then

$$\int_{S_R} F(x) \cdot n dS = 4\pi R^2 f(R) = \int_{B_R} \nabla F(x)\, dx = m,$$

where S_R is the boundary of the ball $B_R = \{x \in \mathbb{R}^3 : \|x\| \leq R\}$. Conclude that $f(R) = \frac{m}{4\pi R^2}$, and thus that $F(x) = \frac{m}{4\pi}\frac{-x}{\|x\|^3}$ for $\|x\| > r$. This gives an alternative way of handling of Newton's nightmare. Note the change of normalization with the factor $1/4\pi$ appearing here.

232.9. To analyze the convergence of the fixed point iteration for the system of equations (232.15), we need to show that $\|I - \alpha A\| < 1$, where $A = (a_{ij})$ is the $(N-1) \times (N-1)$ matrix with $a_{ii} = 2$, $a_{i,i-1} = a_{i-1,i} = -1$ and $a_{ij} = 0$ if $|i-j| > 1$. Since A is symmetric, we have recalling the Chapter The Spectral Theorem:

$$\|I - \alpha A\| = \max_i |1 - \alpha\lambda_i|,$$

where λ_i, $i = 1, ..., N-1$, are the eigenvalues of A. To see this, diagonalize. Prove that for all nonzero $V \in \mathbb{R}^{N-1}$

$$AV \cdot V = \sum_{i,j=1}^{N-1} a_{ij} V_i V_j > 0,$$

and conclude that $\lambda_i > 0$ for all i (Hint: complete squares!). Show similarly that for all $V \in \mathbb{R}^{N-1}$

$$(I - \alpha A)V \cdot V \geq 0$$

if $\alpha \leq \frac{1}{4}$ (Hint: same as before!). Conclude that Fixed point iteration converges if $0 < \alpha \leq \frac{1}{4}$. Can you prove convergence if $\alpha < \frac{1}{2}$? What about convergence if $\alpha < 0$? Hint: Use that if A is a symmetric $m \times m$ matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_m$, then $\lambda_1 = \min_{V \in \mathbb{R}^m}(AV \cdot V)/(V \cdot V)$ and $\lambda_m = \max_{V \in \mathbb{R}^m}(AV \cdot V)/(V \cdot V)$, where $V \neq 0$.

232.10. Extend the above analysis to the 5-point scheme for the Laplacian and show that fixed point iteration converges if $0 < \alpha < \frac{1}{8}$ (or better $\alpha < \frac{1}{4}$).

232.11. Gather some friends and arrange them in a square regular grid, and ask them to keep updating their own value according to a Svensson's formula as the mean value of their neighbors (starting with zero), and assigning certain given values to the people at the boundary. Collect the values obtained after convergence. You have solved Laplace equation on a square with Dirichlet boundary values numerically. What value of α in fixed point iteration did you effectively use?

232.12. Prove Bernoulli's theorem stating that in stationary Euler flow satisfying $(u \cdot \nabla)u + \nabla p = 0$ the quantity $\frac{1}{2}\|u\|^2 + p$ is constant along streamlines.

232.13. Explain the *Magnus effect* causing a top-spin tennis ball curve downwards (see also Chapter *Analytic functions*).

232.14. Prove that the hydrogen atom is stable in the sense that the *Rayleigh quotient*

$$RQ(\psi) = \frac{\frac{1}{2}\int_\Omega |\nabla\psi|^2\,dx - \int_\Omega \psi^2/r\,dx}{\int_\Omega \psi^2\,dx},$$

satisfies

$$\min_{\psi\in V} RQ(\psi) \geq -2,$$

showing that the electron does not fall into the proton. Hint: estimate $\int_\Omega \psi\frac{\psi}{r}$ using Cauchy's inequality and the following Poincaré inequality for functions $\psi \in V$:

$$\int_\Omega \frac{\psi^2}{r^2}\,dx \leq 4\int_\Omega |\nabla\psi|^2\,dx. \tag{232.42}$$

This shows that the potential energy cannot outpower the kinetic energy in the Rayleigh quotient. To prove the last inequality, use the representation

$$\int_\Omega \frac{\psi^2}{r^2}\,dx = -\int_\Omega 2\psi\nabla\psi\cdot\nabla\ln(|x|)\,dx.$$

resulting from Green's formula, together with Cauchy's inequality.

232.15. (a) Show that the eigenvalue problem for the hydrogen atom for eigenfunctions with radial dependence only, may be formulated as the following one-dimensional problem

$$-\frac{1}{2}\varphi_{rr} - \frac{1}{r}\varphi_r - \frac{1}{r}\varphi = \lambda\varphi, \quad r > 0, \quad \varphi(0) \text{ finite}, \quad \int_{\mathbb{R}} \varphi^2 r^2\,dr < \infty, \tag{232.43}$$

where $\varphi_r = \dfrac{d\varphi}{dr}$. (b) Show that $\psi(r) = \exp(-r)$ is an eigenfunction corresponding to the eigenvalue $\lambda = -\frac{1}{2}$. (b) Is this the smallest eigenvalue? (c) Determine λ_2 and the corresponding eigenfunction by using a change of variables of the form $\varphi(r) = v(r)\exp(-\frac{r}{2})$. (d) Solve (250.17) numerically.

> The idea of the continuum seems simple to us. We have somehow lost sight of the difficulties it implies ... We are told such a number as the square root of 2 worried Pythagoras and his school almost to exhaustion. Being used to such queer numbers from early childhood, we must be careful not to form a low idea of the mathematical intuition of these ancient sages; their worry was highly credible. (Schrödinger)

233

Piecewise Linear Polynomials in $\mathbb{R}^2$ and $\mathbb{R}^3$

...usually he sat in a comfortable attitude, looking down, slightly stooped, with hands folded above his lap. He spoke quite freely, very clearly, simply and plainly: but when he wanted to emphasize a new viewpoint ... then he lifted his head, turned to one of those sitting next to him, and gazed at him with his beautiful, penetrating blue eyes during the emphatic speech. ... If he proceeded from an explanation of principles to the development of mathematical formulas, then he got up, and in a stately very upright posture he wrote on a blackboard beside him in his peculiarly beautiful handwriting: he always succeeded through economy and deliberate arrangement in making do with a rather small space. For numerical examples, on whose careful completion he placed special value, he brought along the requisite data on little slips of paper. (Dedekind about Gauss)

233.1 Introduction

In this chapter, we prepare for the application of FEM to partial differential equations by discussing approximation of functions by piecewise linear functions in in $\mathbb{R}^2$ and $\mathbb{R}^3$. We consider three main topics: (i) the construction of a mesh, or *triangulation*, for a domain in $\mathbb{R}^2$ or $\mathbb{R}^3$, (ii) the construction piecewise linear functions on a triangulation, and (iii) estimation of interpolation errors.

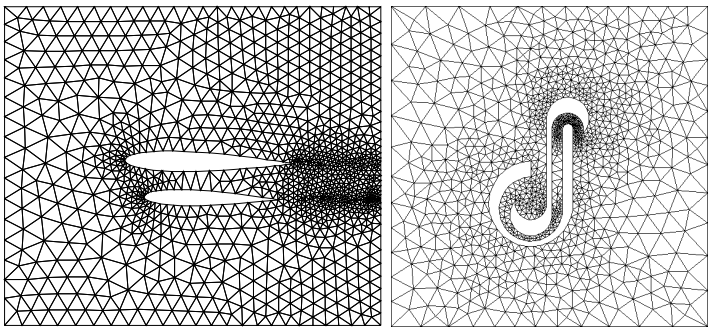


FIGURE 233.1. The mesh on the left was used in a computation of the flow of air around two airfoils. The mesh on the right was used to discretize a piece of metal punched with a fancy character. In both cases, the meshes are adapted to allow accurate computation, taking into account both the behavior of the solution and the shape of the domain.

233.2 Triangulation of a Domain in $\mathbb{R}^2$

We start by considering a two-dimensional domain Ω with a polygonal boundary Γ. A *triangulation* $\mathcal{T}_h = \{K\}$ is a sub-division of Ω into a non-overlapping set of triangles, or *elements*, K constructed so that no vertex of one triangle lies on the edge of another triangle, see Fig. 233.2. We use $\mathcal{N}_h = \{N\}$ to denote the set of *nodes* N or corners of the triangles, usually numbered $N_1, N_2, ..., N_M$, where M is the total number of nodes. A triangulation is specified by a list of the coordinates of the nodes, together with a list containing the numbers of the nodes of each triangle. We may also list the set of triangle sides or *edges* $\mathcal{S}_h = \{S\}$, with each edge S specified by the node numbers of its two end-points, and a list of the nodes and edges on the boundary Γ.

We measure the size of a triangle $K \in \mathcal{T}_h$, by the length h_K of its largest side, which is called the *diameter* of the triangle. The *mesh function* $h(x)$ associated to a triangulation $\mathcal{T}_h$ is the piecewise constant function defined so $h(x) = h_K$ for $x \in K$ for each $K \in \mathcal{T}_h$. We measure the degree of *isotropy* of an element $K \in \mathcal{T}_h$ by its smallest angle τ_K. If $\tau_K \approx \pi/3$ then K is almost isosceles, while if τ_K is small then K is thin, see Fig. 233.3. We use the smallest angle among the triangles in $\mathcal{T}_h$, i.e.

$$\tau = \min_{K \in \mathcal{T}_h} \tau_K$$

as a measure of the degree of anistropy of the triangulation $\mathcal{T}_h$. We shall see below that certain interpolation errors related to approximation with piecewise linear functions on a given triangulation get larger as τ tends to zero, corresponding to allowing the triangles to very thin.

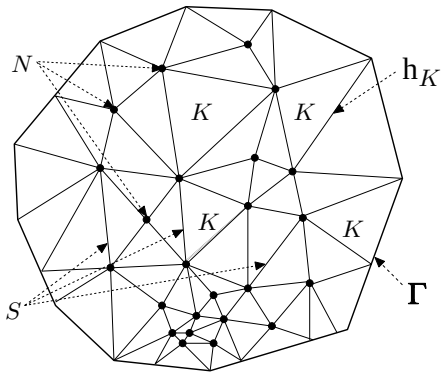


FIGURE 233.2. A triangulation of a domain Ω.

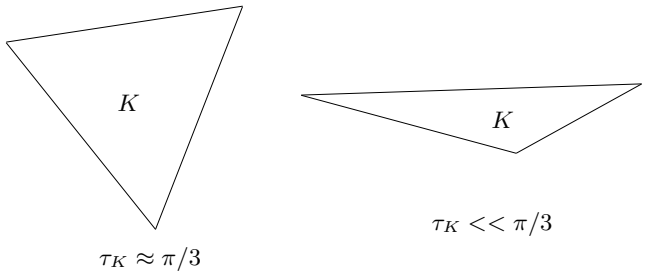


FIGURE 233.3. Measuring the isotropy of a triangle.

The basic problem of *mesh generation* is to generate a triangulation of a given domain with mesh size given by a prescribed mesh function $h(x)$. This problem arises in each step of an adaptive algorithm, where a new mesh function is computed from an approximate solution on a given mesh, and a new mesh is constructed with mesh size given by the new mesh function. The process is then repeated until a stopping criterion is satisfied. The new mesh may be constructed from scratch or by modification of the previous mesh including local refinement or coarsening.

In the *advancing front* strategy a mesh with given mesh size is constructed beginning at some point (often on the boundary) by successively adding one triangle after another, each with a mesh size determined by the mesh function. The curve dividing the domain into a part already triangulated and the remaining part is called the *front*. The front sweeps through the domain during the triangulation process. An alternative is to use a *h-refinement* strategy, where a mesh with a specified local mesh size is constructed by successively dividing elements of an initial coarse triangulation with the elements referred to as *parents*, into smaller elements, called the *children*. We illustrate the refinement and advancing front strategies in

Fig. 233.4. It is often useful to combine the two strategies using the advanc-

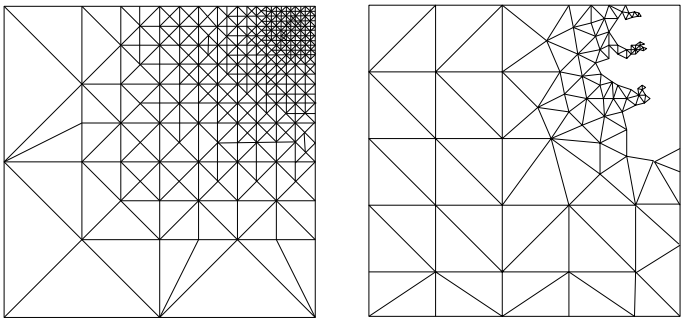


FIGURE 233.4. The mesh on the left is being constructed by successive h refinement starting from the coarse parent mesh drawn with thick lines. The mesh on the right is being constructed by an advancing front strategy. In both cases, high resolution is required near the upper right-hand corner.

ing front strategy to construct an initial mesh that represents the geometry of the domain with adequate accuracy, and use adaptive h-refinement.

There are various strategies for performing the division in an h-refinement aimed at limiting the degree of anisotropy of the elements. After the refinements are completed, the resulting mesh is fixed up by the addition of edges aimed at avoiding nodes that are located in the middle of element sides. This causes a mild "spreading" of the adapted region. We illustrate one technique for h-refinement in Fig. 233.5. In general, refining a mesh tends

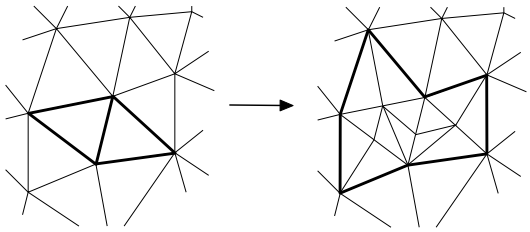


FIGURE 233.5. On the left, two elements in the mesh have been marked for refinement. The refinement uses the Rivara algorithm in which an element is divided into two pieces by inserting a side connecting the node opposite the longest side to the midpoint of the longest side. Additional sides are added to avoid having a node of one element on the side of another element. The refinement is shown in the mesh on the right along with the boundary of all the elements that had to be refined in addition to those originally marked for refinement.

to introduce elements with small angles, as can be seen in Fig. 233.5 and

it is an interesting problem to construct algorithms for mesh refinement that avoid this tendency in situations where the degree of anisotropy has to be limited. On the other hand, in certain circumstances, it is important to use "stretched" meshes that have regions of thin elements aligned together to give a high degree of refinement in one direction. In these cases, we also introduce mesh functions that give the local stretching, or degree of anisotropy, and the orientation of the elements. We discuss the construction and use of such meshes in the advanced companion volume.

233.3 Mesh Generation in $\mathbb{R}^3$

Mesh generation in three dimensions is analogous to that in two dimensions with the triangles being replaced by *tetrahedra*. In practice, the geometric constraints involved become more complicated and the number of elements also increases drastically. We show some examples in Fig. **??** and Fig. 233.6, and further examples in Fig. **??** and Fig. **??**.

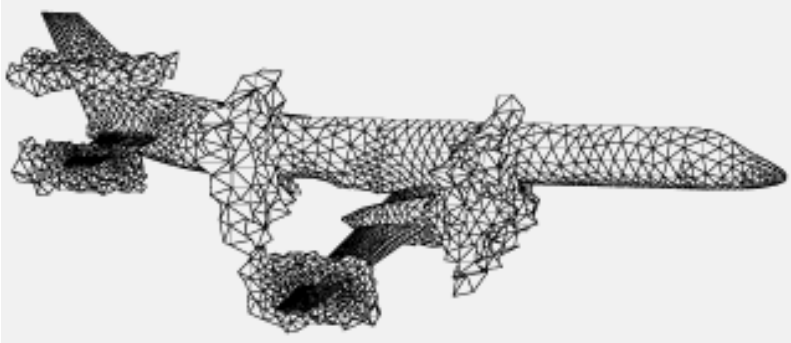


FIGURE 233.6. The surface mesh on the body, and parts of a tetrahedral mesh around a Saab 2000

233.4 Piecewise Linear Functions

Let $\mathcal{T}_h = \{K\}$ be a triangulation of a two-dimensional domain Ω with piecewise polynomial boundary Γ, let $\mathcal{N}_h = \{N\}$ denote the nodes of $\mathcal{T}_h$ and introduce the corresponding the finite dimensional vector space V_h consisting of the continuous piecewise linear functions on $\mathcal{T}_h$. In other words,

$$V_h = \big\{v : v \text{ is continuous on } \Omega, v|_K \in \mathcal{P}(K) \text{ for } K \in \mathcal{T}_h\big\},$$

where $\mathcal{P}(K)$ denotes the set of linear functions on K, i.e., the set of functions $v(x) = v(x_1, x_2)$ of the form $v(x) = c_0 + c_1 x_1 + c_2 x_2$ for some constants c_i. We can describe a function $v(x)$ in V_h by the nodal values $v(N)$ with $N \in \mathcal{N}_h$ because of two facts. The first is that a linear function is uniquely determined by its values at three points, as long as they don't lie on a straight line. To prove this claim, let $K \in \mathcal{T}_h$ have vertices $a^i = (a_1^i, a_2^i)$,

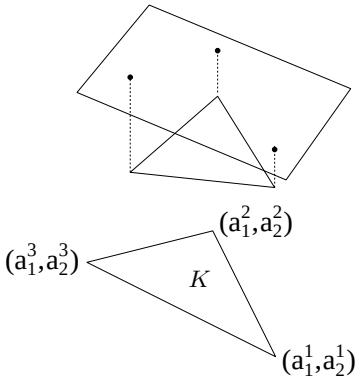


FIGURE 233.7. On the left, we show that the three nodal values on a triangle determine a linear function. On the right, we show the notation used to describe the nodes of a typical triangle.

$i = 1, 2, 3$, see Fig. 233.7. We want to show that $v \in \mathcal{P}(K)$ is determined uniquely by $\{v(a^1), v(a^2), v(a^3)\} = \{v_1, v_2, v_3\}$. A linear function v can be written $v(x_1, x_2) = c_0 + c_1 x_1 + c_2 x_2$ for some constants c_0, c_1, c_2. Substituting the nodal values of v into this expression yields a linear system of equations:

$$\begin{pmatrix} 1 & a_1^1 & a_2^1 \\ 1 & a_1^2 & a_2^2 \\ 1 & a_1^3 & a_1^3 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}.$$

The determinant of the coefficient matrix is equal to the determinant of the following matrix resulting from subtracting the first row from the second and third row:

$$\begin{pmatrix} 1 & a_1^1 & a_2^1 \\ 0 & a_1^2 - a_1^1 & a_2^2 - a_2^1 \\ 0 & a_1^3 - a_1^1 & a_2^2 - a_2^1 \end{pmatrix},$$

which is equal to the twice the area of the triangle K (up to the sign). The determinant of the coefficient matrix is thus non-zero, and we conclude that the system of equations(233.4) has a unique solution. We conclude that at linear function is uniquely specified by its values at three non-colinear points.

The second fact is that if a function is linear in each of two neighboring triangles and its nodal values on the two common nodes of the triangles are equal, then the function is continuous across the common edge. To

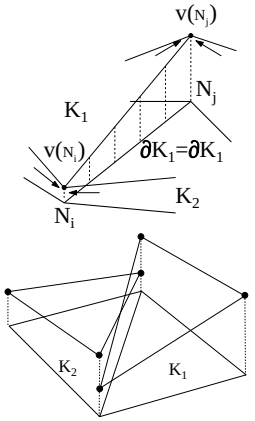


FIGURE 233.8. On the left, we show that a function that is piecewise linear on triangles reduces to a linear function of one variable on triangle edges. On the right, we plot a function that is piecewise linear on triangles whose values at the common nodes on two neighboring triangles do not agree.

see this, let K_1 and K_2 be adjoining triangles with common boundary $\partial K_1 = \partial K_2$; see the figure on the left in Fig. 233.8. Parameterizing v along this boundary, we see that v is a linear function of one variable there. Such functions are determined uniquely by the value at two points, and therefore since the values of v on K_1 and K_2 at the common nodes agree, the values of v on the common boundary between K_1 and K_2 agree, and v is indeed continuous across the boundary.

To construct a set of basis functions for V_h, we begin by describing a set of *element basis functions* for triangles. Once again, assuming that a triangle K has nodes at $\{a^1, a^2, a^3\}$, the element nodal basis is the set of functions $\lambda_i \in \mathcal{P}(K)$, $i = 1, 2, 3$, such that

$$\lambda_i(a^j) = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

We show these functions in Fig. 233.9.

We construct the *global* basis functions for V_h by piecing together the element basis functions on neighboring elements using the continuity requirement, i.e. by matching element basis functions on neighboring triangles that have the same nodal values on the common edge. The resulting set of basis functions $\{\varphi_j\}_{j=1}^M$, where N_1, N_2,..., N_M is an enumeration of the nodes $N \in \mathcal{N}_h$, is called the set of *tent* functions. The tent functions

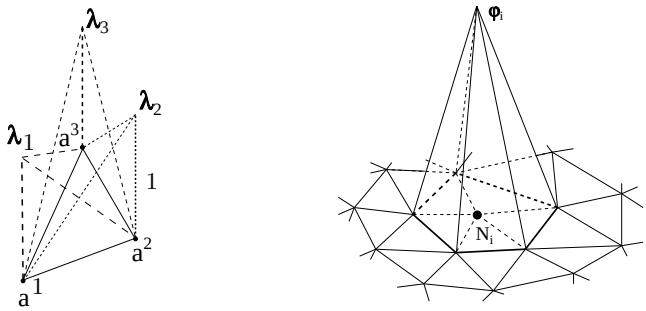


FIGURE 233.9. On the left, we show the three element nodal basis functions for the linear functions on K. On the right, we show a typical global basis "tent" function.

can also be defined by specifying that $\varphi_j \in V_h$ satisfy

$$\varphi_j(N_i) = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

for $i, j = 1, ..., M$. We illustrate a typical tent function in Fig. 233.9. We see in particular that the *support* of φ_i is the set of triangles that share the common node N_i.

The tent functions are a nodal basis for V_h because if $v \in V_h$ then

$$v(x) = \sum_{i=1}^{M} v(N_i)\varphi_i(x).$$

233.5 Max-Norm Error Estimates

In this section we prove the basic pointwise maximum norm error estimate for linear interpolation on a triangle, which states that the interpolation error depends on the second order partial derivatives of the function being interpolated, i.e. on the "curvature" of the function, the mesh size and the shape of the triangle. Analogous results hold for other norms. The results also extend directly to more than two space dimensions.

Let K be a triangle with vertices $a^i, i = 1, 2, 3$. Given a continuous function v defined on K, let the linear interpolant $\pi_K v \in \mathcal{P}(K)$ be defined by

$$\pi_K v(a^i) = v(a^i), \quad i = 1, 2, 3.$$

We illustrate this in Fig. 233.10.

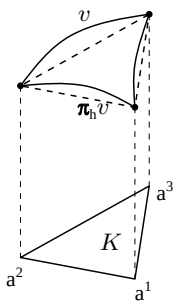


FIGURE 233.10. The nodal interpolant $\pi_K v$ of v.(change from π_h to π_K)

Theorem 233.1 *If v has continuous second derivatives, then*

$$\|v - \pi_K v\|_{L_\infty(K)} \leq \frac{1}{2} h_K^2 \|D^2 v\|_{L_\infty(K)}, \tag{233.1}$$

$$\|\nabla(v - \pi_K v)\|_{L_\infty(K)} \leq \frac{3}{\sin(\tau_K)} h_K \|D^2 v\|_{L_\infty(K)}, \tag{233.2}$$

where h_K is the largest side of K, τ_K is the smallest angle of K, and

$$D^2 v = \left(\sum_{i,j=1}^{2} \left(\frac{\partial^2 v}{\partial x_i \partial x_j} \right)^2 \right)^{1/2}.$$

If ∇v is continuous, then

$$\|v - \pi_K v\|_{L_\infty(K)} \leq h_K \|Dv\|_{L_\infty(K)}, \tag{233.3}$$

Note that the gradient estimate depends on the reciprocal of the sine of the smallest angle of K, and therefore this error bound deteriorates as the the triangle gets thinner.

The proof follows the same general outline as the proofs of corresponding results in the Chapter Piecewise linear approximation. Let λ_i, $i = 1, 2, 3$, be the element basis functions for $\mathcal{P}(K)$ defined by $\lambda_i(a^j) = 1$ if $i = j$, and $\lambda_i(a^j) = 0$ otherwise. A function $w \in \mathcal{P}(K)$ has the representation

$$w(x) = \sum_{i=1}^{3} w(a^i) \lambda_i(x) \quad \text{for } x \in K,$$

and thus

$$\pi_K v(x) = \sum_{i=1}^{3} v(a^i) \lambda_i(x) \quad \text{for } x \in K, \tag{233.4}$$

since $\pi_K v(a^i) = v(a^i)$. We shall derive representation formulas for the interpolation errors $v - \pi_K v$ and $\nabla(v - \pi_K v)$, using a Taylor expansion at $x \in K$:

$$v(y) = v(x) + \nabla v(x) \cdot (y - x) + R(x, y),$$

where

$$R(x,y) = \frac{1}{2}\sum_{i,j=1}^{2} \frac{\partial^2 v}{\partial x_i \partial x_j}(\xi)(y_i - x_i)(y_j - x_j),$$

is the remainder term of order 2 and ξ is a point on the line segment between x and y. In particular choosing $y = a^i = (a_1^i, a_2^i)$, we have

$$v(a^i) = v(x) + \nabla v(x)\cdot(a^i - x) + R_i(x), \tag{233.5}$$

where $R_i(x) = R(x, a^i)$. Inserting (233.5) into (233.4) gives for $x \in K$

$$\pi_K v(x) = v(x)\sum_{i=1}^{3}\lambda_i(x) + \nabla v(x)\cdot\sum_{i=1}^{3}(a^i - x)\lambda_i(x) + \sum_{i=1}^{3} R_i(x)\lambda_i(x). \tag{233.6}$$

We shall use the following identities that hold for $j,k = 1,2$, and $x \in K$,

$$\sum_{i=1}^{3}\lambda_i(x) = 1, \quad \sum_{i=1}^{3}(a_j^i - x_j)\lambda_i(x) = 0, \tag{233.7}$$

$$\sum_{i=1}^{3}\frac{\partial}{\partial x_k}\lambda_i(x) = 0, \quad \sum_{i=1}^{3}(a_j^i - x_j)\frac{\partial \lambda_i}{\partial x_k} = \delta_{jk}, \tag{233.8}$$

where $\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 0$ otherwise. The first of the identities in (233.7) follows by choosing $v(x) = 1$ in (233.6), and the second follows by choosing $v(x) = d_1x_1 + d_2x_2$ with $d_i \in \mathbb{R}$. Finally, (233.8) follows by differentiating (233.7).

Using (233.7), we obtain the following representation of the interpolation error,

$$v(x) - \pi_K v(x) = -\sum_{i=1}^{3} R_i(x)\lambda_i(x).$$

Since $|a^i - x| \le h_K$, we can estimate the remainder term $R_i(x)$ as

$$|R_i(x)| \le \frac{1}{2}h_K^2\|D^2 v\|_{L_\infty(K)}, \quad i = 1,2,3.$$

where we used Cauchy's inequality twice to estimate an expression of the form $\sum_{ij} x_i c_{ij} x_j = \sum_i x_i \sum_j c_{ij}x_j$.

Now, using the fact that $0 \le \lambda_i(x) \le 1$ if $x \in K$, for $i = 1,2,3$, we obtain

$$|v(x) - \pi_K v(x)| \le \max_i |R_i(x)| \sum_{i=1}^{3}\lambda_i(x) \le \frac{1}{2}h_K^2\|D^2 v\|_{L_\infty(K)} \quad \text{for } x \in K,$$

which proves (233.1).

To prove (233.2), we differentiate (233.4) with respect to $x_k, k = 1, 2$ to get

$$\nabla(\pi_K v)(x) = \sum_{i=1}^{3} v(a^i)\nabla\lambda_i(x),$$

which together with (233.5) and (233.8) gives the following error representation:

$$\nabla(v - \pi_K v)(x) = -\sum_{i=1}^{3} R_i(x)\nabla\lambda_i(x) \quad \text{for } x \in K.$$

We now note that

$$\max_{x\in K} |\nabla\lambda_i(x)| \leq \frac{2}{h_K \sin(\tau_K)},$$

which follows by an easy estimate of the shortest height (distance from a vertex to the opposite side) of K. We now obtain (233.2) and (233.3) finally follows using tha Mean Value theorem. The proof is now complete.

Let now $\mathcal{T}_h = \{K\}$ be a triangulation of a domain Ω with mesh function $h(x)$, and let π_h denote nodal interpolation into the corresponding space of continuous piecewise linear functions V_h on $\mathcal{T}_h$. The interpolation error estimates of Theorem 233.1 then take the form

$$\|v - \pi_h v\|_{L_\infty(\Omega)} \leq \frac{1}{2}\|h^2 D^2 v\|_{L_\infty(\Omega)}, \tag{233.9}$$

$$\|v - \pi_h v\|_{L_\infty(\Omega)} \leq \|hDv\|_{L_\infty(\Omega)}, \tag{233.10}$$

$$\|\nabla(v - \pi_h v)\|_{L_\infty(\Omega)} \leq \frac{3}{\sin(\tau)}\|hD^2 v\|_{L_\infty(\Omega)}, \tag{233.11}$$

where τ is the smallest of the τ_K. Below we shall use analogs of these estimates with the $L_\infty(\Omega)$ replaced by $L_2(\Omega)$.

FIGURE 233.11. Sergei Lvovich Sobolev (1908-1989), creator of Functional Analysis and inventor of Sobolev spaces:"I wonder if my space of functions $H^1(\Omega)$ is large enough to contain the solution?"

233.6 Sobolev and his Spaces

Sergei Sobolev (1908-1989) played a leading role in the mathematical world of the former Soviet Union and made important contributions to the theory and practice of partial differential equations, in particular on questions of existence, uniqueness, stability and regularity of solutions by developing tools of *Functional Analysis*. He also worked on numerical methods and gave important results on interpolation and quadrature of functions of several variables by developing techniques of *Sobolev spaces*. A basic Sobolev space is the space of real-valued functions defined on a domain Ω in $\mathbb{R}^d$, which are square integrable together with their first partial derivatives, denoted by $H^1(\Omega)$.

233.7 Quadrature in $\mathbb{R}^2$

To compute the stiffness matrix and load vector a FEM, we have to compute integrals of the form $\int_K g(x)\,dx$, where K is a triangle or tetrahedron and g a given function. Sometimes we may evaluate these integrals exactly, but usually this is either impossible or inefficient. Instead we usually evaluate the integrals approximately using quadrature formulas. We briefly present some quadrature formulas for integrals over triangles.

In general, we would like to use quadrature formulas that do not affect the accuracy of the underlying finite element method, which of course requires an estimate of the error due to quadrature. A quadrature formula for an integral over an element K has the form

$$\int_K g(x)\,dx \approx \sum_{i=1}^{q} g(y^i)\omega_i, \tag{233.12}$$

for a specified choice of *nodes* $\{y^i\}$ in K and *weights* $\{\omega_i\}$. We now list some possibilities using the notation a_K^i to denote the vertices of a triangle K, a_K^{ij} to denote the midpoint of the side connecting a_K^i to a_K^j, and a_K^{123} to denote the center of mass of K, and denote by $|K|$ the area of K:

$$\int_K g\,dx \approx g\big(a_K^{123}\big)|K|, \tag{233.13}$$

$$\int_K g(x)\,dx \approx \sum_{j=1}^{3} g(a_K^j)\frac{|K|}{3}, \tag{233.14}$$

$$\int_K g\,dx \approx \sum_{1\le i<j\le 3} g\big(a_K^{ij}\big)\frac{|K|}{3}, \tag{233.15}$$

$$\int_K g\,dx \approx \sum_{j=1}^{3} g\big(a_K^i\big)\frac{|K|}{20} + \sum_{1\le i<j\le 3} g\big(a_K^{ij}\big)\frac{2|K|}{15} + g\big(a_K^{123}\big)\frac{9|K|}{20}. \tag{233.16}$$

We refer to (233.13) as the center of gravity quadrature, to (233.14) as the vertex quadrature, and to (233.15) as the midpoint quadrature. Recall that the accuracy of a quadrature formula is related to the *precision* of the formula. A quadrature formula has precision r if the formula gives the exact value of the integral if the integrand is a polynomial of degree at most $r-1$, but there is some polynomial of degree r such that the formula is not exact. The quadrature error for a quadrature rule of precision r is proportional to h^r, where h is the mesh size. More precisely, the error of a quadrature rule of the form (233.12) satisfies

$$\left| \int_K g\,dx - \sum_{i=1}^{q} g(y^i)\omega_i \right| \leq C h_K^r \sum_{|\alpha|=r} \int_K |D^\alpha g|\,dx,$$

where C is a constant. Vertex and center of gravity quadrature have precision 2, midpoint quadrature has precision 3, while (233.16) has precision 4.

In finite element methods based on continuous piecewise linear functions, we often use nodal or vertex quadrature, often also referred to as *lumped mass* quadrature, because the mass matrix computed this way becomes diagonal.

EXAMPLE 233.1. In Fig. **??** and Fig. **??** we give two examples, one from fluid mechanics. and the other from electromagnetics.

Chapter 233 Problems

233.1. For a given triangle K, determine the relation between the smallest angle τ_K, the triangle diameter h_K and the diameter ρ_K of the largest inscribed circle.

233.2. Draw the refined mesh that results from sub-dividing the smallest two triangles in the mesh on the right in Fig. 233.5.

233.3. Let K be a tetrahedron with vertices $\{a^i, i = 1, ..., 4\}$. Show that a linear polynomial $v(x) = c_0 + c_1x_1 + c_2x_2 + c_3x_3$ on K is uniquely determined by the nodal values $\{v(a^i), i = 1, ..., 4\}$. Show that the corresponding finite element space V_h consists of continuous functions.

233.4. Prove that the quadrature formulas (233.13),(233.14), (233.15) and (233.16) have the indicated precision.

233.5. Prove that using nodal quadrature to compute a mass matrix for piecewise linears, gives a diagonal mass matrix where a diagonal term is the sum of the terms in the corresponding row in the exactly computed mass matrix. Motivate the term "lumped".

234

FEM for Boundary Value Problems in $\mathbb{R}^2$ and $\mathbb{R}^3$

> ...were very confused, skipping suddenly from one idea to another, from one formula to the next, with no attempt to give a connection between them. His presentations were obscure clouds, illuminated from time to time by flashes of pure genius. ... of the thirty who enrolled with me, I was the only one to see it through. (Menabrea about Cauchy 1832)

234.1 Introduction

In this chapter, we extend the cG(1) FEM for reaction-diffusion-convection problems in one space dimension to corresponding boundary value problems in $\mathbb{R}^2$ and $\mathbb{R}^3$ of the form: Find $u : \Omega \to \mathbb{R}$ such that

$$-\nabla \cdot (a\nabla u) + \nabla \cdot (ub) + cu = f \quad \text{in } \Omega, \tag{234.1}$$

together with boundary conditions of Dirichlet, Neumann or Robin type, where $a(x) > 0$, $b(x)$ and $c(x)$ are given variable coefficients, $f(x)$ is a given right hand side, and Ω is a bounded open domain in $\mathbb{R}^2$ or $\mathbb{R}^3$. Note that the coefficient b is a vector (typically corresponding to a given convection velocity), and that the equation (234.1) can alternatively be written

$$-\nabla \cdot (a\nabla u) + b \cdot \nabla u + \hat{c}u = f \quad \text{in } \Omega, \tag{234.2}$$

with $\hat{c} = c + \nabla \cdot b$. In general, problems of this form cannot be solved analytically and we have to rely on a numerical method such as FEM for computing the solution $u(x)$ for given data.

We consider below the extension to corresponding time dependent problems of the form

$$\dot{u} - \nabla \cdot (a\nabla u) + \nabla \cdot (ub) + cu = f, \tag{234.3}$$

together with initial and boundary value problems, including extensions to systems of such equations, using the material in the Chapters *The General Initial Value Problem* and *Adaptive IVP-Solvers*.

The most fundamental example of the form (234.1) is Poisson's equation with homogeneous Dirichlet boundary conditions corresponding to setting $a = 1$, $b = 0$ and $c = 0$:

$$\begin{cases} -\Delta u(x) = f(x) & \text{for } x \in \Omega, \\ u(x) = 0 & \text{for } x \in \Gamma, \end{cases} \tag{234.4}$$

where Ω is a bounded domain in $\mathbb{R}^2$ with polygonal boundary Γ. We recall that $\nabla \cdot (\nabla u) = \Delta u$. We shall now present the cG(1) method for (240.1) generalizing cG(1) for the two-point boundary value problem (216.9), and then extend to the general problem (234.1).

FIGURE 234.1. Richard Courant (1888-1972), pioneer of finite elements: "In fact, already when writing my 1910 Ph D thesis on using the Dirichlet minimum principle to prove the existence of solutions to Poisson's equation on a domain Ω, I had in mind of seeking approximate solutions in a subspace of the Sobolev space $H^1(\Omega)$ consisting of piecewise linear functions on a triangulation of Ω...".

234.2 Richard Courant: Inventor of FEM

Richard Courant (1888-1972) was a student of Hilbert and published together with him the monumental work *Methoden der Mathematischen Physik*. In the mid 1930s he fled to New York away from the Nazis and created the Courant Institute of Mathematical Sciences, since 1964 occupying a 13 storey building close to Washington Square in Greenwich Village on

Manhattan. Courant presented in a famous paper from 1943 the basics of finite element approximation of differential equations, as an expansion of a foot-note in the 1924 *Methoden*. This foot-note must be one of the most productive remarks in the history of science generating hundreds of thousands of scientific articles and a flood of software from the mid 1960s and on.

234.3 Variational Formulation

We let $\mathcal{T}_h = \{K\}$ be a triangulation of Ω with mesh function $h(x)$ and internal nodes $N_1, ...N_M$, and we let V_h be the corresponding finite element space of continuous piecewise linear functions that vanish on the boundary Γ. We first give (240.1) the following preliminary variational formulation:

$$-\int_\Omega \Delta u\, v\, dx = \int_\Omega f\, v\, dx \tag{234.5}$$

for all suitable test functions v, which results from multiplying (240.1) by $v(x)$ and integrating over Ω. We now want to rewrite the left-hand side to move a derivative from Δu onto v. Assuming that the test function v is zero on Γ, Green's formula implies

$$-\int_\Omega \Delta u\, v\, dx = -\int_\Gamma \partial_n u v\, ds + \int_\Omega \nabla u \cdot \nabla v\, dx = \int_\Omega \nabla u \cdot \nabla v\, dx,$$

where $\partial_n = \frac{\partial}{\partial n}$ denotes the outward unit normal derivative on Γ. We find that a solution $u(x)$ of (240.1) satisfies

$$\int_\Omega \nabla u \cdot \nabla v\, dx = \int_\Omega f\, v\, dx, \tag{234.6}$$

for all test functions v with $v = 0$ on Γ.

234.4 The cG(1) FEM

We are thus led to the following formulation of the cG(1) FEM for (240.1): Find $U \in V_h$ such that

$$\int_\Omega \nabla U \cdot \nabla v\, dx = \int_\Omega f\, v\, dx \quad \text{for all } v \in V_h, \tag{234.7}$$

where V_h is the space of continuous piecewise linear functions on a triangulation $\mathcal{T}_h$ of Ω that vanish on the boundary Γ. Using the notation

$$(w, v) = \int_\Omega wv\, dx, \quad (\nabla w, \nabla v) = \int_\Omega \nabla w \cdot \nabla v\, dx,$$

we can write cG(1) in the form: Find $U \in V_h$ such that

$$(\nabla U, \nabla v) = (f, v) \quad \text{for all } v \in V_h. \tag{234.8}$$

We see that that the trial space and test spaces are equal ($= V_h$) and include the homogenous Dirichlet boundary condition. The Galerkin orthogonality is expressed by

$$(\nabla u - \nabla U, \nabla v) = 0 \quad \text{for all } v \in V_h, \tag{234.9}$$

which results upon subtracting (240.6) from (240.2) with $v \in V_h$.

We recall that the nodal basis functions $\{\varphi_i\}_{i=1}^M$ associated with the internal nodes $N_1, ...N_M$ of $\mathcal{T}_h$ is a basis for V_h. Expressing U in terms of this basis,

$$U(x) = \sum_{j=1}^{M} U(N_j)\varphi_j(x), \tag{234.10}$$

substituting into (240.6), and choosing $v = \varphi_i$ for $i = 1, ..., M$, gives

$$\sum_{j=1}^{M} (\nabla\varphi_j, \nabla\varphi_i) U(N_j) = (f, \varphi_i), \quad i = 1, ..., M.$$

This is equivalent to the linear system of equations

$$A\xi = b, \tag{234.11}$$

where $\xi = (\xi_j)$ is the vector of internal nodal values $\xi_j = U(N_j)$, $A = (a_{ij})$ is the *stiffness matrix* with elements $a_{ij} = (\nabla\varphi_j, \nabla\varphi_i)$ and $b = (b_i)$ with $b_i = (f, \varphi_i)$ is the *load vector*..

The stiffness matrix A is obviously symmetric and it is also positive-definite since for any $v = \sum_{i=1}^M \eta_i\varphi_i$ in V_h,

$$\sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j = \sum_{i,j=1}^{M} \eta_i (\nabla\varphi_i, \nabla\varphi_j)\eta_j$$
$$= \left(\nabla \sum_{i=1}^{M} \eta_i\varphi_i, \nabla \sum_{j}^{M} \eta_j\varphi_j \right) = (\nabla v, \nabla v) > 0,$$

unless $\eta_i = 0$ for all i. This means in particular that (240.8) has a unique solution vector U and thus the cG(1) finite element problem (240.6) has a unique solution $U \in V_h$.

A triangle with associated linear approximation, i.e. the basic *finite element* of cG(1), is also called the *Courant element*, as a recognition of its inventor.

Uniform Triangulation of a Square

We compute the stiffness matrix A and load vector b in (240.8) explicitly on the uniform triangulation of the square $\Omega = [0,1] \times [0,1]$ pictured in Fig. 240.1. We choose an integer $m \geq 1$ and set $h = 1/(m+1)$, then construct the triangles as shown. The diameter of the triangles in $\mathcal{T}_h$ is $\sqrt{2}h$ and there are $M = m^2$ internal nodes. We number the nodes starting from the lower left and moving right, then working up across the rows.

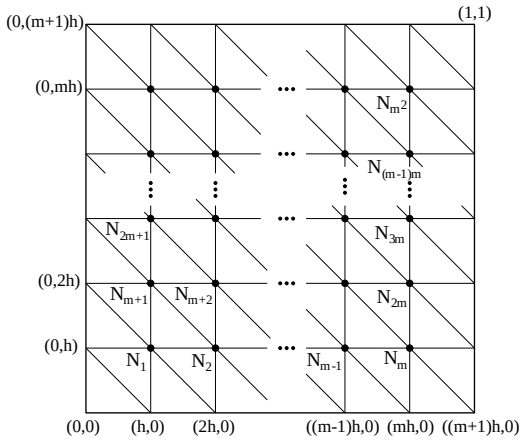


FIGURE 234.2. The standard triangulation of the unit square.

In Fig. 240.2, we show the support of the basis function corresponding to the node N_i along with parts of the basis functions for the neighboring nodes. As in one dimension, the basis functions are "almost" orthogonal in

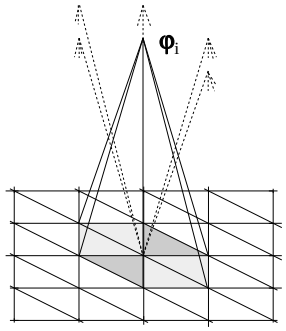


FIGURE 234.3. The support of the basis function φ_i together with parts of the neighboring basis functions.

the sense that only basis functions φ_i and φ_j sharing a common triangle

in their supports yield a non-zero value in $(\nabla\varphi_i, \nabla\varphi_j)$. We show the nodes neighboring N_i in Fig. 240.3. The support of any two neighboring basis

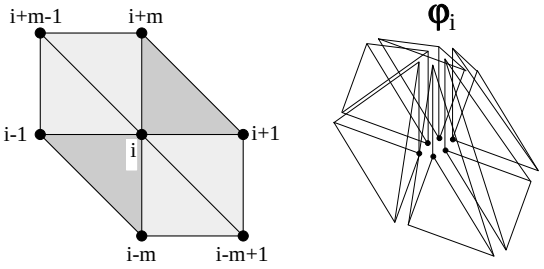


FIGURE 234.4. The indices of the nodes neighboring N_i and an "exploded" view of φ_i.

functions overlap on just two triangles, while a basis function "overlaps itself" on six triangles.

We first compute

$$a_{ii} = (\nabla\varphi_i, \nabla\varphi_i) = \int_\Omega |\nabla\varphi_i|^2\, dx = \int_{\text{support of } \varphi_i} |\nabla\varphi_i|^2\, dx,$$

for $i = 1, ..., m^2$. As noted, we only have to consider the integral over the domain pictured in Fig. 240.3, which is written as a sum of integrals over the six triangles making up the domain. Examining φ_i on these triangles, see Fig. 240.3, we see that there are only two different integrals to be computed since φ_i looks the same, except for orientation, on two of the six triangles and similarly the same on the other four triangles. We shade the corresponding triangles in Fig. 240.2. The orientation affects the direction of $\nabla\varphi_i$ of course, but does not affect $|\nabla\varphi_i|^2$.

We compute $(\nabla\varphi_i, \nabla\varphi_i)$ on the triangle shown in Fig. 240.4. In this case, φ_i is one at the node located at the right angle in the triangle and zero at the other two nodes. We change coordinates to compute $(\nabla\varphi_i, \nabla\varphi_i)$ on the *reference triangle* shown in Fig. 240.4. Again, changing to these coordinates does not affect the value of $(\nabla\varphi_i, \nabla\varphi_i)$ since $\nabla\varphi_i$ is constant on the triangle. On the triangle, φ_i can be written $\varphi_i = ax_1 + bx_2 + c$ for some constants a, b, c. Since $\varphi_i(0,0) = 1$, we get $c = 1$. Similarly, we compute a and b to find that $\varphi_i = 1 - x_1/h - x_2/h$ on this triangle. Therefore, $\nabla\varphi_i = \left(-h^{-1}, -h^{-1}\right)$ and the integral is

$$\int_\triangleright |\nabla\varphi_i|^2\, dx = \int_0^h \int_0^{h-x_1} \frac{2}{h^2}\, dx_2\, dx_1 = 1.$$

In the second case, φ_i is one at a node located at an acute angle of the triangle and is zero at the other nodes. We illustrate this in Fig. 240.5.

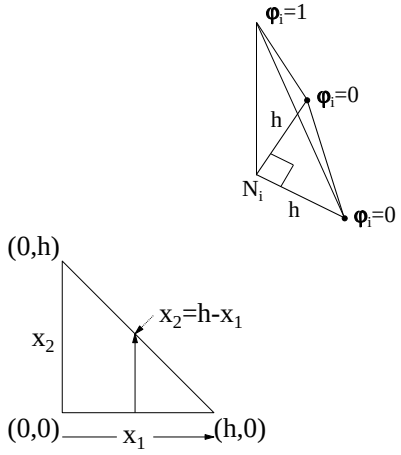


FIGURE 234.5. First case showing φ_i on the left together with the variables used in the reference triangle.

We use the coordinate system shown in Fig. 240.5 to write $\varphi_i = 1 - x_1/h$. When we integrate over the triangle, we get $1/2$.
Summing the contributions from all the triangles gives

$$a_{ii} = (\nabla\varphi_i, \nabla\varphi_i) = 1 + 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 4.$$

Next, we compute $(\nabla\varphi_i, \nabla\varphi_j)$ for indices corresponding to neighboring nodes. For a general node N_i, there are two cases of inner products (see Fig. 240.2 and Fig. 240.3):

$$a_{i\,i-1} = (\nabla\varphi_i, \nabla\varphi_{i-1}) = (\nabla\varphi_i, \nabla\varphi_{i+1}) = (\nabla\varphi_i, \nabla\varphi_{i-m}) = (\nabla\varphi_i, \nabla\varphi_{i+m}),$$

and

$$a_{i\,i-m+1} = (\nabla\varphi_i, \nabla\varphi_{i-m+1}) = (\nabla\varphi_i, \nabla\varphi_{i+m-1}).$$

The orientation of the triangles in each of the two cases are different, but the inner product of the gradients of the respective basis functions is not affected by the orientation. Note that the the equations corresponding to nodes next to the boundary are special, because the nodal values on the boundary are zero, see Fig. 240.1. For example, the equation corresponding to N_1 only involves N_1, N_2 and N_{m+1}.

For the first case, we next compute $(\nabla\varphi_i, \nabla\varphi_{i+1})$. Plotting the intersection of the respective supports shown in Fig. 240.6, we conclude that there are equal contributions from each of the two triangles in the intersection. We choose one of the triangles and construct a reference triangle as above. Choosing suitable variables, we find that

$$\nabla\varphi_i \cdot \nabla\varphi_{i+1} = \big(-\frac{1}{h}, -\frac{1}{h}\big) \cdot \big(\frac{1}{h}, 0\big) = -\frac{1}{h^2},$$

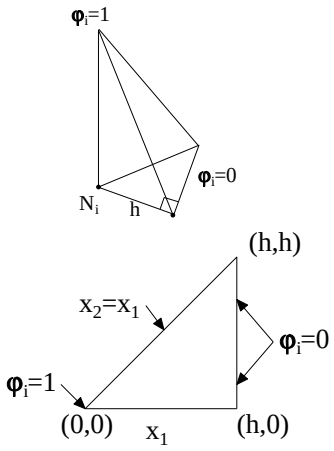


FIGURE 234.6. Second case showing φ_i and the reference triangle.

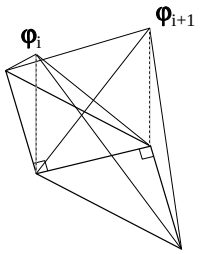


FIGURE 234.7. The overlap of φ_i and φ_{i+1}.

and integrating over the triangle gives $-1/2$. Similarly, we see that

$$(\nabla\varphi_i, \nabla\varphi_{i-m+1}) = (\nabla\varphi_i, \nabla\varphi_{i+m-1}) = 0.$$

We can now determine the stiffness matrix A using the information above. We start by considering the first row. The first entry is $(\nabla\varphi_1, \nabla\varphi_1) = 4$ since N_1 has no neighbors to the left or below. The next entry is $(\nabla\varphi_1, \nabla\varphi_2) = -1$. The next entry after that is zero, because the supports of φ_1 and φ_3 do not overlap. This is true in fact of all the entries up to and including φ_m. However, $(\nabla\varphi_1, \nabla\varphi_{m+1}) = -1$, since these neighboring basis functions do share two supporting triangles. Finally, all the rest of the entries in that row are zero because the supports of the corresponding basis functions do not overlap. We continue in this fashion working row by row. The result is pictured in Fig. 240.7. We see that A has a *block structure* consisting of banded $m \times m$ sub-matrices, most of which consist only of zeros. Note the pattern of entries around corners of the diagonal block matrices; it is a common mistake to program these values incorrectly.

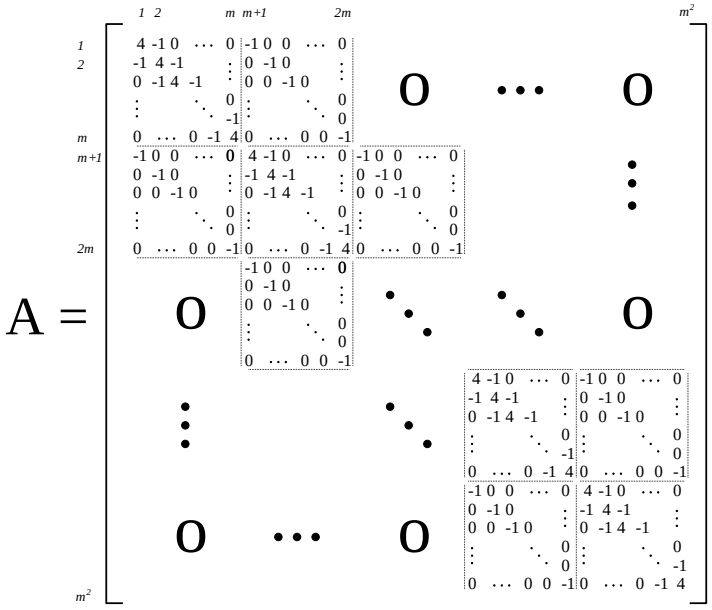


FIGURE 234.8. The stiffness matrix.

The storage of a sparse matrix and the solution of a sparse system are both affected by the *structure* or *sparsity pattern* of the matrix. The sparsity pattern is affected in turn by the enumeration scheme used to mark the nodes.

There are several algorithms for reordering the coefficients of a sparse matrix to form a matrix with a smaller bandwidth. Reordering the coefficients is equivalent to computing a new basis for the vector space.

The load vector b is computed in the same fashion, separating each integral

$$\int_{\Omega} f\varphi_i \, dx = \int_{\text{support of } \varphi_i} f(x)\varphi_i(x) \, dx$$

into integrals over the triangles making up the support of φ_i. To compute the elements (f, φ_i) of the load vector, we often use one of the quadrature formulas presented in Chapter **??**.

234.5 Basic Data Structures

To compute the finite element approximation U, we have to compute the coefficients of the stiffness matrix A and load vector b and solve the linear system of equations (240.8). We just computed A and b for a uniform

triangulation of the unit square, and we now discuss the case of a general triangulation of a general domain.

We have to compute the non-zero elements $a_{ij} = (\nabla\varphi_j, \nabla\varphi_i)$ of the stiffness matrix A. We know that $a_{ij} = 0$ unless both N_i and N_j are nodes of the same triangle K, because only then the supports of basis functions φ_i and φ_j overlap. The common support corresponding to a non-zero element a_{ij} is equal to the support of φ_i if $j = i$, and is equal to the two triangles with the common edge connecting N_j and N_i if $i \neq j$. In each case a_{ij} is the sum of contributions

$$a_{ij}^K = \int_K \nabla\varphi_j \cdot \nabla\varphi_i \, dx \tag{234.12}$$

over the triangles K in the common support. The process of adding up the contributions a_{ij}^K from the relevant triangles K to get the element a_{ij}, is called *assembling* the stiffness matrix A. Arranging for a given triangle K the numbers a_{ij}^K, where N_i and N_j are nodes of K, into a 3×3 matrix, we obtain the *element stiffness matrix* for the triangle K. We refer to the assembled matrix A as the *global stiffness matrix*. Notice that we use *element* with two different meanings: as an element a_{ij} of the stiffness matrix A, and as a finite element or triangle of the triangulation.

To compute the element stiffness matrix a_{ij}^K for a given triangle K, we need the physical coordinates of the nodes of K. To perform the assembly where we loop over all elements and add the corresponding contributions to the global stiffness matrix, we need the node numbers of each triangle. Similar information is needed to compute the load vector. The required information is arranged in a *data structure*, or data base, containing (i) a list of the coordinates of the nodes numbered in some way, and (ii) a list of the node numbers of each triangle. A list of the numbers of the nodes on the boundary is also needed to handle the boundary conditions This information is typically the output of the mesh generator generating a triangulation of the domain.

234.6 Solving the Discrete System

Once we have assembled the stiffness matrix A and computed the load vector b, we have to solve the linear system $AU = b$ to obtain the finite element approximation $U(x)$. We now discuss this topic briefly based on the material presented in Chapter 94. The stiffness matrix resulting from discretizing the Laplacian is symmetric and positive-definite and therefore invertible. These properties also mean that there is a wide choice in the methods used to solve the linear system $AU = b$, which take advantage of the fact that A is sparse.

In the case of the standard uniform discretization of a square, we saw that A is a banded matrix with five non-zero diagonals and bandwidth $m + 1$,

where m is the number of nodes on a side. The dimension of A is m^2 and the asymptotic operations count for using Gaussian elimination is $O\big(m^4\big) = O\big(h^{-4}\big)$. Note that even though A has mostly zero diagonals inside the band, fill-in occurs as the elimination is performed, so we may as well treat A as if it has non-zero diagonals throughout the band. Clever rearrangement of A to reduce the amount of fill-in leads to a solution algorithm with an operations count on the order of $O(m^3) = O(h^{-3})$. In contrast, if we treat A as a full matrix, we get an asymptotic operations count of $O\big(h^{-6}\big)$, which is considerably larger for a large number of elements.

In general, we get a sparse stiffness matrix, though there may not be a band structure. If we want to use a Gaussian elimination method efficiently in general, then it is necessary to first reorder the system to bring the matrix into banded form.

We can also apply both the Jacobi and Gauss-Seidel methods to solve the linear system arising from discretizing the Poisson equation. In the case of the uniform standard discretization of a square for example, the operations count is $O\big(M\big)$ per iteration for both methods if we make use of the sparsity of A. Therefore a single step of either method is much cheaper than a direct solve. The question is: How many iterations do we need to compute in order to obtain an accurate solution?

Typically the spectral radius of the iteration matrix of the Jacobi or Gauss-Seidel method is equal to $1 - Ch^2$ with C some moderate positive constant, which means that the convergence rate quickly gets slow as h decreases: to reduce the error a certain factor, we need of the order of $O(h^{-2})$ iterations, and since each iteration takes $O(h^{-2})$ operations, the total number of operations is $O(h^{-4})$, which is the same order as using a banded Gaussian elimination solver.

There has been a lot of activity in developing iterative methods that converge more quickly than Jacobi and Gauss-Seidel. In recent years, very efficient *multi-grid methods* have been developed and are now becoming a standard tool. A multi-grid method is based on a sequence of Gauss-Seidel or Jacobi steps performed on a hierarchy of successively coarser meshes and are optimal in the sense that the solution work is proportional to the total number of unknowns (that is h^{-2} in the model problem).

234.7 An Equivalent Minimization Problem

The variational problem (240.6) is equivalent to the following quadratic *minimization problem*: find $U \in V_h$ such that

$$F(u) \leq F(v) \quad \text{for all } v \in V_h, \tag{234.13}$$

where

$$F(v) = \frac{1}{2}\int_\Omega |\nabla v|^2\,dx - \int_\Omega fv\,dx = \frac{1}{2}(\nabla v, \nabla v) - (f, v).$$

The quantity $F(v)$ may be interpreted as the *total energy* of the function $v(x)$ composed of the *internal energy* $\frac{1}{2}(\nabla v, \nabla v)$ and the *load potential* $-(f, v)$. Thus, the solution U minimizes the total energy $F(v)$ with v varying over V_h.

To see the equivalence of (240.6) and (240.5), we assume first that $U \in V_h$ satisfies (240.6). Let then $v \in V_h$ and write $v = U + (v - U) = U + w$ with $w = v - U \in V_h$. Using $(\nabla U, \nabla w) = (\nabla w, \nabla U)$, we get

$$F(v) = F(U + w) = \frac{1}{2}(\nabla U, \nabla U) + (\nabla U, \nabla w) + \frac{1}{2}(\nabla w, \nabla w) - (f, U) - (f, w)$$
$$= F(U) + \frac{1}{2}(\nabla w, \nabla w) \geq F(U),$$

with equality only if $w = 0$. We conclude that U satisfies (240.5).

Conversely, if U is the solution of (240.5), then we have for all $v \in V_h$

$$g_v(\epsilon) = F(U + \epsilon v) \geq g(0) = F(U) \quad \text{for all } \epsilon \in \mathbb{R},$$

and thus $\epsilon = 0$ is an interior minimum point of $g_v(\epsilon)$ with v fixed, and thus $g_v'(0) = 0$. Computing we get

$$0 = g_v'(0) = (\nabla U, \nabla v) - (f, v)$$

and thus U satisfies (240.6). We sum up in the following theorem:

Theorem 234.1 *The problems* (240.6) *and* (240.5) *are equivalent in the sense that they have the same unique solution.*

234.8 An Energy Norm A Priori Error Estimate

In this section, we derive a priori and a posteriori estimates of the error $u - U$ in the *energy norm* $\|\nabla(u - U\|$ with

$$\|\nabla v\| = (\int_\Omega |\nabla v|^2 \, dx)^{1/2}, \tag{234.14}$$

where u is the exact solution and U a finite element solution of Poisson's equation with homogeneous Dirichlet boundary conditions. The energy norm, which is the L_2 norm of the gradient of a function in this problem, arises naturally in the error analysis of the finite element method because it is closely tied to the variational problem. The gradient of the solution, representing heat flow, electric field, flow velocity, or stress for example, can be a variable of physical interest as much as the solution itself, representing temperature, potential or displacement for example, and in this case, the energy norm is the relevant error measure.

We first prove that the Galerkin finite element approximation is the best approximation of the true solution in V_h with respect to the energy norm.

Theorem 234.2 *Assume that u satisfies the Poisson equation (240.1) and U is the Galerkin finite element approximation satisfying (240.6). Then*

$$\|\nabla(u-U)\| \leq \|\nabla(u-v)\| \quad \text{for all } v \in V_h. \tag{234.15}$$

The proof goes as follows: Using the Galerkin orthogonality (234.9) with v replaced by $U - v \in V_h$, we can write

$$\|\nabla e\|^2 = (\nabla e, \nabla(u-U)) = (\nabla e, \nabla(u-U)) + (\nabla e, \nabla(U-v)).$$

Adding the terms involving U on the right, whereby U drops out, and using Cauchy's inequality, we get

$$\|\nabla e\|^2 = (\nabla e, \nabla(u-v)) \leq \|\nabla e\|\,\|\nabla(u-v)\|,$$

which proves the theorem after dividing by $\|\nabla e\|$.

Choosing $v = \pi_h u$ and using an $L_2(\Omega)$ analog of the interpolation estimate (233.11), we get the following quantitative a priori error estimate (with $\|v\| = \|v\|_{L_2(\Omega)}$):

Corollary 234.3 *There exists a constant C_i depending only on the minimal angle τ in $\mathcal{T}_h$, such that*

$$\|\nabla(u-U)\| \leq C_i \|hD^2 u\|. \tag{234.16}$$

234.9 An Energy Norm A Posteriori Error Estimate

We now prove an a posteriori error estimate following the strategy used for the two-point boundary value problem in Chapter **??**. A new feature occurring in higher dimensions is the appearance of integrals over the internal edges S in $\mathcal{S}_h$. We start by writing an equation for the error $e = u - U$ using (240.2) and (240.6) to get

$$\begin{aligned}\|\nabla e\|^2 &= (\nabla(u-U), \nabla e) = (\nabla u, \nabla e) - (\nabla U, \nabla e)\\ &= (f, e) - (\nabla U, \nabla e) = (f, e - \pi_h e) - (\nabla U, \nabla(e - \pi_h e)),\end{aligned}$$

where $\pi_h e \in V_h$ is an interpolant of e. We now break up the integrals over Ω into sums of integrals over the triangles K in $\mathcal{T}_h$ and integrate by parts over each triangle in the last term to get

$$\|\nabla e\|^2 = \sum_K \int_K (f + \Delta U)(e - \tilde{\pi}_h e)\,dx - \sum_K \int_{\partial K} \frac{\partial U}{\partial n_K}(e - \pi_h e)\,ds, \tag{234.17}$$

where $\partial U/\partial n_K$ denotes the derivative of U in the outward normal direction n_K of the boundary ∂K of K. In the boundary integral sum in (240.16),

each internal edge $S \in \mathcal{S}_h$ occurs twice as a part of each of the boundaries ∂K of the two triangles K that have S as a common side. Of course the outward normals n_K from each of the two triangles K sharing S point in opposite directions. For each side S, we choose one of these normal directions and denote by $\partial_S v$ the derivative of a function v in that direction on S. We note that if $v \in V_h$, then in general $\partial_S v$ is different on the two triangles sharing S; see Fig. 233.8, which indicates the "kink" over S in the graph of v. We can express the sum of the boundary integrals in (240.16) as a sum of integrals over edges of the form

$$\int_S [\partial_S U](e - \pi_h e)\, ds,$$

where $[\partial_S U]$ is the difference, or jump, in the derivative $\partial_S U$ computed from the two triangles sharing S. The jump appears because the outward normal directions of the two triangles sharing S are opposite. We further note that $e - \tilde{\pi}_h e$ is continuous across S, but in general does not vanish on S even if π_h is the nodal interpolant. This is different than the one-dimensional case, where the corresponding sum over nodes does indeed vanish because $e - \pi_h e$ vanishes at the nodes. We may thus rewrite (240.16) as follows with the second sum replaced by a sum over internal edges S:

$$\|\nabla e\|^2 = \sum_K \int_K (f + \Delta U)(e - \pi_h e)\, dx + \sum_{S \in \mathcal{S}_h} \int_S [\partial_S U](e - \pi_h e)\, ds.$$

Next, we return to a sum over element edges ∂K by just distributing each jump equally to the two triangles sharing it, to obtain an *error representation* of the energy norm of the error in terms of the residual error:

$$\|\nabla e\|^2 = \sum_K \int_K (f + \Delta U)(e - \pi_h e)\, dx + \sum_K \frac{1}{2} \int_{\partial K} h_K^{-1} [\partial_S U](e - \pi_h e) h_K\, ds,$$

where we have prepared to estimate the second sum by inserting a factor h_K and compensating. In crude terms, the residual error results from substituting U into the differential equation $-\Delta u - f = 0$, but in reality straightforward substitution is not possible because U is not twice differentiable in Ω. The integral on the right over K is the remainder from substituting U into the differential equation inside each triangle K, while the integral over ∂K arises because $\partial_S U$ in general is different when computed from the two triangles sharing S.

We estimate the first term in the error representation by inserting a factor h, compensating and using the estimate $\|h^{-1}(e - \pi_h e)\| \leq C_i \|\nabla e\|$ analogous to (233.11), to obtain

$$\begin{aligned} &|\sum_K \int_K h(f + \Delta U) h^{-1}(e - \pi_h e)\, dx| \\ &\qquad \leq \|hR_1(U)\| \|h^{-1}(e - \pi_h e)\| \leq C_i \|hR_1(U)\| \|\nabla e\|, \end{aligned}$$

where $R_1(U)$ is the function defined on Ω by setting $R_1(U) = |f + \Delta U|$ on each triangle $K \in \mathcal{T}_h$. We estimate the contribution from the jumps on the edges similarly. Formally, the estimate results from replacing $h_K\, ds$ by dx corresponding to replacing the integrals over element boundaries ∂K by integrals over elements K. Dividing by $\|\nabla e\|$, we obtain the following a posteriori error estimate:

Theorem 234.4 *There is an interpolation constant C_i only depending on the minimal angle τ such that the error of the Galerkin finite element approximation U of the solution u of the Poisson equation satisfies*

$$\|\nabla u - \nabla U\| \leq C_i \|hR(U)\|, \tag{234.18}$$

where $R(U) = R_1(U) + R_2(U)$ with

$$R_1(U) = |f + \Delta U| \quad \text{on } K \in \mathcal{T}_h,$$
$$R_2(U) = \frac{1}{2} \max_{S \subset \partial K} h_K^{-1} \big|[\partial_S U]\big| \quad \text{on } K \in \mathcal{T}_h.$$

Note that $R_1(U)$ is the contribution to the total residual from the interior of the elements K. In the present case of piecewise linear approximation, $R_1(U) = |f|$. Further, $R_2(U)$ is the contribution to the residual from the jump of the normal derivative of U across edges. In the one dimensional problem considered in Chapter **??**, this contribution does not appear because the interpolation error may be chosen to be zero at the node points.

234.10 Adaptive Error Control

The basic goal of adaptive error control is to find a triangulation $\mathcal{T}_h$ with a least number of nodes such that the corresponding finite element approximation U satisfies

$$\|\nabla u - \nabla U\| \leq \text{ TOL}. \tag{234.19}$$

Using the a posteriori error estimate we are thus led to find a triangulation $\mathcal{T}_h$ with a least number of nodes such that the corresponding finite element approximation U satisfies

$$C_i \|hR(U)\| \leq \text{ TOL}. \tag{234.20}$$

This is a nonlinear constrained minimization problem with U depending on $\mathcal{T}_h$. If (240.17) is a reasonably sharp estimate of the error, then a solution of this optimization problem will meet our original goal.

We cannot expect to be able to solve this minimization problem analytically. Instead, a solution has to be sought by an iterative process in which we start with a coarse initial mesh and then successively modify the mesh by seeking to satisfy the stopping criterion (240.19) with a minimal number of elements. More precisely, we follow the following *adaptive algorithm*:

1. Choose an initial triangulation $\mathcal{T}_h^{(0)}$.
2. Given the j^{th} triangulation $\mathcal{T}_{h^{(j)}}$ with mesh function $h^{(j)}$, compute the corresponding finite element approximation $U^{(j)}$.
3. Compute the corresponding residuals $R_1(U^{(j)})$ and $R_2(U^{(j)})$ and check whether or not (240.19) holds. If it does, stop.
4. Find a new triangulation $\mathcal{T}_{h^{(j+1)}}$ with mesh function $h^{(j+1)}$ and with a minimal number of nodes such that $C_i\|h^{(j+1)}R(U^{(j)})\| \leq \text{TOL}$, and then proceed to #2.

The success of this iteration hinges on the mesh modification strategy used to perform step #4. A natural strategy for error control based on the L_2 norm uses the *principle of equidistribution* of the error in which we try to equalize the contribution from each element to the integral defining the L_2 norm. The rationale is that refining an element with large contribution to the error norm gives a large pay-off in terms of error reduction per new degree of freedom.

In other words, the approximation computed on the optimal mesh $\mathcal{T}_h$ in terms of computational work satisfies

$$\|\nabla e\|_{L_2(K)}^2 \approx \frac{\text{TOL}^2}{M} \quad \text{for all } K \in \mathcal{T}_h,$$

where M is the number of elements in $\mathcal{T}_h$. Based on (240.17), we would therefore like to compute the triangulation at step #4 so that

$$C_i^2\big(\big\|h^{(j+1)}R\big(U^{(j+1)}\big)\big\|_{L_2(K)}^2 \approx \frac{\text{TOL}^2}{M^{(j+1)}} \quad \text{for all } K \in \mathcal{T}_{h^{(j+1)}}, \tag{234.21}$$

where $M^{(j+1)}$ is the number of elements in $\mathcal{T}_{h^{(j+1)}}$. However, (240.20) is a nonlinear equation, since we don't know $M^{(j+1)}$ and $U^{(j+1)}$ until we have chosen the triangulation. Hence, we replace (240.20) by

$$C_i^2\big(\big\|h^{(j+1)}R\big(U^{(j)}\big)\big\|_{L_2(K)}^2 \approx \frac{\text{TOL}^2}{M^{(j)}} \quad \text{for all } K \in \mathcal{T}_h{}^{(j+1)}, \tag{234.22}$$

and use this formula to compute the new mesh size $h^{(j+1)}$.

There are several questions we may ask about the process described here: How much efficiency is lost by replacing (240.18) by (240.19)? Does the iterative process #1–#4 converge? Is the approximation (240.21) justified? We address such issues in the advanced companion volumes.

234.11 An Example

We want to compute the the solution

$$u(x) = \frac{a}{\pi}\,\exp\bigl(-a(x_1^2 + x_2^2)\bigr), \quad a = 400,$$

of Poisson's equation $-\Delta u = f$ on the square $(-.5, .5) \times (-.5, .5)$ with $f(x)$ being the following "approximate delta function":

$$f(x) = \frac{4}{\pi}\,a^2\,\bigl(1 - ax_1^2 - ax_2^2\bigr)\,\exp\bigl(-a(x_1^2 + x_2^2)\bigr),$$

We plot f in Fig. 240.10 (note the vertical scale), together with the initial mesh with 224 elements. The adaptive algorithm took 5 steps to achieve

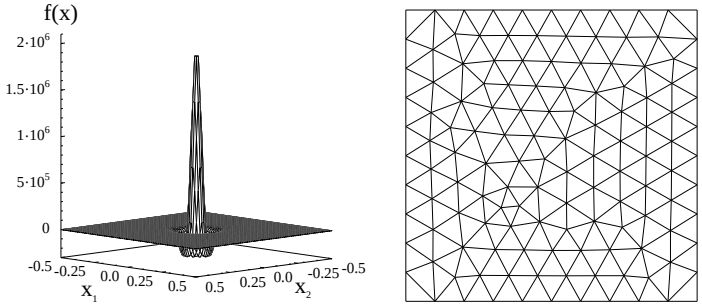


FIGURE 234.9. The approximate delta forcing function f and the initial mesh used for the finite element approximation.

an estimated .5% relative error. We plot the final mesh together with the associated finite element approximation in Fig. 240.11. The algorithm produced meshes with 224, 256, 336, 564, 992, and 3000 elements respectively.

234.12 Non-Homogeneous Dirichlet Boundary Conditions

We now consider Poisson's equation with non-homogeneous Dirichlet boundary conditions:

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = g & \text{on } \Gamma, \end{cases} \tag{234.23}$$

where g is the given boundary data.

We compute a finite element approximation on a triangulation $\mathcal{T}_h$, where we now also include the nodes on the boundary, denoting the internal nodes

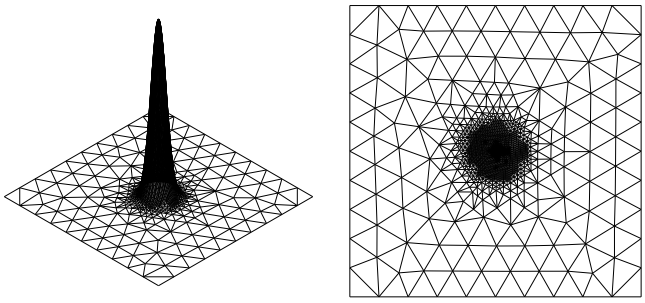


FIGURE 234.10. The finite element approximation with a relative error of .5% and the final mesh used to compute the approximation. The approximation has a maximum height of roughly 5.

by $\mathcal{N}_h$ as above and the set of nodes on the boundary by $\mathcal{N}_b$. We compute an approximation U of the form

$$U = \sum_{N_j \in \mathcal{N}_b} \xi_j \varphi_j + \sum_{N_j \in \mathcal{N}_h} \xi_j \varphi_j, \tag{234.24}$$

where φ_j denotes the basis function corresponding to node N_j in an enumeration $\{N_j\}$ of all the nodes, and, because of the boundary conditions, $\xi_j = g(N_j)$ for $N_j \in \mathcal{N}_b$. Thus the boundary values of U are given by g on Γ and only the coefficients of U corresponding to the interior nodes remain to be found. To this end, we substitute (240.24) into the variational formulation of (240.22) with the test functions being all the basis functions for the internal nodes, and we then get the following a square system of linear equations for the unknown coefficients of U:

$$\sum_{N_j \in \mathcal{N}_h} \xi_j (\nabla\varphi_j, \nabla\varphi_i) = (f, \varphi_i) - \sum_{N_j \in \mathcal{N}_b} g(N_j)(\nabla\varphi_j, \nabla\varphi_i), \quad N_i \in \mathcal{N}_h.$$

where the terms with known boundary values of U are shifted to the right hand side as data.

234.13 An L-shaped Membrane

We present an example that shows the performance of the adaptive algorithm on a problem with a *boundary singularity* with the derivatives of the exact solution being infinite at a corner of the boundary. We consider the Laplace equation in an L-shaped domain that has a non-convex corner at the origin satisfying homogeneous Dirichlet boundary conditions at the sides meeting at the origin and non-homogeneous conditions on the other sides, see Fig. 240.13. We choose the boundary conditions so that the exact

solution is $u(r,\theta) = r^{2/3}\sin(2\theta/3)$ in polar coordinates (r,θ) centered at the origin, which has a typical singularity of a corner problem:

$$\frac{\partial u}{\partial r}(r,\theta) = \frac{2}{3}r^{-1/3}\sin(2\theta/3),$$

which tends to infinity as r tends to zero (unless $\theta = 0$ or $\theta = \frac{3\pi}{2}$).

We use the knowledge of the exact solution to evaluate the performance of the adaptive algorithm.

We compute using an adaptive FEM-solver with energy norm control based on (240.17) to achieve an error tolerance of TOL$=$ .005 using h refinement mesh modification. In Fig. 240.13, we show the initial mesh $\mathcal{T}_{h^{(0)}}$ with 112 nodes and 182 elements. In Fig. 240.14, we show the level

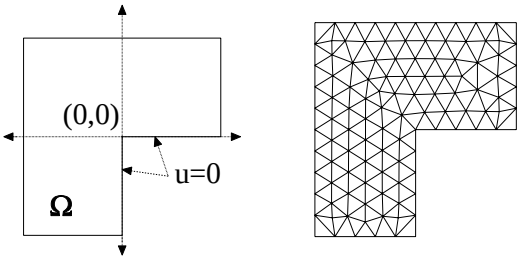


FIGURE 234.11. The L-shaped domain and the initial mesh.

curves of the solution and the final mesh with 295 nodes and 538 elements that achieves the desired error bound. The interpolation constant was set to $C_i = 1/8$. The quotient between the estimated and true error on the final mesh was 1.5.

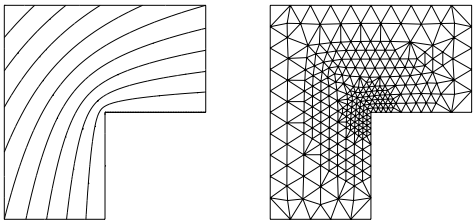


FIGURE 234.12. Level curves of the solution and final adapted mesh on the L-shaped domain.

Since the exact solution is known in this example, we can also use the a priori error estimate to determine a mesh that gives the desired accuracy.

We do this by combining the a priori error estimate (240.15) and the principle of equidistribution of error to determine $h(r)$ so that $C_i\|hD^2u\| = \text{TOL}$ while keeping h as large as possible (and keeping the number of elements at a minimum). Since $D^2u(r) \approx r^{-4/3}$, as long as $h \leq r$, that is up to the elements touching the corner, we determine that

$$\left(hr^{-4/3}\right)^2 h^2 \approx \frac{\text{TOL}^2}{M} \qquad \text{or} \quad h^2 = \text{TOL}\, M^{-1/2} r^{4/3},$$

where M is the number of elements and h^2 measures the element area. To compute M from this relation, we note that $M \approx \int_\Omega h^{-2}\, dx$, since the number of elements per unit area is $O(h^{-2})$, which gives

$$M \approx M^{1/2}\text{TOL}^{-1} \int_\Omega r^{-4/3}\, dx.$$

Since the integral is convergent (prove this), it follows that $M \propto \text{TOL}^{-2}$, which implies that $h(r) \propto r^{1/3}\,\text{TOL}$. Note that the total number of unknowns, up to a constant, is the same as that required for a smooth solution without a singularity, namely TOL^{-2}. This depends on the very local nature of the singularity in the present case. In general, of course solutions with singularities may require a much larger number of elements than smooth solutions do.

234.14 Robin and Neumann Boundary Conditions

Next, we consider Poisson's equation with homogeneous Dirichlet conditions on part Γ_1 of the boundary and non-homogeneous Robin conditions on the remaining part of the boundary Γ_2:

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_1, \\ \partial_n u + \kappa u = g & \text{on } \Gamma_2, \end{cases} \tag{234.25}$$

where $\kappa \geq 0$ is a given coefficient, and f and g are given data. Setting $\kappa = 0$ gives the Neumann condition $\partial_n u + \kappa u = g$. To find a variational formulation, we multiply the Poisson equation by a test function v satisfying the homogenous Dirichlet boundary condition, integrate over Ω, and use Green's formula to move derivatives from u to v:

$$\begin{aligned}(f, v) = -\int_\Omega \Delta u\, v\, dx &= \int_\Omega \nabla u \cdot \nabla v\, dx - \int_\Gamma \partial_n u v\, ds \\ &= \int_\Omega \nabla u \cdot \nabla v\, dx + \int_{\Gamma_2} \kappa u v\, ds - \int_{\Gamma_2} g v\, ds,\end{aligned}$$

where we use the boundary conditions to rewrite the boundary integral. We are thus led to the following cG(1) FEM based on a space V_h of continuous piecewise linear functions vanishing on Γ_1: find $U \in V_h$ such that

$$(\nabla U, \nabla v) + \int_{\Gamma_2} \kappa U v \, ds = (f, v) + \int_{\Gamma_2} g v \, ds \quad \text{for all } v \in V_h. \qquad (234.26)$$

We recall that boundary conditions like the Dirichlet condition that are enforced explicitly in the choice of the space V_h are called *essential boundary conditions*. Boundary conditions like the Robin condition that are implicitly contained in the weak formulation are called *natural boundary conditions*. (To remember that we must assume essential conditions: there are two "ss" in assume and essential.)

Note that the stiffness matrix and load vector related to (240.28) contain contributions from both integrals over Ω and Γ_2 related to the basis functions corresponding to the nodes on the boundary Γ_2.

To illustrate, we compute the solution of Laplace's equation with a combination of Dirichlet, Neumann and Robin boundary conditions on the domain shown in Fig. 240.15 using an adaptive FEM-solver. We show the boundary conditions in the illustration. The problem models e.g. stationary heat flow around a hot water pipe in the ground. We show the mesh

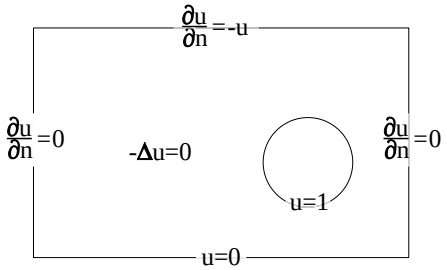


FIGURE 234.13. A problem with Robin boundary conditions.

used to compute the approximation so that the error in the L_2 norm is smaller than .0013 together with a contour plot of the approximation in Fig. 240.16. We notice that the level curves are parallel to a boundary with a homogeneous Dirichlet condition, and orthogonal to a boundary with a homogeneous Neumann condition.

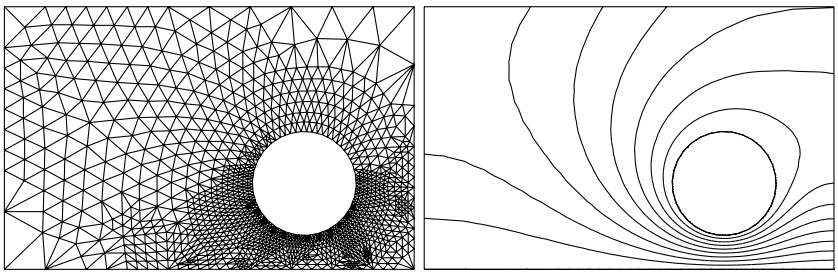


FIGURE 234.14. The adaptive mesh and contour lines of the approximate solution of the problem shown in Fig. 240.15 computed with error tolerance .0013.

234.15 Stationary Convection-Diffusion-Reaction

We now consider the extension to a convection-diffusion-reaction problem of the form

$$\begin{aligned} -\nabla\cdot(a\nabla u)+\nabla\cdot(ub)+cu &= f \quad \text{in } \Omega,\\ a\partial_n u+\kappa u &= g \quad \text{on } \Gamma. \end{aligned} \tag{234.27}$$

with Robin boundary conditions, where f and g are given data, and $a>0$, b, c and $\kappa\geq 0$ are given coefficients, and Ω is a given domain in $\mathbb{R}^2$ with boundary Γ. The term cu models absorption if $c\geq 0$ and production $c<0$.

Let V_h be the space of continuous piecewise linear functions on a triangulation of Ω with no restriction on the nodal values on the boundary. The cG(1) FEM for (234.27) takes the form: Find $U\in V_h$ such that

$$\begin{aligned} \int_\Omega a\nabla U\cdot\nabla v\,dx+\int_\Omega \nabla\cdot(Ub)v\,dx+\int_\Omega cUv\,dx+\int_\Gamma \kappa Uv\,ds&\\ =\int_\Omega fv\,dx+\int_\Gamma gv\,ds,& \end{aligned} \tag{234.28}$$

for all $v\in V_h$. Note that the extension to include the terms $\nabla\cdot(Ub)$ and cu is very natural and that the corresponding terms in the variational formulation are obtained by multiplying by the test function v without any partial integration. For the term $-\nabla\cdot(a\nabla u)$ we note that multiplication by $v(x)$ and integration over Ω gives using the Divergence theorem

$$-\int_\Omega \nabla\cdot(a\nabla u)v\,dx=\int_\Omega a\nabla U\cdot\nabla v\,dx-\int_\Gamma a\partial_n u\,v\,ds$$

and the variational formulation results from replacing $-a\partial_n u\,ds$ by $ku-g$ using the Robin boundary condition.

The matrix equation corresponding to (234.28) has a banded and sparse stiffness matrix, but the symmetry is lost if $b\neq 0$, as is evident from the presence of the non-symmetric term $\int_\Omega \nabla\cdot(ub)v\,dx$. The non-symmetry of

the convection term eliminates the best approximation property of FEM, but FEM still may give good results. If $c < 0$ then solutions may be non-unique corresponding to non-zero solutions (eigen-functions) of the homogeneous problem $-\nabla \cdot (a\nabla u) + \nabla \cdot (ub) + cu = 0$.

The Convection-Dominated Case: Streamline Diffusion

If $|b| > \frac{a}{h}$, where $h(x)$ is the mesh size, which we refer to as a *convection-dominated case*, and the exact solution is non-smooth with rapid variation, then the FEM-solution may exhibit spurious oscillations. In such cases the cG(1)-method (234.28) will have to be modified by changing the test functions from v to $v + \delta\nabla \cdot (vb)$ in all terms but the diffusion and boundary terms, where $\delta = \frac{h}{2|b|}$ acts as a parameter. The presence of the modification $\delta\nabla \cdot (vb)$ introduces the positive quadratic term $\int_\Omega \delta(\nabla \cdot (Ub))^2\,dx$ upon choosing $v = U$, which gives enhanced stability and (almost) eliminates spurious oscillations. The fact that the modification is not made in the diffusion term does not destroy accuracy, because in the convection dominated case the diffusion coefficient is small. The modified method is referred to as *the streamline diffusion method* or *weighted least squares-stabilization*.

234.16 Time-Dependent Convection-Diffusion-Reaction

We now consider the time-dependent analog of (234.27), that is the problem

$$\begin{aligned} \dot{u} - \nabla \cdot (a\nabla u) + \nabla \cdot (ub) + cu &= f \quad \text{in } \Omega \times (0,T], \\ a\partial_n u + \kappa u &= g \quad \text{on } \Gamma \times (0,T], \\ u(\cdot) &= u^0 \quad \text{in } \Omega, \end{aligned} \tag{234.29}$$

where $[0,T]$ is a given time interval, and u^0 a given initial value. For the time discretization we may use e.g. dG(0) or cG(1) on a subdivision $0 = t_0 < t_1 < \cdots < t_N = T$ into time intervals $I_n = (t_{n-1}, t_n]$ with time steps $k_n = t_n - t_{n-1}$. Using dG(0) we seek $U^n \in V_h$ for $n = 1, ..., N$, such that for $n = 1, ..., N$,

$$\begin{aligned} &\int_\Omega U^n v\,dx + \int_{\Omega\times I_n} a\nabla U^n \cdot \nabla v\,dx\,dt \\ &+ \int_{\Omega\times I_n} \nabla \cdot (U^n b)v\,dx\,dt + \int_{\Omega\times I_n} cU^n v\,dx\,dt + \int_{\Gamma\times I_n} \kappa U^n v\,ds\,dt \\ &= \int_\Omega U^{n-1} v\,dx + \int_{\Omega\times I_n} fv\,dx\,dt + \int_{\Gamma\times I_n} gv\,ds\,dt, \end{aligned} \tag{234.30}$$

for all $v \in V_h$, where $U^0 = u^0$. The corresponding discrete system for U^n takes the form

$$M\xi^n + k_n A_n \xi^n = M\xi^{n-1} + k_n b^n$$

where the vector ξ^n contains the nodal values of $U^n \in V_h$, M is the mass matrix related to V_h, A_n is the relevant stiffness matrix connected to the convection-diffusion-reaction terms, and b^n the relevant load vector.

In a convection-dominated case, the test functions v are again modified to $v + \delta\nabla\cdot(vb)$ in all terms with integration over $\Omega \times I_n$, but the diffusion term.

234.17 The Wave Equation

We now consider the extension to the wave equation with homogeneous Dirichlet boundary conditions:

$$\begin{aligned} \ddot{u} - \Delta u &= f \quad \text{in } \Omega \times (0,T], \\ u &= 0 \quad \text{on } \Gamma \times (0,T], \\ u &= u^0, \quad \dot{u} = \dot{u}^0 \quad \text{in } \Omega, \end{aligned} \tag{234.31}$$

where u^0 and $\dot{u}^0$ are given initial conditions. As above we let V_h be the set of piecewise linears functions on a triangulation of Ω satisfying the homogeneous Dirichlet boundary conditions, and we let $0 = t_0 < t_1 < \cdots < t_N = T$ be a subdivision of $[0,T]$ into time intervals $I_n = (t_{n-1}, t_n]$ with time steps $k_n = t_n - t_{n-1}$. We apply cG(1) in space and cG(1) in time and seek a discrete solution U in the space of functions W_h spanned by the functions

$$v(x,t) = \sum_{n=0}^{N}\sum_{j=1}^{M} \eta_j^n \varphi_j(x)\psi_n(t),$$

where $\{\varphi_j(x)\}_{j=1}^M$ is a basis for V_h, and $\{\psi_n(t)\}_{n=0}^N$ is a basis for the space of continuous piecewise linear functions on the subdivision $0 = t_0 < t_1 < \cdots < t_N = T$. The corresponding discrete system takes the following explicit form if mass lumping is used in space as well as time and the time step is constant $k_n = k$:

$$\xi^{n+1} = 2\xi^n - \xi^{n-1} + k^2 A\xi^n \quad \text{for } n = 1, ..., N-1,$$

with appropriate starting values ξ^0 and ξ^1 computed from the initial conditions, and A the relevant stiffness matrix related to the Laplacian.

234.18 Examples

We present some examples of systems of nonlinear reaction-diffusion-convection equations (234.27) arising in physics, chemistry and biology of

the form

$$\dot{u}_i - \nabla \cdot (a_i \nabla u_i) + \nabla \cdot (u_i b_i) + c_i u_i = f_i(u_1, ..., u_d) \quad \text{in } \Omega \times I,\ i = 1, ..., d, \tag{234.32}$$

where the $a_i > 0$, b_i and c_i are given coefficients, and the $f_i : \mathbb{R}^d \to \mathbb{R}$. These systems may be solved numerically by a direct extension of the cG(1) method in space and time presented above. We will return in detail to this issue below. In all the examples, a is a positive constant.

EXAMPLE 234.1. *The bistable equation for ferro-magnetism*

$$\dot{u} - a\Delta u = u - u^3. \tag{234.33}$$

EXAMPLE 234.2. *Superconductivity of fluids*

$$\begin{aligned} \dot{u}_1 - a\Delta u_1 &= (1 - |u|^2)u_1, \\ \dot{u}_2 - a\Delta u_2 &= (1 - |u|^2)u_2. \end{aligned} \tag{234.34}$$

EXAMPLE 234.3. *Flame propagation*

$$\begin{aligned} \dot{u}_1 - a\Delta u_1 &= -u_1 e^{-\alpha_1/u_2}, \\ \dot{u}_2 - a\Delta u_2 &= \alpha_2 u_1 e^{-\alpha_1/u_2}, \end{aligned} \tag{234.35}$$

where $\alpha_1, \alpha_2 > 0$ are constants.

EXAMPLE 234.4. *Interaction of two species*

$$\begin{aligned} \dot{u}_1 - a\Delta u_1 &= u_1 M(u_1, u_2), \\ \dot{u}_2 - a\Delta u_2 &= u_2 N(u_1, u_2), \end{aligned} \tag{234.36}$$

where $M(u_1, u_2)$ and $N(u_1, u_2)$ are given functions describing various situations such as (i) predator-prey ($M_{u_2} < 0$, $N_{u_1} > 0$) (ii) competing species ($M_{u_2} < 0$, $N_{u_1} < 0$) and (iii) symbiosis ($M_{u_2} > 0$, $N_{u_1} > 0$).

EXAMPLE 234.5. *Morphogenesis of patterns (zebra)*

$$\begin{aligned} \dot{u}_1 - a\Delta u_1 &= -u_1 u_2^2 + \alpha_1(1 - u_1), \\ \dot{u}_2 - a\Delta u_2 &= u_1 u_2^2 - (\alpha_1 + \alpha_2)u_2. \end{aligned} \tag{234.37}$$

EXAMPLE 234.6. *Belousov-Zhabotinski reaction in chemical kinetics*

$$\begin{aligned} \dot{u}_1 - a\Delta u_1 &= \alpha_1(u_2 - u_1 u_2 + u_1 - \alpha_2 u_2^2), \\ \dot{u}_2 - a\Delta u_2 &= \alpha_1^{-1}(\alpha_3 u_3 - u_2 - u_1 u_2). \\ \dot{u}_3 - a\Delta u_3 &= \alpha_4(u_1 - u_3), \end{aligned} \tag{234.38}$$

where $\alpha \approx 10^2$, $\alpha_2 \approx 10^{-2}$, $\alpha_3 \approx 1$, $\alpha_4 \approx 10^{-1}$.

Chapter 234 Problems

234.1. Compute the coefficients of the mass matrix M on the standard triangulation of the square of mesh size h. Hint: it is possible to use quadrature based on the midpoints of the sides of the triangle because this is exact for quadratic functions. The diagonal terms are $h^2/2$ and the off-diagonal terms are all equal to $h^2/12$. The sum of the elements in a row is equal to h^2.

234.2. Compute the stiffness matrix for cG(1) for the problem $-\Delta u = 1$ in $\Omega = (0,1) \times (0,1)$ with $u = 0$ on the side with $x_2 = 0$ and $\partial_n u + u = 1$ on the other three sides of Ω using the standard triangulation. Note the contribution to the stiffness matrix from the nodes on the boundary.

234.3. Describe the sparsity pattern of the stiffness matrices A for the Poisson equation with homogeneous Dirichlet data on the unit square corresponding to the continuous piecewise linear finite element method on the standard triangulation using the three numbering schemes pictured in Fig. 240.9.

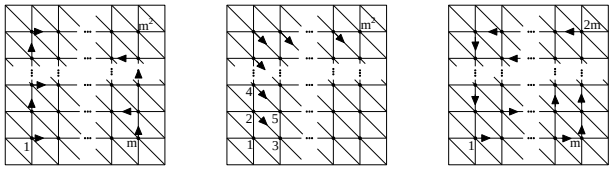


FIGURE 234.15. Three node numbering schemes for the standard triangulation of the unit square.

234.4. Compute the load vector for $f(x) = x_1 + x_2^2$ on the standard triangulation of the unit square using exact integration and the lumped mass (trapezoidal rule) quadrature.

234.5. Write a code to solve $A\xi = b$ using both the Jacobi and Gauss-Seidel iteration methods, making use of the sparsity of A in storage and operations. Compare the convergence rate of the two methods using the result from a direct solver as a reference value.

234.6. Write a code to solve the system $A\xi = b$ with A a band matrix.

234.7. Compute the stiffness matrix for the Poisson equation with homogeneous Dirichlet boundary conditions for (a) the *union jack* triangulation of a square shown in Fig. 240.8 and (b) the triangulation of triangular domain shown in Fig. 240.8.

234.8. Compute the discrete equations for the finite element approximation for $-\Delta u = 1$ on $\Omega = (0,1) \times (0,1)$ with boundary conditions $u = 0$ for $x_1 = 0$, $u = x_1$ for $x_2 = 0$, $u = 1$ for $x_1 = 1$ and $u = x_1$ for $x_2 = 1$ using the standard triangulation (Fig. 240.1).

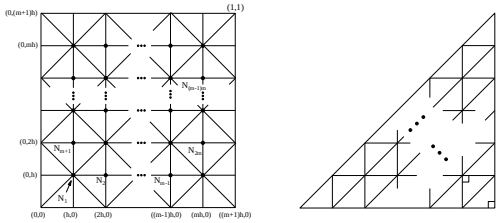


FIGURE 234.16. The "union jack" triangulation of the unit square and a uniform triangulation of a right triangle.

234.9. (a) Show that the element stiffness matrix (240.13) for the linear polynomials on a triangle K with vertices at $(0,0)$, $(h,0)$, and $(0,h)$ numbered 1, 2 and 3, is given by

$$\begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1/2 & 0 \\ -1/2 & 0 & 1/2 \end{pmatrix}.$$

(b) Use this result to verify the formula computed for the stiffness matrix A for the continuous piecewise linear finite element method for the Poisson equation with homogeneous boundary conditions on the unit square using the standard triangulation. (c) Compute the element stiffness matrix for a triangle K with nodes $\{a^i\}$.

234.10. Compute the asymptotic operations count for the direct solution of the system $A\xi = b$ using the three A computed in Problem 240.15.

234.11. Apply the finite element method with piecewise linear approximation to the Poisson equation in three dimensions with a variety of boundary conditions. Compute the stiffness matrix and load vector in some simple cases.

234.12. Derive a priori error bound in the energy norm for cG(1) FEM for Poisson's equation with Robin boundary conditions. Generalize to problems of the form $-\nabla \cdot (a\nabla u) + cu = f$, where $a(x) > 0$ and $c \geq 0$.

234.13. Derive a posteriori error bound in the energy norm for cG(1) FEM for Poisson's equation with Robin boundary conditions. Generalize to problems of the form $-\nabla \cdot (a\nabla u) + cu = f$, where $a(x) > 0$ and $c \geq 0$.

234.14. Implement adaptive energy norm error control for cG(1) for Poisson's equation based on an a posteriori error estimate.

234.15. Find an exact solution of the L-shaped membrane problem with the Dirichlet condition replaced by a Neumann condition on one of the sides meeting at $\frac{3\pi}{2}$ corner. What is the nature of the singularity?

234.16. Let $\omega(x)$ be a positive weight function defined on the domain $\Omega \subset \mathbb{R}^2$. Assume that the mesh function $h(x)$ minimizes the integral $\int_\Omega h^2(x)\omega(x)\,dx$ under the constraint $\int_\Omega h^{-2}(x)\,dx = N$, where N is a given positive integer. Prove that $h^4(x)\omega(x)$ must be constant. Interpret the result as equidistribution in the context of error control. Hint: argue that $h^4(x)\omega(x)$ is the gain adding one more node.

235
Inverse Problems

I never guess. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.

When you have eliminated the impossible, whatever remains, however improbable, must be the truth. (Sherlock Holmes in the *The Sign of Four*, 1888)

235.1 Introduction

We have above in our study of Poisson's equation studied "forward" problems of the form: Given the function $f : \Omega \to \mathbb{R}$, find a function $u : \Omega \to \mathbb{R}$ such that

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ \partial_n u + \kappa u = 0 & \text{on } \Gamma, \end{cases} \tag{235.1}$$

where $\kappa \geq 0$ is a given constant. A corresponding "inverse" problem would be to assume knowledge of $u(x)$ and seek the corresponding function $f(x)$ so that (235.1) is satisfied! If we know $u(x)$ in the entire region Ω, this is a problem of differentiation: we just compute $\Delta u(x)$ from $u(x)$. We then have $-\Delta u = f$! We have studied this problem in Chapter *The derivative*, and we recall that this problem is a bit delicate and that we have to balance the step length h in a difference approximation of Δ to the precision in the given data $u(x)$.

Suppose now that we know $u(x)$ only on the boundary Γ. Can we then determine $f(x)$ in Ω? This type of problem connects to a wealth of important applications of the following form: Suppose we can measure something on the *boundary* of an object. Can we then say something about what is *inside* the object? For example, suppose the object is a human body, and that we can measure something on the boundary (or outside) the body. Can we than get some information on what is inside the body? Or, suppose we can accumulate data on the surface of the Earth, can we then say something about what is in the interior of the Earth, such as the presence of layers of oil. These are all examples of *inverse problems*.

The nature of an inverse problem is to be "ill-posed" in the sense that solutions may be non-unique and/or that small changes in the data may cause large changes in the solution. To single out a unique solution which is not too sensitive to little errors in data, we may have to "regularize" the inverse problem e.g. by smoothing of the data and/or restricting the size of (derivatives of) the solution. Differentiation is such an ill-posed problem where we may need to "smooth" or regularize a given function before attempting to compute its derivative.

A typical forward problem is "well-posed" in the sense that small changes in data cause small changes in the solution. A basic example of a well-posed problem is integration corresponding to solving a differential equation. The corresponding inverse problem is differentiation which is ill-posed as we just noted. Solving a differential equation does not always correspond to a well-posed problem: in Chapter *Lorenz and the Essence of Chaos* we met a simple differential equation with solutions being highly sensitive to changes in data.

EXAMPLE 235.1. An *electrocardiogram* ECG produces a curve reflecting the electrical activity of the heart from measurements of electric potentials on the chest, and the curve gives a specialist information on abnormal activities of the heart such as abnormal heart rhythm (arrhythmias). Similarly, an *electroencephalogram* EEG gives information on the electrical activity of the brain from measurements of electric potentials on the scalp. These techniques are however too imprecise for many diagnoses, and more recently techniques of *electrocardiographic imaging* have been developed, which build on solving inverse problems for Poisson-like equations. The geometry of the individual patient is then obtained from computer tomography, and a picture of the electrical activity inside the body is obtained from measurements of electric potentials on the boundary (e.g. the chest or scalp) by solving an inverse problem for a Poisson-like equation (using the finite element method). Electrocardiographic imaging may give more accurate information on e.g. abnormal cardiac or brain activity than ECG and EEG, and is now a part of practice in advanced neurological and radiological depart-

ments. Further development of in particular the computational process (adaptivity, geometric modeling) is needed to increase the accuracy.

EXAMPLE 235.2. Another inverse problem of importance to mankind occurs in inverse seismic prospecting: Explosions on the surface of the Earth are set off and the reflections of the induced waves in the mantle of the Earth are recorded on the surface, and from this information one tries to determine subsurface structures such as layers of oil-bearing rock. To solve this reconstruction problem one uses computational methods based on solving the wave equation involving a wave speed coefficient characteristic of different materials, and through optimization one tries to find the local wave speed coefficient which gives best least squares fit to measured data on the surface of the Earth, and which then gives information on the unknown subsurface layering.

235.2 An Inverse Problem for One-Dimensional Convection

We start considering the simplest boundary value problem:

$$u'(x) = f(x), \text{ for } x \in (0,1],\ u(0) = 0, \tag{235.2}$$

modeling convection with $u : [0,1] \to \mathbb{R}$ representing a concentration and $f : [0,1] \to \mathbb{R}$ a source. We seek to determine or *reconstruct* the function $f : [0,1] \to \mathbb{R}$ from the boundary value *observation* $u(1)$ of the corresponding solution $u(x)$ of (235.2). It is clear that we cannot hope to determine $f(x)$ for all $x \in (0,1)$ from this observation alone. This is because there are many functions $f(x)$ such that the corresponding function $u(x)$ satisfies (235.2) and $u(1) = 0$. To see this it is sufficient to choose a non-zero function $u(x)$ on $[0,1]$ satisfying $u(0) = u(1) = 0$ and define $f(x) = u'(x)$. Evidently, the reconstruction is undetermined up to such functions.

The indeterminancy of the reconstruction $f(x)$ reflects the *ill-posed* nature of the inverse problem; even if the measurements of the boundary value $u(1)$ is very precise, the corresponding source $f(x)$ is not well defined. We thus need some *extra condition* to single out a (hopefully) unique source $f(x)$. We may do this in many ways and depending on the extra condition imposed, we may get different reconstructions $f(x)$. We now indicate one possibility, where we reconstruct under the extra condition that $f(x)$ is *as small as possible* in a least squares sense, which is a common technique of regularization. We then reformulate the inverse problem as the following *least squares optimization problem*: Find the function $f : [0,1] \to \mathbb{R}$ which minimizes the total "cost"

$$J(f) = (u(1) - \bar{u}(1))^2 + \mu \int_0^1 f(x)^2\, dx,$$

where $u(x)$ solves (235.2), $\bar{u}(1)$ is the observed boundary value at $x = 1$, and $\mu > 0$ is a constant. We may view this as a control problem where the objective is to find the *control* $f : [0, 1] \to \mathbb{R}$ which minimizes the total cost $J(f)$, where the μ-term measures the cost of the control f and the first term the cost of a boundary value misfit $u(1) - \bar{u}(1)$.

We can phrase this problem as finding the function $u(x)$ with $u(0) = 0$ which minimizes

$$(u(1) - \bar{u}(1))^2 + \mu \int_0^1 (u')^2 \, dx.$$

Recalling Chapter *FEM for two-point boundary value problems*, we understand that the solution $u(x)$ satisfies $u''(x) = 0$ in $(0, 1)$, $u(0) = 0$ and $u(1) - \bar{u}(1) + \mu u'(1) = 0$. We conclude that $u(x) = \frac{1}{1+\mu}\bar{u}(1)x$, and thus the reconstructed source $f(x) = u'(x)$ takes on a constant value and is given by

$$f(x) = \frac{1}{1+\mu}\,\bar{u}(1) \quad \text{for } x \in [0, 1].$$

Evidently, we are led to choose the regularization parameter μ small; the smaller μ is the more accurately we will fit the boundary value observation $\bar{u}(1)$. Since the reconstructed function $f(x)$ is constant, we have effectively only one constant to determine and we may expect to be able to determine this single value from the single observation $\bar{u}(1)$.

We can also rephrase the optimization problem as follows introducing the integral operator B defined on functions on $[0, 1]$ by $Bf(x) = \int_0^x f(y)\, dy$ for $x \in [0, 1]$: Find the function $f : [0, 1] \to \mathbb{R}$ which minimizes

$$J(f) = (Bf(1) - \bar{u}(1))^2 + \mu \int_0^1 f(x)^2 \, dx.$$

The optimality condition obtained by setting $\frac{d}{d\epsilon}J(f + \epsilon g) = 0$ for $\epsilon = 0$, where $g : [0, 1] \to \mathbb{R}$ is an arbitrary function, takes the form:

$$(Bf(1) - \bar{u}(1))Bg(1) + \mu \int_0^1 f(x)g(x)\, dx = 0 \tag{235.3}$$

for all functions $g : [0, 1] \to \mathbb{R}$. We shall now rewrite this condition by introducing the *adjoint operator* $B^\top$ defined on functions $w(x)$ as follows: for a given $w = w(x)$ we let $B^\top w$ be the function on $[0, 1]$ satisfying $(B^\top w)'(x) = 0$ for $x \in (0, 1)$ and $B^\top w(1) = w(1)$, that is, $B^\top w$ is the constant function on $[0, 1]$ taking the value $w(1)$ for all x. We can then rewrite (235.3) in the form

$$\int_0^1 B^\top (Bf(x) - \bar{u}(1))(x)g(x)\, dx + \mu \int_0^1 f(x)g(x)\, dx = 0 \tag{235.4}$$

for all functions $g : [0,1] \to \mathbb{R}$, because by partial integration

$$\int_0^1 B^\top (Bf(x) - \bar{u}(1))(x) \underbrace{(Bg(x))'}_{=g(x)} dx = (Bf(1) - \bar{u}(1))Bg(1),$$

where as indicated $(Bg(x))' = g(x)$ and $B^\top w(1) = w(1)$. We conclude that

$$\int_0^1 (B^\top Bf + \mu f)g\, dx = \int_0^1 B^\top \bar{u}(1)g\, dx$$

for all functions $g : [0,1] \to \mathbb{R}$, and therefore

$$B^\top Bf + \mu f = B^\top \bar{u}(1) \quad \text{on } (0,1), \tag{235.5}$$

or with I the identity operator:

$$(B^\top B + \mu I)f = B^\top \bar{u}(1). \tag{235.6}$$

We conclude that $f(x)$ is constant on $[0,1]$ and takes the value

$$f(x) = \frac{1}{1+\mu}\,\bar{u}(1) \quad \text{for } x \in [0,1],$$

which is the same result as already derived. We note the form (235.6) of the optimality condition (235.6) with the operator $B^\top B + \mu I$ appearing. We shall meet the same equation below with different solution operators B and adjoints $B^\top$.

235.3 An Inverse Problem for One-Dimensional Diffusion

We continue with the boundary value problem

$$-u'' = f \quad \text{in } (0,1), \quad u'(0) = 0, \quad u'(1) + u(1) = 0, \tag{235.7}$$

where we seek to determine the source $f(x)$ in $(0,1)$ by observing the boundary values $u(0)$ and $u(1)$ of the corresponding solution $u(x)$ of (235.7). Again it is clear that we cannot hope to determine $f(x)$ for all $x \in (0,1)$ from these two observations alone, because there are many functions $f(x)$ such that the corresponding function $u(x)$ satisfies (235.7) and $u(0) = u(1) = 0$. To see this it is sufficient to choose a non-zero function $u(x)$ on $[0,1]$ satisfying $u(0) = u'(0) = u(1) = u'(1) = 0$ and set $f(x) = -u''(x)$.

As above we seek to reconstruct $f(x)$ under the extra condition that $f(x)$ is as small as possible and we therefore reformulate the inverse problem as

the following least squares optimization problem: Find $f(x)$ in $(0,1)$ such that

$$J(f) = (u(0) - \bar{u}(0))^2 + (u(1) - \bar{u}(1))^2 + \mu \int_0^1 f^2(x)\,dx$$

is as small as possible, where $\mu > 0$ is a positive constant acting as a regularization, $\bar{u}(0)$ and $\bar{u}(1)$ are the boundary observations, and of course $u(x)$ solves (235.7). We thus seek $f(x)$ so that in a least squares sense we fit the boundary observations as well as the smallness of $f(x)$ as well as possible.

To state the optimality equations, we introduce the solution operator B corresponding to (235.7), that is, for a given function $f : [0,1] \to \mathbb{R}$ we let $Bf(x)$ be the function on $[0,1]$ satisfying

$$\int_0^1 (Bf)'v'\,dx + Bf(1)v(1) = \int_0^1 fv\,dx, \tag{235.8}$$

for all functions $v(x)$ on $[0,1]$. This follows from Chapter *FEM for two-point boundary value problems*. Setting $\frac{d}{d\epsilon}J(f+\epsilon g) = 0$ for $\epsilon = 0$, we obtain the optimality condition in the form

$$(Bf(0) - \bar{u}(0), Bg(0)) + (Bf(1) - \bar{u}(1), Bg(1)) + \mu \int_0^1 f(x)g(x)\,dx = 0 \tag{235.9}$$

for all functions $g(x)$ on $[0,1]$. Next we introduce the adjoint operator $B^\top$ defined as follows: given the values $w(0)$ and $w(1)$, we let $B^\top w$ be the function on $[0,1]$ which satisfies

$$\int_0^1 (B^\top w)'v'\,dx + B^\top w(1)v(1) = w(0)v(0) + w(1)v(1) \tag{235.10}$$

for all $v(x)$. We see that $(B^\top w)'' = 0$ and $-(B^\top w)'(0) = w(0)$, $B^\top w(1) + (B^\top w)'(1) = w(1)$. In other words, $B^\top w$ is a linear function determined by the two boundary conditions. In particular, if $w(0) = 0$, then $B^\top w = w(1)$ is a constant. Now, setting $v = Bg$ in (235.10), we get

$$\begin{aligned} w(0)Bg(0) + w(1)Bg(1) &= \int_0^1 (B^\top w)'(Bg)'\,dx + B^\top w(1)Bg(1) \\ &= \int_0^1 B^\top w g\,dx, \end{aligned}$$

where we used (235.8) with f replaced by g and v replaced by $B^\top w$, and thus we can write the optimality condition (235.9) in the same form as above:

$$(B^\top B + \mu I)f = B^\top \bar{u}. \tag{235.11}$$

From this equation we can uniquely solve for the function $f(x)$, which will be a linear function defined by two constants, because $f = \frac{1}{\mu}B^\top(Bf - \bar{u})$.

For example if $\bar{u}(0) = 0$ and $\bar{u}(1) = 1$, then we get choosing μ small, $f(x) \approx 10\bar{u}(1)x - 8\bar{u}(1)$ with corresponding solution $u(x) \approx -3x^3 + 4x^2$.

We now comment on the nature of the optimality equation (235.11). The operator B maps a space of sources, say F, into a space of observations, say O, and the adjoint operator $B^\top$ maps O into F. We may think of the dimension of F as large, and that of O as smaller. For the discussion we may assume that the dimension of F is n and the dimension of O is m and thus B corresponds to an $m \times n$ matrix and $B^\top$ to an $n \times m$ matrix with $m << n$. This will be the setting with computational approximations of the solution operators B and $B^\top$. In particular, the columns of B must be severely linearly independent since there are many more columns than rows, and thus the $n \times n$ matrix $B^\top B$ must be singular with many non-zero n-vectors f satisfying $B^\top Bf = 0$. On the other hand, the matrix $B^\top B + \mu I$ with $\mu > 0$ is nonsingular, because if $(B^\top B + \mu I)f = 0$, then scalar multiplication by the n-vector $f^\top$, we obtain $\|Bf\|^2 + \mu\|f\|^2 = 0$ and thus $f = 0$. The non-zero solutions f to $B^\top Bf = 0$ are eigenvectors corresponding to a zero eigenvalue, and by changing to the regularized operator $B^\top B + \mu I$ we shift the spectrum to the interval $[\mu, \infty)$ on the positive real axis.

235.4 An Inverse Problem for Poisson's Equation

We now pass to an inverse problem for Poisson's equation (235.1) assuming Ω, Γ and κ to be known: Given $u(x) = \hat{U}(x)$ for $x \in \Gamma$, find $f(x)$ for $x \in \Omega$.

We approach this problem directly in discrete form as the following least squares problem: Find $F \in V_h$ which minimizes

$$J(F) = \|U - \hat{U}\|_\Gamma^2 + \mu\|F\|_\Omega^2 \tag{235.12}$$

over V_h, where $U \in V_h$ satisfies

$$(\nabla U, \nabla v)_\Omega + (\kappa U, v)_\Gamma = (F, v)_\Omega \quad \text{for all } v \in V_h, \tag{235.13}$$

and V_h is the space of continuous piecewise linear functions on a given triangulation of Ω of mesh size $h(x)$. As above $\mu \geq 0$ acts as a *regularization parameter* which helps to cope with the ill-posed nature of the problem. Further, $\|\cdot\|_\Omega$ and $(\cdot,\cdot)_\Omega$ denote the $L_2(\Omega)$ norm and scalar product, and similarly, $\|\cdot\|_\Gamma$ and $(\cdot,\cdot)_\Gamma$ denote the $L_2(\Gamma)$ norm and scalar product.

We reformulate (235.12) by introducing the solution operator $B_h : V_h \to W_h$ defined by $B_h F = U_F$ on Γ, where $U_F \in V_h$ solves (235.13), and W_h is the restriction of the space V_h to the boundary Γ, that is a set of piecewise linear functions on Γ. By definition, $U_F \in V_h$ satisfies:

$$(\nabla U_F, \nabla v)_\Omega + (\kappa U_F, v)_\Gamma = (F, v)_\Omega \quad \text{for all } v \in V_h. \tag{235.14}$$

We can now formulate the minimization problem (235.12) as follows: Find $F \in V_h$ which minimizes

$$J(F) = \|B_h F - \hat{U}\|_\Gamma^2 + \mu\|F\|_\Omega^2, \tag{235.15}$$

over V_h. This is a quadratic minimization problem with unique solution $F \in V_h$ characterized by a least squares equation of the form

$$(B_h F, B_h G)_\Gamma + (\mu F, G)_\Omega = (\hat{U}, B_h G)_\Gamma \quad \text{for all } G \in V_h, \tag{235.16}$$

which expresses that $\frac{d}{d\epsilon}J(F+\epsilon G) = 0$ for $\epsilon = 0$ for all $G \in V_h$.

We can express (235.16) as

$$(B_h^\top B_h F, G)_\Omega + (\mu F, G)_\Omega = (B_h^\top \hat{U}, G)_\Omega \quad \text{for all } G \in V_h,$$

that is

$$(B_h^\top B_h + \mu I)F = B_h^\top \hat{U}, \tag{235.17}$$

where $B_h^\top : W_h \to V_h$ is the transpose of B_h defined as follows: Given $w \in W_h$, we let $B_h^\top w \in V_h$ satisfy

$$(\nabla v, \nabla B_h^\top w)_\Omega + (\kappa v,\, B_h^\top w)_\Gamma = (v,\, w)_\Gamma \quad \text{for all } v \in V_h. \tag{235.18}$$

In other words, $B_h^\top w$ is an approximation of the solution z of the Poisson-problem:

$$\begin{cases} -\Delta z = 0 & \text{in } \Omega, \\ \partial_n z + \kappa z = w & \text{on } \Gamma. \end{cases} \tag{235.19}$$

Choosing $v = B_h G$ in (235.18), we get using also (235.13) with $v = B_h^\top w$

$$(B_h G, w)_\Gamma = (\nabla B_h G, \nabla B_h^\top w)_\Omega + (\kappa B_h G,\, B_h^\top w)_\Gamma = (G, B_h^\top w)_\Omega$$

and thus as expected from a transpose

$$(B_h G, w)_\Gamma = (G, B_h^\top w)_\Omega,$$

that is, moving B_h from G onto w brings in the transpose $B_h^\top$.

Solving (235.17) gives an approximation $F(x)$ of the function $f(x)$ we are looking for. We may solve (235.17) by direct matrix inversion if the number of nodes is small, and by some iterative method such as the gradient or the conjugate gradient method for larger problems.

The gradient method takes the form:

$$F^{n+1} = F^n - \alpha((B_h^\top B_h + \mu I)F^n - B_h^\top \hat{U}) = F^n - \alpha(B_h^\top (B_h F^n - \hat{U}) + \mu F^n).$$

In each step we have to compute first $B_h F$ and then $B_h^\top (B_h F - \hat{U})$ corresponding to solving two Poisson problems.

EXAMPLE 235.3. In our first application realized using Matlab we interpret (235.17) as a matrix equation explicitly formed by computing the inverses of the stiffness matrices for the problem (235.14) and the adjoint (235.18), and we then solve this matrix equation to get the nodal values of $F(x)$. One may handle a couple of hundreds of nodes this way. For simplicity, we have considered the case $\Omega = \{(x_1, x_2) : 0 < x_1, x_2 < 1\}$ with $\kappa = 1$ and $f = 0.5 + (x - y)(x + y - 1)$, observed the boundary values of the resulting solution u, and then solved for a reconstruction of the given data f using $\mu = 0.0001$. The result is shown in Fig. 235.1 with reconstruction error ~ 0.032 in f and ~ 0.000176 in the corresponding state (boundary values).

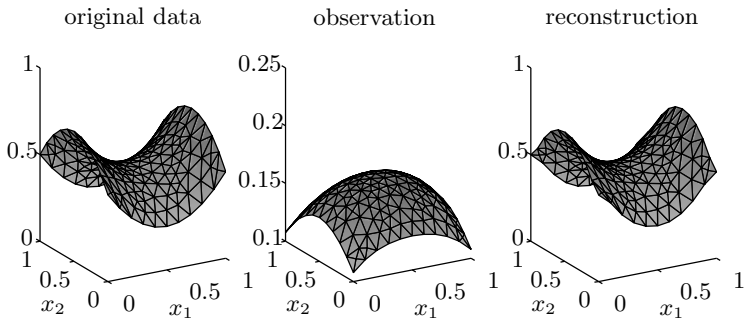


FIGURE 235.1. Original data f (left), resulting state u (middle), and the reconstruction of f (right) with $\mu = 0.0001$ and reconstruction error ~ 0.032.

EXAMPLE 235.4. We next take $\kappa = 50$ and show the resulting state u in Fig. 235.2. The reconstruction using $\mu = 0.0001$ is now rather poor, at least in terms of f with a reconstruction error of order ~ 0.4, while the corresponding state error is of order ~ 0.02. Taking $\mu = 0.00001$ brings the state error down to ~ 0.003, while a reconstruction of f with error ~ 0.04 requires taking $\mu = 0.0000005$.

235.5 An Inverse Problem for Laplace's Equation

Let Ω be a domain in $\mathbb{R}^2$ with boundary Γ composed of three parts Γ_0, Γ_1 and Γ_2. For a given function f defined on Γ_2, let u_f be the solution to the boundary value problem

$$\begin{cases} -\Delta u_f = 0 & \text{in } \Omega, \\ u_f = 0 & \text{on } \Gamma_0 \cup \Gamma_1, \quad u_f = f \quad \text{on } \Gamma_2, \end{cases} \tag{235.20}$$

and define $Bf = \frac{\partial u_f}{\partial n}$ on Γ_1, where n is the unit outward normal to Γ_1. We may think of u_f as a stationary temperature defined in Ω satisfying given

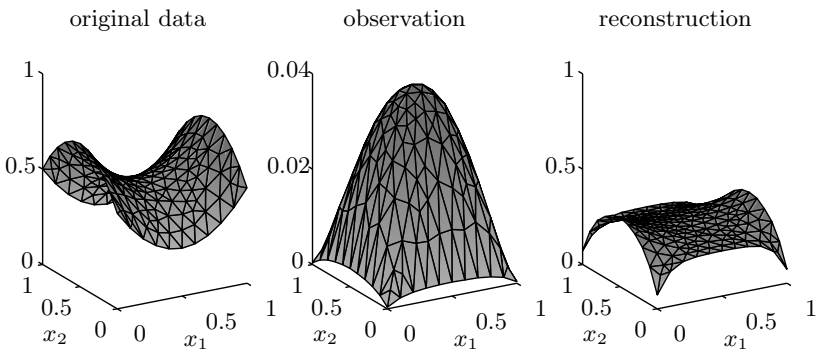


FIGURE 235.2. Original data f (left), resulting state u (middle), and the reconstruction of f (right) with $\mu = 0.0001$, now with reconstruction error ~ 0.43.

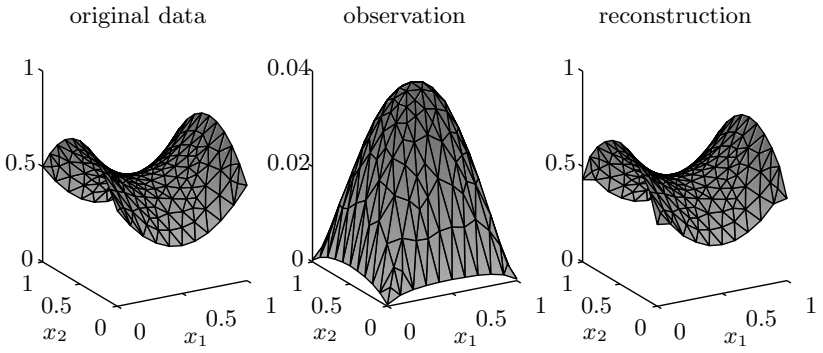


FIGURE 235.3. Original data f (left), resulting state u (middle), and the reconstruction of f (right) after 10 steps of the conjugate gradient method with $\mu = 0.000001$, with a state error ~ 0.0003 in (the boundary values of) u and a reconstruction error ~ 0.07 in f.

boundary conditions on Γ ($= 0$ on $\Gamma_0 \cup \Gamma_1$ and $= f$ on Γ_2) and with Bf representing the heat flux on Γ_1. Suppose now we can measure the heat flux on Γ_1 and that we want to determine the temperature f on Γ_2. We thus have a situation where we have access to the temperature ($= 0$) along $\Gamma_0 \cup \Gamma_1$ and may measure also the heat flux, say $\bar{q}$, along Γ_1, and we want to determine the temperature f on the inaccessible part of the boundary Γ_2. This problem arises in EKG with u being a potential and Γ_1 representing the surface of the chest and Γ_2 that of the heart, and the inverse problem being to reconstruct the potential on the heart from measurements on the chest.

We formulate the reconstruction problem as a least squares optimization problem of the form: Find f on Γ_2 which minimizes

$$J(f) = \|Bf - \bar{q}\|^2_{\Gamma_1} + \mu \|f\|^2_{\Gamma_2},$$

where we use the notation of the previous section, and $\mu > 0$. The optimality equation as usual takes the form

$$(B^\top B + \mu I) = B^\top \bar{q},$$

where $B^\top g = \frac{\partial u^g}{\partial n}$ on Γ_2 and u^g solves the problem

$$\begin{cases} -\Delta u^g = 0 & \text{in } \Omega, \\ u^g = g & \text{on } \Gamma_1, \quad u^g = 0 \quad \text{on } \Gamma_0 \cup \Gamma_2. \end{cases} \tag{235.21}$$

This because by integrations by parts

$$(g, Bf)_{\Gamma_1} = (\nabla u^g, \nabla u_f)_\Omega = (B^\top g, f)_{\Gamma_2}$$

EXAMPLE 235.5. We consider again the domain $\Omega = \{(x_1, x_2) : 0 < x_1, x_2 < 1\}$ now with $\Gamma_1 = \{(0, x_2) : 0 < x_2 < 1\}$, $\Gamma_2 = \{(1, x_2) : 0 < x_2 < 1\}$ and $\Gamma_0 = \Gamma \backslash \Gamma_1$ with an observed flow $\bar{q}$ along Γ_1 corresponding to $f = 6\, x_2^2\, (1 - x_2)$ along Γ_2. The figure shows the original (Dirichlet) boundary values to the left, the resulting state u and the associated observed flux q along Γ_1 in the middle, and the control/reconstruction f after a few conjugate gradient iterations to the right, with $\mu = 0.001$. The error in the (piecewise constant) reconstruction of the boundary values along Γ_2 is ~ 0.2 and the resulting error in flux through Γ_1 is ~ 0.006.

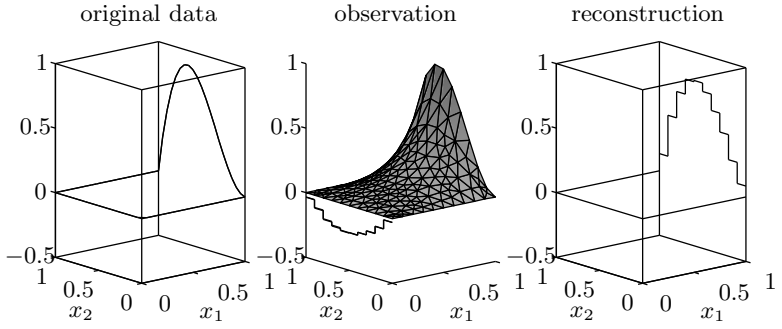


FIGURE 235.4. Original Dirichlet boundary values (left), the corresponding state u with observed flow q along Γ_1 (middle), and the reconstruction of the boundary values along Γ_2 (to the right) using $\mu = 0.001$ and a few (5) conjugate gradient steps.

235.6 The Backward Heat Equation

Another basic inverse problem is the *Backward heat equation*: Given the temperature at final time $t = T$, find the temperature at initial time $t = 0$.

We consider this problem in the following setting: let $f(x)$ be an initial temperature and let $u(x,t)$ be the corresponding solution of the heat equation:

$$\begin{cases} \dot{u} - \Delta u = 0 & \text{in } \Omega \times (0,T], \\ \partial_n u + \kappa u = 0 & \text{on } \Gamma \times (0,T], \\ u(x,0) = f(x) & \text{for } x \in \Omega, \end{cases}$$

where the domain $\Omega \in \mathbb{R}^d$ and the coefficient $\kappa \geq 0$ are given. We consider the following inverse problem: Given the final temperature $u(x,T)$, find the initial temperature $u(x,0) = f(x)$. This corresponds to solving the heat equation "backwards".

We consider the following discrete analog of (235.6) with discretization in space: Let V_h be the usual space of continuous piecewise linear functions on a triangulation of Ω with mesh size $h(x)$, and let $F \in V_h$ and let $U(t) \in V_h$ be the solution of the discrete heat equation

$$\begin{cases} (\dot{U}, v)_\Omega + (\nabla U(t), \nabla v)_\Omega = 0 & \text{for } t \in (0,T],\ v \in V_h, \\ (U(0), v)_\Omega = (F, v)_\Omega & \text{for } v \in V_h. \end{cases} \tag{235.22}$$

The discrete inverse problem, corresponding to a discrete "backwards" heat equation, reads: Given the final temperature $U(T) = \hat{U} \in V_h$, find the initial temperature $U(0) = F \in V_h$.

To formulate this problem as a regularized least squares problem, we introduce the solution operator $B_h : V_h \to V_h$ defined as follows: $B_h F = U(T) \in V_h$, where U solves with $U(0) = F \in V_h$. The operator B_h thus takes an initial temperature F to a corresponding final temperature $B_h F$. The regularized least squares problem is now the same as that above, that is, we seek $F \in V_h$ which minimizes

$$J(F) = \|B_h F - \hat{U}\|_\Omega^2 + \mu \|F\|_\Omega^2, \tag{235.23}$$

over V_h. The unique solution $F \in V_h$ to this quadratic minimization problem is characterized by a least squares equation of the form

$$(B_h F, B_h G)_\Omega + (\mu F, G)_\Omega = (\hat{U}, B_h G)_\Omega \quad \text{for all } G \in V_h, \tag{235.24}$$

which we can express as

$$(B_h^\top B_h F, G)_\Omega + (\mu F, G)_\Omega = (B_h^\top \hat{U}, G)_\Omega \quad \text{for all } G \in V_h,$$

that is

$$(B_h^\top B_h + \mu I) F = B_h^\top \hat{U}, \tag{235.25}$$

where $B_h^\top : V_h \to V_h$ is defined as follows: $B_h^\top G = Z(0) \in V_h$, where $Z(t) \in V_h$ solves the discrete heat equation

$$\begin{cases} -(v, \dot{Z})_\Omega + (\nabla v, \nabla Z)_\Omega = 0 & \text{for } t \in (0,T],\ v \in V_h, \\ (v, Z(T))_\Omega = (G, v)_\Omega & \text{for } v \in V_h. \end{cases} \tag{235.26}$$

Note computing $B_h^\top$ corresponds to solving a heat equation: note the minus sign in the term $-(v, \dot{Z})$ and that we solve starting with $t = T$ and ending with $t = 0$. Changing variables introducing a new time variable $s = T - t$ brings this problem into the form of the usual heat equation. Note that (235.26) is a discrete analog of the problem:

$$\begin{cases} -\dot{z} - \Delta z = 0 & \text{in } \Omega \times (0,T], \\ \partial_n z + \kappa z = 0 & \text{on } \Gamma \times [0,T), \\ z(x,T) = g(x) & \text{for } x \in \Omega, \end{cases}$$

with $G \in V_h$ an approximation of $g(x)$ and $Z(t) \in V_h$ an approximation of $z(\cdot, t)$ for $t \in [0,T]$.

To solve the least squares equation by the gradient or conjugate gradient method, we have to compute $B_h F^n$ and $B_h^\top G$ for given vectors F^n and G in V_h, by using some time stepping method such as the dG(0) or cG(1) method.

EXAMPLE 235.6. We consider the given problem with domain Ω as in the previous examples, $\kappa = 1000$ corresponding to boundary conditions $u \approx 0$, final time $T = 0.1$ and observed state at time T corresponding to initial values $u_0 = 16\, x_1\, (1 - x_1)\, x_2\, (1 - x_2)$ and $u_0 = 2 \min(x_1,\ 1 - x_1,\ x_2,\ 1 - x_2)$, respectively. We then seek to reconstruct these initial data from the observations of the resulting solutions at time $T = 0.1$ with $\mu = 0.001$ and a few conjugate gradient iteration using the cG(1) (initiated by two dG(0) steps to filter out high frequency noise) with timesteps $k = 0.0025$. The results are shown in Fig. 235.5 and 235.6, respectively. The reconstruction error in the first case is ~ 0.057 and in the second case ~ 0.19.

One would think that by decreasing μ it would be possible to better reconstruct the crisp details in the initial data u_0 in the second example. However, the observation we use here is a computed one modelling the fact that observations are imperfect or not considered in full detail in most cases, so that in this case the reconstruction does not get much better by decreasing μ. However, if we decrease T to say 0.02 we can reconstruct also the more detailed structure of u_0 also in the second example with $\mu = 0.00001$ and a resulting reconstruction error ~ 0.068:

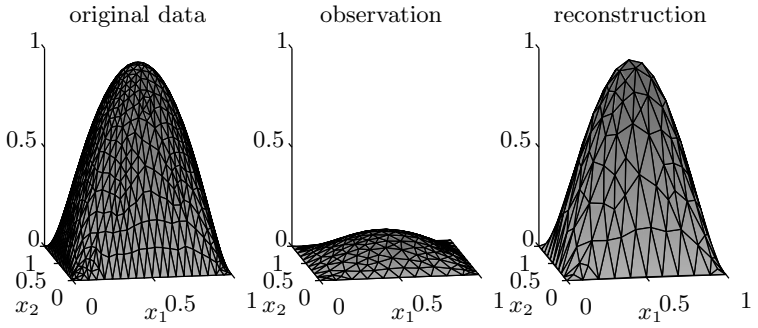


FIGURE 235.5. Original initial data $u_0 = 16\,x_1\,(1-x_1)\,x_2\,(1-x_2)$ (left), corresponding observed state/solution at time $T = 0.1$ (middle), and reconstruction of u_0 (right) obtained with $\mu = 0.001$.

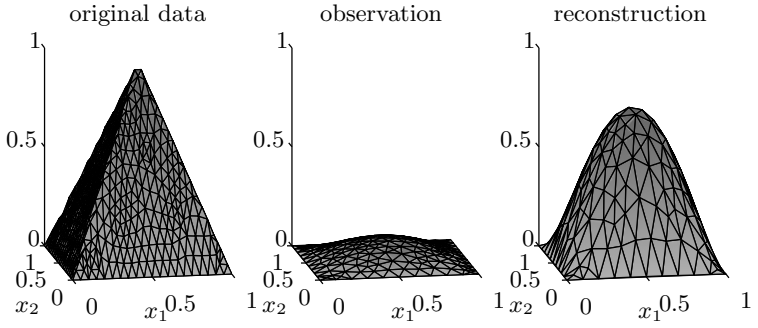


FIGURE 235.6. Original initial data $u_0 = 2\max(\,x_1,\,1-x_1,\,x_2,\,1-x_2)$ (left), corresponding observed state/solution at time $T = 0.1$ (middle), and reconstruction of u_0 (right) obtained with $\mu = 0.001$.

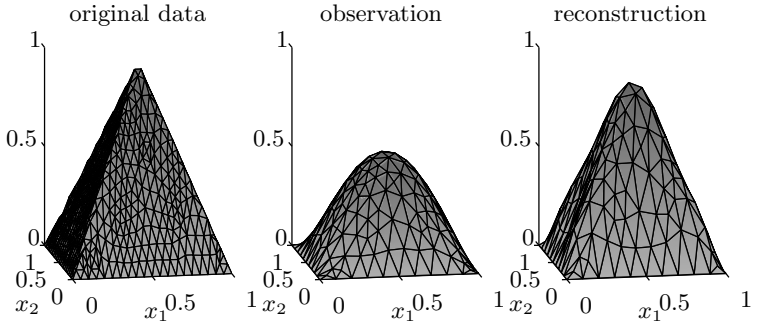


FIGURE 235.7. Original initial data $u_0 = 2\max(\,x_1,\,1-x_1,\,x_2,\,1-x_2)$ (left), corresponding observed state/solution at time $T = 0.02$ (middle), and reconstruction of u_0 (right) obtained with $\mu = 0.00001$.

236
Optimal Control

We're making the right decisions to bring the solution to an end.
(George W. Bush)

236.1 Introduction

In this chapter we continue with aspects of optimization connected to *optimal control* in the following setting: Consider an IVP of the form: Find the *state* $v : [0,T] \to \mathbb{R}^n$ satisfying the *state equation*

$$\dot{v}(t) + f(v(t), q(t)) = 0 \quad 0 < t \leq T, \quad v(0) = u^0, \tag{236.1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is a given function, u^0 a given initial value, and $q : [0,T] \to \mathbb{R}^m$ is a *control*. We seek to determine an *optimal control* $p : [0,T] \to \mathbb{R}^m$ such that $J(p) \leq J(q)$ for all $q : [0,T] \to \mathbb{R}^m$, where

$$J(q) \equiv \frac{1}{2}\|v - \hat{u}\|^2 + \frac{\alpha}{2}\|q\|^2, \tag{236.2}$$

where v solves (236.1), $\hat{u} : [0,T] \to \mathbb{R}^n$ is a given function, and

$$\|w\|^2 = \int_0^T |w(t)|^2 \, dt$$

with $|\cdot|$ denoting the Euclidean norm, and α is a positive constant. We thus seek to choose the control q so that the corresponding state v is as

close as possible to a given state $\hat{u}$ in the $\|\cdot\|$-norm and we also add a cost of the control measured by the factor $\alpha > 0$.

We reformulate this problem as the following *saddle point problem*:

$$\min_{v,q} \max_{\mu} L(v,q,\mu) \tag{236.3}$$

with the *Lagrangian* L defined by

$$L(v,q,\mu) = \frac{1}{2}\|v-\hat{u}\|^2 + \frac{\alpha}{2}\|q\|^2 + (\dot{v} + f(v,q),\mu) \tag{236.4}$$

with $(\cdot,\cdot)$ the scalar product corresponding to the norm $\|\cdot\|$, and (v,q,μ) varying freely (with $v(0) = u^0$ and $\mu(T) = 0$).

The condition for stationarity of $L(v,q,\mu)$ at (u,p,λ) is $L'(u,p,\lambda) = 0$, where L' is the Jacobian of $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, or in component form:

$$(\dot{u} + f(u,p),\mu) = 0 \quad \forall \mu, \tag{236.5}$$

$$(u-\hat{u},v) + (\dot{v} + f'_v(u,p)v,\lambda) = 0 \quad \forall v, \tag{236.6}$$

$$(f'_q(u,p)q,\lambda) + \alpha(p,q) = 0 \quad \forall q, \tag{236.7}$$

where $f'_v(v,q)$ and $f'_q(v,q)$ denote the Jacobians of $f(v,q)$ with respect to v and q at (v,q), respectively, and we assume that $v(0) = 0$ and $\mu(T) = 0$. We can restate these equations in (u,p,λ) pointwise in time as follows:

$$\dot{u} + f(u,p) = 0 \quad \text{on } [0,T], \quad u(0) = u^0, \tag{236.8}$$

$$-\dot{\lambda} + f'_v(u,p)^\top \lambda = \hat{u} - u \quad \text{on } [0,T], \quad \lambda(T) = 0. \tag{236.9}$$

$$f'_q(u,p)^\top \lambda + \alpha p = 0 \quad \text{on } [0,T], \tag{236.10}$$

where $\top$ denotes transpose. Here (236.8) is the state equation, (236.9) is the *costate equation*, and (236.10) is the *feed back control* coupling the *optimal control* p to the *costate* λ.

To solve the stationarity equations we may consider the following *gradient method* in the control p:

$$p^{n+1} = p^n - \kappa(\alpha p^n + f'_q(u^n,p^n)^\top \lambda^n) \tag{236.11}$$

where u^n and λ^n solve the state and costate equations $\dot{u}^n + f(u^n,p^n) = 0$ and $-\dot{\lambda}^n + f'_v(u^n,p^n)^\top \lambda^n = \hat{u} - u^n$, respectively, and $\kappa > 0$ is a *step length*.

EXAMPLE 236.1. If $f(v,p) = Av - Bq$ with A a $n \times n$ and B a $n \times m$ matrix, then the stationarity equations take the form:

$$\dot{u} + Au = Bp \quad \text{on } [0,T], \quad u(0) = u^0, \tag{236.12}$$

$$-\dot{\lambda} + A^\top \lambda = \hat{u} - u \quad \text{on } [0,T], \quad \lambda(T) = 0. \tag{236.13}$$

$$\alpha p = B^\top \lambda \quad \text{on } [0,T]. \tag{236.14}$$

236.2 The Connection Between $\frac{dJ}{dp}$ and $\frac{\partial L}{\partial p}$

We shall now prove that

$$J'(p) = \frac{dJ}{dp}(p) = \frac{\partial L}{\partial p}(u, p, \lambda), \tag{236.15}$$

where the state $u = u(p)$ satisfies the state equation (236.8) with control p, and the costate λ satisfies the costate equation (236.9). We can thus express the gradient $J'(p) = \frac{dJ}{dp}(p)$ of the cost function $J(p)$ in terms of the corresponding state $u = u(p)$ and costate λ, while direct computation of $J'(p)$ requires computation of the derivative $u'(p)$ of the state $u(p)$ with respect to the control p: By the Chain rule we have

$$J'(p)q = \frac{\partial}{\partial \epsilon} J(p + \epsilon q)|_{\epsilon=0} = (u(p) - \hat{u}, u'(p)q) + \alpha(p, q),$$

where we thus want to eliminate $u'(p)$. To do so we differentiate the state equation in the form (assuming for simplicity that $u^0 = 0$),

$$0 = (u, -\dot{\mu}) + (f(u, p), \mu) \quad \forall \mu \text{ with } \mu(T) = 0,$$

with respect to p, to get $\forall \mu$ with $\mu(T) = 0$,

$$0 = \frac{d}{d\epsilon}\big((u(p + \epsilon q), -\dot{\mu}) + (f(u(p + \epsilon q), p + \epsilon q, \mu))\big)|_{\epsilon=0},$$

that is,

$$0 = (u'(p)q, -\dot{\mu}) + (f'_u(u(p), p)u'(p)q + f'_p(u(p), p)q, \mu)$$

or

$$(u'(p)q, -\dot{\mu}) + (u'(p)q, f'_u(u(p), p)^\top \mu) = -(q, f'_p(u(p), p)^\top \mu).$$

Choosing now $\mu = \lambda$ and using that by the costate equation,

$$(u'(p)q, -\dot{\lambda}) + (u'(p)q, f'_u(u(p), p)^\top \lambda) = -(u(p) - \hat{u}, u'(p)q),$$

we can now express $J'(p)$ in the form

$$J'(p)q = (q, f'_p(u(p), p)^\top \lambda) + \alpha(p, q),$$

or

$$J'(p) = f'_p(u(p), p)^\top \lambda + \alpha p = \frac{\partial L}{\partial p}(u, p, \lambda)$$

as we set out to demonstrate.

Through the introduction of the costate λ we are thus able to express the gradient of the cost $J(p)$ with respect to the control p, and we may then apply a gradient method to search for the minimum of $J(p)$.

EXAMPLE 236.2. We consider the problem of balancing an inverted pendulum on a fingertip, when the mass is subject to perturbations of horizontal force and initial condition. Assuming small displacements around the vertical position, the state equation takes the form $\dot{u}_2(t) - u_1(t) = f(t)$ and $\dot{u}_1(t) - u_2(t) = p(t)$ for $0 < t \leq T$, $u_1(0) = u_1^0$, $u_2(0) = u_2^0$, where $f(t)$ is the perturbation and $p(t)$ the control. The optimal control problem of keeping the pendulum in upright position with u_1 and u_2 close to zero, takes the form: Find $p : [0, T] \to \mathbb{R}$ which minimizes the cost

$$J(p) = \frac{1}{2}\int_0^T (a_1 u_1^2(t) + a_2 u_2^2(t))\, dt + \frac{\alpha}{2}\int_0^T p^2(t)\, dt,$$

where (u_1, u_2) solves the state equation with control p, and a_1, a_2 and α are positive constants. In Fig. 236.1 we show the result of applying the gradient method (236.11) for this problem with $f(t) = \sin(2t)+\sin(10t)$, $u_1^0 = 0.3$, $u_2^0 = 0$, $T = 2$, $a_1 = 100$, $a_2 = 1$, $\alpha = 0.0001$ and $\kappa = 0.005$. We note that the weighting with $a_1 >> a_2$ gears the control towards keeping the position $u_1(t)$ close to zero, rather than the velocity $u_2(t)$.

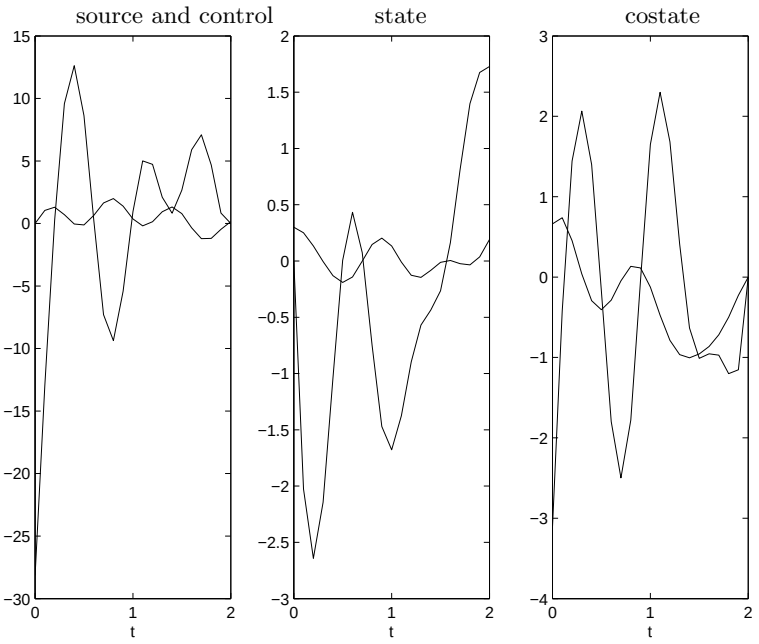


FIGURE 236.1. Source, control, state and costate for the inverse pendulum problem

237
Differential Equations Tool Bag

It seems to me that there are at least four different viewpoints– or extremes of viewpoint– that one may reasonably hold:

1. All thinking is computation; in particular, feelings of conscious awareness are evoked merely by the carrying out of appropriate computations. (Hard AI)

2. Awareness is a feature of the brain's physiological action; and whereas any physical action can be simulated computationally, computational simulation cannot by itself evoke awareness. (Soft AI)

3. Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally. (Penrose's view)

4. Awareness cannot be explained by physical, computational, or any other scientific terms.

(R. Penrose in *Shadows of the Mind*)

237.1 Introduction

We here collect basic facts about solving differential equations analytically and numerically.

237.2 The Equation $u'(x) = \lambda(x)u(x)$

The solution to the scalar initial value problem

$$u'(x) = \lambda(x)u(x) \quad \text{for } x > a, \quad u(a) = u_a,$$

where $\lambda(x)$ is a given function of x, and u_a a given initial value, is

$$u(x) = \exp(\Lambda(x))u_a = e^{\Lambda(x)}u_a,$$

where $\Lambda(x)$ is a primitive function of $\lambda(x)$ such that $\Lambda(a) = 0$. In particular, if λ is a constant, then $u(x) = \exp(\lambda x)u_a$.

237.3 The Equation $u'(x) = \lambda(x)u(x) + f(x)$

The solution the scalar initial value problem

$$u'(x) = \lambda(x)u(x) + f(x) \quad \text{for } x > a, \quad u(a) = u_a,$$

where $\lambda(x)$ and $f(x)$ are given functions of x, and u_a a given initial value, can be expressed using Duhamel's principle in the form

$$u(x) = e^{\Lambda(x)}u_a + e^{\Lambda(x)} \int_a^x e^{-\Lambda(y)} f(y)\, dy.$$

where $\Lambda(x)$ is a primitive function of $\lambda(x)$ such that $\Lambda(a) = 0$.

237.4 The Differential Equation $\sum_{k=0}^n a_k D^k u(x) = 0$

A solution to the constant coefficient differential equation

$$p(D)u(x) = \sum_{k=0}^{n} a_k D^k u(x) = 0, \quad \text{for } x \in I,$$

where I is an interval of real numbers, has the form

$$u(x) = \alpha_1 \exp(\lambda_1) + ... + \alpha_n \exp(\lambda_n),$$

where the α_i are arbitrary constants and the λ_i are the roots of the polynomial equation $p(\lambda) = 0$ with $p(\lambda) = \sum_{k=0}^n a_k \lambda^k$, assuming there are n distinct roots. If $p(\lambda) = 0$ has a multiple root λ_i of multiplicity r, then the solution is the sum of terms of the form form $q(x)\exp(\lambda_i x)$, where $q(x)$ is a polynomial of degree at most $r-1$. For example, if $p(D) = (D-1)^2$, then a solution of $p(D)u = 0$ has the form $u(x) = (a_0 + a_1 x)\exp(x)$.

237.5 The Damped Linear Oscillator

A solution $u(t)$ to

$$\ddot{u} + \mu\dot{u} + ku = 0, \text{for } t > 0,$$

where μ and k are constants, has the form

$$u(t) = ae^{-\frac{1}{2}(\mu+\sqrt{\mu^2-4k})t} + be^{-\frac{1}{2}(\mu-\sqrt{\mu^2-4k})t},$$

if $\mu^2 - 4k > 0$, and

$$u(t) = ae^{-\frac{1}{2}\mu t}\cos(\frac{t}{2}\sqrt{4k-\mu^2}) + be^{-\frac{1}{2}\mu t}\sin(\frac{t}{2}\sqrt{4k-\mu^2}),$$

if $\mu^2 - 4k < 0$, and

$$u(t) = (a+bt)e^{-\frac{1}{2}\mu t},$$

if $\mu^2 - 4k = 0$, where a and b are arbitrary constants.

237.6 The Matrix Exponential

The solution to the initial value problem linear system

$$u'(x) = Au(x) \quad \text{for } 0 < x \leq T,\ u(0) = u_0,$$

where A is a *constant* $d \times d$ matrix, $u_0 \in \mathbb{R}^d$, $T > 0$, is given by

$$u(x) = \exp(xA)u_0 = e^{xA}u_0.$$

If A is diagonalizable so that $A = SDS^{-1}$, where S is nonsingular and D is diagonal with diagonal elements d_i (the eigenvalues of A), then

$$\exp(xA) = S\exp(xD)S^{-1}.$$

where $\exp(xD)$ be the diagonal matrix with diagonal elements equal to $\exp(xd_i)$.

The solution to the initial value problem

$$u'(x) = Au(x) + f(x) \quad \text{for } 0 < x \leq 1,\ u(0) = u_0,$$

where $f(x)$ is a given function, is given by Duhamel's principle:

$$u(x) = \exp(xA)u_0 + \int_0^x \exp((x-y)A)f(y)\,dy.$$

237.7 Fundamental Solutions of the Laplacian

The function $\Phi(x) = \frac{1}{4\pi}\frac{1}{\|x\|}$ for $x \in \mathbb{R}^3$ satisfies the differential equation $-\Delta\Phi = \delta_0$ in $\mathbb{R}^3$, where δ_0 represents a point mass at the origin. The function $\Phi(x) = \frac{1}{2\pi}\log(\frac{1}{\|x\|})$ for $x \in \mathbb{R}^2$ satisfies the differential equation $-\Delta\Phi = \delta_0$ in $\mathbb{R}^2$, where δ_0 represents a point mass at the origin.

237.8 The wave equation in 1d

The general solution to the one-dimensional wave equation

$$\ddot{u} - u'' = 0 \quad \text{for } x, t \in \mathbb{R},$$

is given by $u(x,t) = v(x-t) + w(x+t)$ where $v, w : \mathbb{R} \to \mathbb{R}$ are arbitrary functions.

237.9 Numerical Methods for IVPs

The dG(O), the discontinuous Galerkin method with piecewise constants, for the initial value problem $\dot{u}(t) = f(u(t),t)$ for $t > 0$, $u(0) = u^0$, with $f : \mathbb{R}^{d+1} \to \mathbb{R}^d$, takes the form

$$U^n = U^{n-1} + \int_{t_{n-1}}^{t_n} f(U^n, t)\, dt, \quad n = 1, 2, ...,$$

where $U(t)$ is piecewise constant on a partition $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} < \cdots$, with $U(t) = U^n$ for $t \in (t_{n-1}, t_n]$ and $U(0) = u^0$. With right-end point quadrature we obtain the implicit backward-Euler method:

$$U^n = U^{n-1} + k_n f(U^n, t_n)\, dt, \quad n = 1, 2, ...,$$

where $k_n = t_n - t_{n-1}$. The explicit forward Euler method reads:

$$U^n = U^{n-1} + k_n f(U^{n-1}, t_{n-1})\, dt, \quad n = 1, 2, ...,$$

The cG(1), the continuous Galerkin method with continuous piecewise linear functions, takes the form

$$U(t_n) = U(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(U(t), t)\, dt, \quad n = 1, 2, ...,$$

where $U(t)$ is continuous piecewise linear with nodal values $U(t_n) \in \mathbb{R}^d$ and $U(0) = u^0$.

237.10 cg(1) for Convection-Diffusion-Reaction

The cG(1) finite element method for the scalar convection-diffusion-reaction problem

$$-\nabla \cdot (a\nabla u) + \nabla \cdot (ub) + cu = f \quad \text{in } \Omega,$$
$$a\frac{\partial u}{\partial n} + \kappa u = g \quad \text{on } \Gamma,$$

with Robin boundary conditions, where f and g are given data, and $a > 0$, b, c and $\kappa \geq 0$ are given coefficients, and Ω is a given domain in $\mathbb{R}^2$ with boundary Γ, takes the form: Find $U \in V_h$ such that

$$\int_\Omega a\nabla U \cdot \nabla v \, dx + \int_\Omega \nabla \cdot (ub)v \, dx + \int_\Omega cuv \, dx + \int_\Gamma \kappa uv \, ds$$
$$= \int_\Omega fv \, dx + \int_\Gamma gv \, ds,$$

where V_h is a space of continuous piecewise linear functions on a triangulation of Ω with no restriction on the nodal values on the boundary.

237.11 Svensson's Formula for Laplace's Equation

$$U_{i,j} = \frac{1}{4}(U_{i-1,j} + U_{i+1,j} + U_{i,j-1} + U_{i,j+1}), \quad \text{for } i, j \in \mathbb{Z},$$

where $U_{i,j}$ approximates $u(ih, jh)$ with $h > 0$ and $u : \mathbb{R}^2 \to \mathbb{R}$ solves $\Delta u = 0$.

237.12 Optimal Control

The stationary equations for the saddle point problem $\min_{v,q} \max_\mu L(v, q, \mu)$, with

$$L(v, q, \mu) = \frac{1}{2}\|v - \hat{u}\|^2 + \frac{\alpha}{2}\|q\|^2 + (\dot{v} + f(v, q), \mu)$$

with $(v, w) = \int_0^T v \cdot w \, dt$ and (v, q, μ) varying freely (with $v(0) = u^0$ and $\mu(T) = 0$), take the form:

$$\dot{u} + f(u, p) = 0 \quad \text{on } [0, T], \quad u(0) = u^0, \tag{237.1}$$
$$-\dot{\lambda} + f'_v(u, p)^\top \lambda = \hat{u} - u \quad \text{on } [0, T], \quad \lambda(T) = 0. \tag{237.2}$$
$$f'_q(u, p)^\top \lambda + \alpha p = 0 \quad \text{on } [0, T], \tag{237.3}$$

where $\top$ denotes transpose. Here (237.1) is the state equation, (237.2) is the *costate equation*, and (237.3) is the *feed back control* coupling the *optimal control* p to the *costate* λ.

238
Applications Tool Bag

238.1 Introduction

In this section we collect the basic models of engineering and science expressed as differential equations. For specification of boundary and initial values we refer to the text.

238.2 Malthus' Population Model

$$\dot{u} = \lambda u - \mu u,$$

where $u(t)$ is the population at time t, $\lambda \geq 0$ the birth rate and $\mu \geq 0$ the death rate.

238.3 The Logistics Equation

$$\dot{u} = u(1 - u)$$

238.4 Mass-Spring-Dashpot System

$$m\ddot{u} + \mu\dot{u} + ku = f, \quad \text{((force balance)},$$

where $u(t)$ is the displacement, m is the mass, μ the viscosity, and k the spring constant.

238.5 LCR-Circuit

$$L\ddot{u} + R\dot{u} + \frac{u}{C} = f, \quad \text{((balance of potentials),}$$

where $u(t)$ is a primitive function of the current, L is the inductance, R the resistance, C the capacitance, and f a potential.

238.6 Laplace's Equation for Gravitation

$$-\Delta u = \rho,$$

where $u : \mathbb{R}^3 \to \mathbb{R}$ is the gravitational potential and $\rho(x)$ the mass density.

238.7 The Heat Equation

$$\dot{u} - \nabla \cdot q = f, \quad q = k\nabla u \quad \text{(heat balance and Fourier's law)}$$

where $u(x,t)$ is a temperature, $q(x,t)$ a heat flux, $k(x,t) > 0$ a conduction coefficient and $f(x,t)$ a heat source. If $k = 1$, then we get the heat equation: $\dot{u} - \Delta u = f$.

238.8 The Wave Equation

$$\ddot{u} - \Delta u = f.$$

238.9 Convection-Diffusion-Reaction

$$\dot{u} + \nabla \cdot (\beta u) + \alpha u - \nabla \cdot (\epsilon \nabla u) = f.$$

where $u(x,t)$ a concentration, $\beta(x,t)$ is a convection velocity, $\alpha(x,t)$ a reaction coefficient, $\epsilon(x,t)$ a diffusion coefficient, and $f(x,t)$ a production rate.
.

238.10 Maxwell's Equations

$$\begin{cases} \dfrac{\partial B}{\partial t} + \nabla \times E = 0, & \text{(Faraday's law)} \\ -\dfrac{\partial D}{\partial t} + \nabla \times H = J, & \text{(Ampère's law)} \\ \nabla \cdot B = 0, \quad \nabla \cdot D = \rho, & \text{(Gauss' and Coulomb's laws} \\ B = \mu H, \quad D = \epsilon E, \quad J = \sigma E, & \text{(constitutive laws and Ohm's law)} \end{cases}$$

where E is the electric field, H is the magnetic field , D is the electric displacement, B is the em magnetic flux , J is the electric current, ρ is the charge, μ is the magnetic permeability , ϵ is the dielectric constant, and σ is the electric conductivity.

238.11 The Incompressible Navier-Stokes Equations

$$\frac{\partial u}{\partial t} + (u \cdot \nabla)u + \nabla p - \nu \Delta u = f, \quad \nabla \cdot u = 0,$$

where $u(x,t)$ is the fluid velocity, $p(x,t)$ the pressure, $f(x,t)$ a given force and $\nu > 0$ a constant viscosity.

238.12 Schrödinger's Equation

$$i\frac{\partial \varphi}{\partial t} = \big(-\frac{1}{2}\sum_j \Delta_j + V(r_1, ..., r_N)\big)\varphi(r_1, ..., r_N), \quad r_j \in \mathbb{R}^3.$$

$$i\frac{\partial \varphi}{\partial t} = \big(-\frac{1}{2}\Delta + \frac{1}{|x|}\big)\varphi(x), \quad x \in \mathbb{R}^3, \quad \text{(Hydrogen atom)}.$$

Part XIV

Canon of PDEs

239
Poisson's Equation Analysis

Nature resolves everything to its component atoms and never reduces anything to nothing. (Lucretius)

... on aura donc $\Delta u = 0$; cette équation remarquable nous sera de la plus grande utilité dans la theorie de la figure des corps célestes. (Laplace)

One time I was sitting visiting the show at the Old Copley Theatre, an idea came into my mind which simply distracted all my attention from the performance. It was the notion of an optical machine for harmonic analysis. I had already learned not to disregard these stray ideas, no matter when they came to my attention, and I promptly left the theatre to work out some of the details of my new plan....The projected machine will solve boundary value problems in the field of partial differential equations. (Wiener)

239.1 Introduction

In this chapter, we extend the material of Chapter **??** to Poisson's equation $-\Delta u = f$ in a domain $\Omega \subset \mathbb{R}^d$, where $d = 2$ or $d = 3$, together with various boundary conditions. We begin by presenting some models from physics and mechanics that are modeled by Poisson's equation and describing some of the properties of its solutions. We then discuss the finite

element method for the Poisson equation: constructing the discrete system of linear equations determining the approximation, deriving a priori and a posteriori error estimates, formulating an adaptive error control algorithm, and briefly addressing some implementation issues. The material directly extends e.g. to problems with variable coefficients of the form (232.8) and to three space dimensions using piecewise linear approximation based on tetrahedral meshes.

239.2 Applications of Poisson's equation

We derived Poisson's equation in Chapter **??** as a model of stationary heat conduction. Poisson's equation is the prototype of the class of elliptic equations and has numerous applications in physics and mechanics. These include

- *Elasticity.* The model (**??**) of the deflection of an elastic string discussed in Chapter **??** can be extended to describe the transversal deflection due to a transverasal load of a horizontal elastic membrane of uniform tension stretched over a plane curve Γ enclosing a region Ω in $\mathbb{R}^2$; see Fig. 239.1. The equation takes the form of the Poisson equation $-\Delta u = f$ in Ω together with the boundary condition $u = 0$ on Γ, where $f(x)$ is the transversal load.

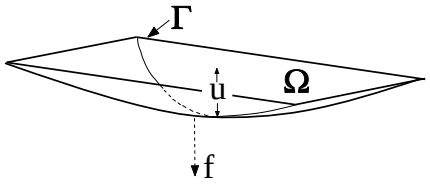


FIGURE 239.1. An elastic membrane under the load f supported at Γ.

- *Electrostatics.* A basic problem in *electrostatics* is to describe the *electric field* $E(x)$ in a volume Ω containing charges of density $\rho(x)$ and enclosed by a perfectly conducting surface Γ. *Coulomb's law*, one of the famous *Maxwell equations* describing electromagnetic phenomena, can be written

$$\nabla \cdot E = \rho \quad \text{in } \Omega. \tag{239.1}$$

  It follows from Faraday's law $\nabla \times E = 0$ (see Chapter **??** below), that the electric field E is the gradient of a scalar *electric potential* φ, i.e. $E = \nabla\varphi$. This leads to the Poisson equation $\Delta\varphi = \rho$ with a Dirichlet boundary condition $\varphi = c$ on Γ, where c is a constant.

- *Fluid mechanics.* The velocity field u of rotation-free fluid flow satisfies $\nabla \times u = 0$, from which it follows that $u = \nabla\varphi$ where φ is a (scalar) velocity potential. If the fluid is incompressible, then $\nabla \cdot u = 0$, and we obtain the Laplace equation $\Delta\varphi = 0$ for the potential of rotation-free incompressible flow. At a solid boundary, the normal velocity is zero, which translates to a homogeneous Neumann boundary condition for the potential. Note that fluid flow is rarely rotation-free in the whole region occupied by the fluid. In particular, if the fluid is viscous, then rotation is generated at solid boundaries.

- *Statistical physics.* The problem is to describe the motion of particles inside a container Ω that move at random until they hit the boundary where they stop. We illustrate this in Fig. 239.2. Suppose the

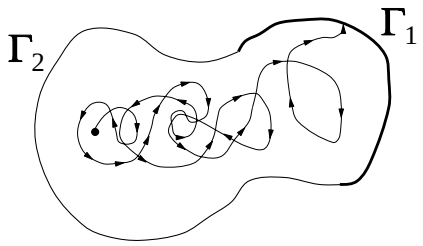


FIGURE 239.2. An illustration of Brownian motion.

boundary Γ of Ω is divided into two pieces $\Gamma = \Gamma_1 \cup \Gamma_2$. Let $u(x)$ be the probability that a particle starting at $x \in \Omega$ winds up stopping at some point on Γ_1, so that $u(x) = 1$ means that it is certain and $u(x) = 0$ means it never happens. It turns out that u solves $\Delta u = 0$ in Ω together with $u = 1$ on Γ_1 and $u = 0$ on Γ_2. Note that the solution of this problem is not continuous on the boundary.

239.3 Solution by Fourier series

For special domains, it is possible to write down a formula for the solution of Poisson's equation using Fourier series. For example in Cartesian coordinates, this is possible if the domain is a square or cube. Using polar, cylindrical or spherical coordinates, the set of domains for which Fourier's method may be used includes discs, cylinders, and spheres.

As an illustration, we use Fourier series to solve Poisson's equation $-\Delta u = f$ in a cube $\Omega = (0, \pi) \times (0, \pi) \times (0, \pi)$ with homogeneous Dirichlet boundary conditions. Because the sides of the cube are parallel to the coordinate axes, we can use *separation of variables* to reduce the problem to

finding a Fourier series in each variable independently. We start by seeking a solution of the eigenvalue problem $-\Delta v = \lambda v$ in Ω, with $v = 0$ on the boundary of Ω, of the form

$$v(x_1, x_2, x_3) = V_1(x_1)V_2(x_2)V_3(x_3),$$

where each factor satisfies an independent boundary condition $V_i(0) = V_i(\pi) = 0$, $i = 1, 2, 3$. Substituting this into the differential equation yields

$$\frac{V_1''}{V_1} + \frac{V_2''}{V_2} + \frac{V_3''}{V_3} = -\lambda.$$

Because x_1, x_2, and x_3 vary independently, each term V_i''/V_i must be constant. Denoting this constant by λ_i we find that each V_i must solve

$$V_i'' + \lambda_i V_i = 0 \quad \text{in } (0, \pi), \quad V_i(0) = V_i(\pi) = 0.$$

This is the one-dimensional eigenvalue problem considered in Section **??** with solution $V_i(x_i) = \sin(jx_i)$ and $\lambda_i = j^2$, where j is an arbitrary integer. It follows that

$$\lambda = \lambda_{jkl} = j^2 + k^2 + l^2, \tag{239.2}$$

for integers j, k, and l with the corresponding eigenfunction

$$v = v_{jkl} = \sin(jx_1)\sin(kx_2)\sin(lx_3).$$

Using the orthogonality of the eigenfunctions, the solution u can be expressed as a Fourier series

$$u(x) = \sum_{j,k,l} A_{jkl} \sin(jx_1)\sin(kx_2)\sin(lx_3),$$

with Fourier coefficients

$$A_{jkl} = \lambda_{jkl}^{-1} \left(\frac{2}{\pi}\right)^3 \int_\Omega f(x)\sin(jx_1)\sin(kx_2)\sin(lx_3)\,dx.$$

The discussion about convergence is nearly the same as in one dimension. In particular, if $f \in L_2(\Omega)$ then the Fourier series of u converges and defines a solution of the given Poisson equation.

239.1. Prove the formula for A_{jkl}.

239.2. Prove that the set of eigenfunctions $\{v_{jkl}\}$ are pairwise orthogonal.

239.3. (a) Compute the Fourier series for the solution of $-\Delta u = 1$ in the square $(0, \pi) \times (0, \pi)$ with homogeneous Dirichlet boundary conditions. (b) Do the same with the Dirichlet condition replaced by a Neumann condition on one side of the square.

Note that there can be several different eigenfunctions for a specific eigenvalue. The *multiplicity* of an eigenvalue is the number of linearly independent eigenvectors that share that eigenvalue. Computing the multiplicity of an eigenvalue λ given by (239.2) is equivalent to determining the number of ways λ be written as a sum of the squares of three integers counting order. For example, $\lambda = 6$ has multiplicity three because $6 = 2^2 + 1 + 1 = 1 + 2^2 + 1 = 1 + 1 + 2^2$.

239.4. Show that $\lambda = 17$ is an eigenvalue of the Poisson equation posed on $(0, \pi)^3$ with Dirichlet boundary conditions and compute its multiplicity.

239.4 Gravitational fields and fundamental solutions

In his famous treatise *Mécanique Céleste* in five volumes published 1799-1825, Laplace extended Newton's theory of gravitation and in particular developed a theory for describing gravitational fields based on using gravitational potentials that satisfy Laplace's equation, or more generally Poisson's equation.

We consider a gravitational field in $\mathbb{R}^3$ with gravitational force $F(x)$ at position x, generated by a distribution of mass of density $\rho(x)$. We recall that the work of a unit mass, moving along a curve Γ joining a point A to a point B, is given by

$$\int_\Gamma F_\tau \, ds,$$

where F_τ is the component of F in the direction of the tangent to the curve. We illustrate this in Fig. 239.3. If the path Γ is closed, then the total work

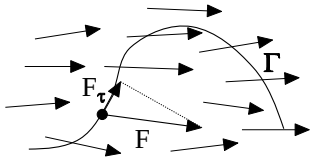


FIGURE 239.3. The motion of a particle in a field F along a curve Γ.

performed is zero. By Stokes' theorem, it follows that a gravitational field F satisfies $\nabla \times F = 0$ and using the results in Chapter **??**, we conclude that F is the gradient of a scalar potential u, i.e.

$$F = \nabla u. \tag{239.3}$$

Laplace proposed the following relation between the gravitational field F and the mass distribution ρ:

$$-\nabla \cdot F = g\rho, \tag{239.4}$$

where g is a gravitational constant. This is analogous to Coulomb's law $\nabla \cdot E = \rho$ in electrostatics, see (239.1), and also to the energy balance equation $\nabla \cdot q = f$ for stationary heat conduction, where q is the heat flux and f a heat source, which we derived in Chapter **??**. A corresponding "derivation" of (239.4) does not appear to be available, reflecting that the nature of gravitation is not yet understood. In particular, (239.4) suggests that $\nabla \cdot F(x) = 0$ at points x where there is no mass so that $\rho(x) = 0$. Combining (239.3) and (239.4), we obtain Poisson's equation $-\Delta u = g\rho$ for the gravitational potential u. In particular, the potential satisfies Laplace's equation $\Delta u = 0$ in empty space.

Newton considered gravitational fields generated by *point masses*. We recall that a unit point mass at a point $z \in \mathbb{R}^3$ is represented mathematically by the *delta function* δ_z at z, which is defined by the property that for any smooth function v,

$$\int_{\mathbb{R}^3} \delta_z \, v \, dx = v(z),$$

where the integration is to be interpreted in a generalized sense. Actually, δ_z is a *distribution*, not a proper function, and there is no conventional "formula" for it; instead we define the delta function by its action inside an average of a smooth function.

Formally, the gravitational potential $E(x)$ (avoid confusion with the notation for an electric field used above) corresponding to a unit point mass at the origin should satisfy

$$-\Delta E = \delta_0 \quad \text{in } \mathbb{R}^3, \tag{239.5}$$

where we assumed that the gravitational constant is equal to one. To give a precise meaning to this equation, we first formally multiply by a smooth test function v vanishing outside a bounded set, to get

$$-\int_{\mathbb{R}^3} \Delta E(x) v(x) \, dx = v(0). \tag{239.6}$$

Next, we rewrite the left-hand side formally integrating by parts using Green's formula (**??**) to move the Laplacian from E to v, noting that the boundary terms dissappear since v vanishes outside a bounded set. We may thus reformulate (239.5) as seeking a potential $E(x)$ satisfying

$$-\int_{\mathbb{R}^3} E(x) \Delta v(x) \, dx = v(0), \tag{239.7}$$

for all smooth functions $v(x)$ vanishing outside a bounded set. This is a weak formulation of (239.5), which is perfectly well defined since now the Laplacian acts on the smooth function $v(x)$ and the potential E is assumed to be integrable. We also require the potential $E(x)$ to decay to zero as $|x|$

tends to infinity, which corresponds to a "zero Dirichlet boundary condition at infinity".

In Chapter **??**, we showed that the function $1/|x|$ satisfies Laplace's equation $\Delta u(x) = 0$ for $0 \neq x \in \mathbb{R}^3$, while it is singular at $x = 0$. We shall prove that the following scaled version of this function satisfies (239.7):

$$E(x) = \frac{1}{4\pi}\frac{1}{|x|}. \tag{239.8}$$

We refer to this function as the *fundamental solution* of $-\Delta$ in $\mathbb{R}^3$. We conclude in particular that the gravitational field in $\mathbb{R}^3$ created by a unit point mass at the origin is proportional to

$$F(x) = \nabla E(x) = -\frac{1}{4\pi}\frac{x}{|x|^3},$$

which is precisely Newton's inverse square law of gravitation. Laplace thus gives a motivation why the exponent should be two, which Newton did not (and therefore was criticized by Leibniz). Of course, it still remains to motivate (239.4). In the context of heat conduction, the fundamental solution $E(x)$ represents the stationary temperature in a homogeneous body with heat conductivity equal to one filling the whole of $\mathbb{R}^3$, subject to a concentrated heat source of strength one at the origin and with the temperature tending to zero as $|x|$ tends to infinity.

We now prove that the function $E(x)$ defined by (239.8) satisfies (239.7). We then first note that since Δv is smooth and vanishes outside a bounded set, and $E(x)$ is integrable over bounded sets, we have

$$\int_{\mathbb{R}^3} E\Delta v\,dx = \lim_{a\to 0^+}\int_{D_a} E\Delta v\,dx, \tag{239.9}$$

where $D_a = \{x \in \mathbb{R}^3 : a < |x| < a^{-1}\}$, with a small, is a bounded region obtained from $\mathbb{R}^3$ by removing a little sphere of radius a with boundary surface S_a and also points further away from the origin than a^{-1}, see Fig. 239.4. We now use Green's formula (**??**) on D_a with $w = E$. Since v is zero for $|x|$ large, the integrals over the outside boundary vanish when a is sufficiently small. Using the fact that $\Delta E = 0$ in D_a, $E = 1/(4\pi a)$ on S_a and $\partial E/\partial n = 1/(4\pi a^2)$ on S_a with the normal pointing in the direction of the origin, we obtain

$$-\int_{D_a} E\Delta v\,dx = \int_{S_a}\frac{1}{4\pi a^2}\,v\,ds - \int_{S_a}\frac{1}{4\pi a}\frac{\partial v}{\partial n}\,ds = I_1(a) + I_2(a),$$

with the obvious definitions of $I_1(a)$ and $I_2(a)$. Now, $\lim_{a\to 0} I_1(a) = v(0)$ because $v(x)$ is continuous at $x = 0$ and the surface area of S_a is equal to $4\pi a^2$, while $\lim_{a\to 0} I_2(a) = 0$. The desired equality (239.7) now follows recalling (239.9).

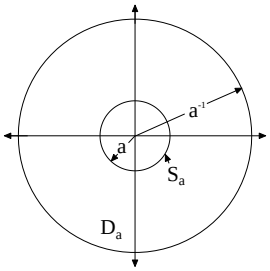


FIGURE 239.4. A cross-section of the domain D_a.

The corresponding fundamental solution of $-\Delta$ in $\mathbb{R}^2$ is given by

$$E(x) = \frac{1}{2\pi}\log(\frac{1}{|x|}). \tag{239.10}$$

In this case the fundamental solution is not zero at infinity.

239.5. Prove that (239.10) is a fundamental solution of $-\Delta$ in $\mathbb{R}^2$.

239.6. Because the presented mathematical models of heat flow and gravitation, namely Poisson's equation, are the same, it opens the possibility of thinking of a gravitational potential as "temperature" and a gravitational field as "heat flux". Can you "understand" something about gravitation using this analogy?

Replacing 0 by an arbitrary point $z \in \mathbb{R}^3$, (239.7) becomes

$$-\int_{\mathbb{R}^3} E(z-x)\Delta v(x)\,dx = v(z), \tag{239.11}$$

which leads to a solution formula for Poisson's equation in $\mathbb{R}^3$. For example, if u satisfies the Poisson equation $-\Delta u = f$ in $\mathbb{R}^3$ and $|u(x)| = O(|x|^{-1})$ as $|x| \to \infty$, then u may be represented in terms of the fundamental solution E and the right-hand side f as follows:

$$u(z) = \int_{\mathbb{R}^3} E(z-x)f(x)\,dx = \frac{1}{4\pi}\int_{\mathbb{R}^3}\frac{f(x)}{|z-x|}\,dx. \tag{239.12}$$

We see that $u(z)$ is a mean value of f centered around z weighted so that the influence of the values of $f(x)$ is inversely proportional to the distance from z.

239.7. Present a corresponding solution formula in the case $d = 2$.

Similarly, the potential u resulting from a distribution of mass of density $\rho(x)$ on a (bounded) surface Γ in $\mathbb{R}^3$ is given by

$$u(z) = \frac{1}{4\pi}\int_\Gamma \frac{\rho(\cdot)}{|z-\cdot|}\,ds, \tag{239.13}$$

where the dot indicates the integration variable. Formally we obtain this formula by simply adding the potentials from all the different pieces of mass on Γ. One can show that the potential u defined by (239.13) is continuous in $\mathbb{R}^3$ if ρ is bounded on Γ, and of course u satisfies Laplace's equation away from Γ. Suppose now that we would like to determine the distribution of mass ρ on Γ so that the corresponding potential u defined by (239.13) is equal to a given potential u_0 on Γ, that is we seek in particular a function u solving the boundary value problem $\Delta u = 0$ in Ω and $u = u_0$ on Γ, where Ω is the volume enclosed by Γ. This leads to the following *integral equation*: given u_0 on Γ find the function ρ on Γ, such that

$$\frac{1}{4\pi}\int_\Gamma \frac{\rho(y)}{|x-y|}\,ds = u_0(x) \quad \text{for } x \in \Gamma. \tag{239.14}$$

This is a *Fredholm integral equation of the first kind*, named after the Swedish mathematician Ivar Fredholm (1866-1927). In the beginning of the 20th century, Fredholm and Hilbert were competing to prove the existence of solutions of the basic boundary value problems of mechanics and physics using integral equation methods. The integral equation (239.14) is an alternative way of formulating the boundary value problem of finding u such that $\Delta u = 0$ in Ω, and $u = u_0$ on Γ. Integral equations may also be solved using Galerkin methods. We return to the topic of integral equations and their numerical solution in the advanced volume.

239.8. Show that the potential from a uniform distribution of mass on the surface of a sphere is given as follows: (a) outside the sphere the potential is the same as the potential from a point mass at the origin of the sphere with the same mass as the total surface mass. (b) inside the sphere the potential is constant. Hint: rewrite the surface integral in spherical coordinates and consult a calculus book to evaluate the resulting standard integral.

239.5 Green's functions

There is an analog of the formula (239.12) for the solution of Poisson's equation in a bounded domain Ω based on using a *Green's function*, which is the analog of the fundamental solution on a domain different from $\mathbb{R}^d$. The Green's function $G_z(x)$ for the Laplace operator with homogeneous Dirichlet boundary conditions on a bounded domain Ω with boundary Γ satisfies:

$$\begin{cases} -\Delta G_z(x) = \delta_z(x) & \text{for } x \in \Omega, \\ G_z(x) = 0 & \text{for } x \in \Gamma. \end{cases}$$

$G_z(x)$ has a singularity at z corresponding to that of the fundamental solution and in this sense, it is a modified fundamental solution that satisfies the Dirichlet boundary condition. In heat conduction, the Green's function

$G_z(x)$ represents the stationary temperature in a homogeneous heat conducting body occupying Ω with zero temperature at its boundary subjected to a concentrated heat source at $z \in \Omega$. It is possible to compute G_z for special domains. For example if $\Omega = \{x : |x| < a\}$ is the ball of radius a in $\mathbb{R}^3$ centered at the origin, then

$$G_z(x) = \frac{1}{4\pi|x-z|} - \frac{1}{4\pi|\ |z|x/a - az/|z|\ |}. \tag{239.15}$$

239.9. Verify (239.15).

239.10. Determine the Green's function for a "half space" defined as a part of $\mathbb{R}^3$ that has a given plane as a boundary. Hint: consider the function $(|x-z|^{-1} - |x-z^*|^{-1})/(4\pi)$, where z^* is obtained from z by reflection in the plane defining the half space.

If u satisfies $-\Delta u = f$ in Ω and $u = g$ on Γ, then using Green's formula as above we find that the solution u can be represented as

$$u(z) = -\int_\Gamma g\,\partial_n G_z\,ds + \int_\Omega f\,G_z\,dx. \tag{239.16}$$

In the case Ω is the ball of radius a and $f = 0$, so that

$$u(z) = \frac{a^2 - |z|^2}{2^{d-1}\pi a}\int\limits_{S_a} gK_z\,ds, \tag{239.17}$$

with $S_a = \{x : |x| = a\}$ and $K_z(x) = |x-z|^{-d}$, the representation (239.16) is called *Poisson's formula* for harmonic functions. We note in particular that the value at the center of the sphere S_a is equal to the mean value of u on the surface of the sphere, i.e.

$$u(0) = \frac{1}{(2a)^{d-1}\pi}\int\limits_{S_a} u\,ds.$$

Thus a harmonic function has the property that the value at a point is equal to its spherical mean values.

239.11. Verify (239.16) and (239.17).

In general it is difficult to use (239.16) to compute a solution of Poisson's equation, since finding a formula for the Green function for a general domain is difficult. Moreover, integrals over the entire domain and its boundary have to be evaluated for each value $u(z)$ desired.

239.6 The differentiability of solutions

The Poisson formula may be used to show that a bounded function u satisfying $\Delta u = 0$ in a domain Ω has derivatives of any order inside Ω. Thus a harmonic function is *smooth* inside the domain where it is harmonic. This is because the function $|z-x|^{-d}$ is differentiable with respect to z any number of times as long as $x \neq z$, and if x is strictly inside Ω then the sphere $|z - x| = a$ is contained in Ω for a sufficiently small, so that the Poisson representation formula may be used. Thus a bounded solution u of $\Delta u = 0$ in Ω is smooth away from the boundary of Ω. On the other hand, it may very well have singularities on the boundary; we discuss this below. These results carry over to solutions of the Poisson equation $-\Delta u = f$ in Ω: if f is smooth inside Ω then so is u.

240
Poisson's Equation FEM

A man who was famous as a tree climber was guiding someone in climbing a tall tree. He ordered the man to cut the top branches, and, during this time, when the man seemed in great danger, the expert said nothing. Only when the man was coming down and had reached the height of the eaves did the expert call out, "Be careful! Watch your step coming down!" I asked him, "Why did you say that? At that height he could jump the rest of the way if he chose."
"That's the point," said the expert. "As long as the man was up at a dizzy height and the branches were threatening to break, he himself was so afraid I said nothing. Mistakes are always made when people get to easy places." (Kenko, translated by D. Keene)

240.1 Variational Formulation

We present the finite element method with piecewise linear approximation for the Poisson equation with homogeneous Dirichlet boundary conditions

$$\begin{cases} -\Delta u(x) = f(x) & \text{for } x \in \Omega, \\ u(x) = 0 & \text{for } x \in \Gamma, \end{cases} \tag{240.1}$$

where Ω is a bounded domain in $\mathbb{R}^2$ with polygonal boundary Γ,

Generalizing the procedure used in one dimension from Chapter **??**, we first give (240.1) the following variational formulation: find $u \in V$ such that

$$(\nabla u, \nabla v) = (f, v) \quad \text{for all } v \in V, \tag{240.2}$$

where

$$(w, v) = \int_\Omega wv\, dx, \quad (\nabla w, \nabla v) = \int_\Omega \nabla w \cdot \nabla v\, dx,$$

and

$$V = \left\{ v : \int_\Omega (|\nabla v|^2 + v^2) dx < \infty \text{ and } v = 0 \text{ on } \Gamma \right\}. \tag{240.3}$$

A detailed motivation for the choice of V is given in Chapter **??**. Here we note that if v and w belong to V, then $(\nabla v, \nabla w)$ is well defined, and if $v \in V$ and $f \in L_2(\Omega)$, then (f, v) is well defined. This follows from Cauchy's inequality. Thus, (240.2) makes sense. In fact, we may think of V as the largest space with this property.

As in the one-dimensional case, we now seek to show that (240.1) and (240.2) have the same solution if f is smooth. First, to see that a solution u of (240.1) with continuous second derivatives (requiring f to be continuous) also is a solution of the variational problem (240.2), we multiply $-\Delta u = f$ by $v \in V$ and use Green's formula to get

$$\int_\Omega fv\, dx = -\int_\Omega \Delta u\, v\, dx = -\int_\Gamma \partial_n uv\, ds + \int_\Omega \nabla u \cdot \nabla v\, dx = \int_\Omega \nabla u \cdot \nabla v\, dx,$$

where the boundary condition $v = 0$ on Γ was used to eliminate the boundary integral over Γ. Conversely, assuming that the solution of (240.2) has continuous second derivatives, we can use Green's formula in (240.2) to put two derivatives back on u, again using the boundary conditions on v, to get

$$\int_\Omega (-\Delta u - f)\, v\, dx = 0 \quad \text{for all } v \in V. \tag{240.4}$$

Now suppose that $-\Delta u - f$ is non-zero, say positive, at some point $x \in \Omega$. Since $-\Delta u - f$ is continuous, it is therefore positive in some small neighborhood of x contained in Ω. We choose v to be a smooth "hill" that is zero outside the neighborhood and positive inside. It follows that $(-\Delta u - f)v$ is positive in the small neighborhood and zero outside, which gives a contradiction in (240.4). It remains to show that the solution u of (240.2) in fact has continuous second order derivatives if f is continuous; we prove such a regularity result in Chapter **??**. We conclude that the differential equation (240.1) and the variational problem (240.2) have the same solution if the data f is continuous. As in the one-dimensional case, the variational problem (240.2) is meaningful for a wider set of data including $f \in L_2(\Omega)$.

240.1. Prove that the set of functions that are continuous and piecewise differentiable on Ω and vanish on Γ, is a subspace of V.

240.2. Assuming that a solution of (240.2) is continuous on $\Omega \cup \Gamma$, show that it is unique. Hint: choose $v = u$ and use the continuity of u.

240.3. Provide the details of the equivalence of (240.1) and (240.2).

The variational problem (240.2) is equivalent to the following quadratic *minimization problem*: find $u \in V$ such that

$$F(u) \leq F(v) \quad \text{for all } v \in V, \tag{240.5}$$

where

$$F(v) = \frac{1}{2}\int_\Omega |\nabla v|^2\,dx - \int_\Omega f v\,dx.$$

The quantity $F(v)$ may be interpreted as the *total energy* of the function $v \in V$ composed of the *internal energy* $\frac{1}{2}\int_\Omega |\nabla v|^2\,dx$ and the *load potential* $-\int_\Omega f v\,dx$. Thus, the solution u minimizes the total energy $F(v)$ over V. In Chapter **??** we prove existence of a unique solution to the minimization problem (240.5) and thus existence of a unique solution to the variational problem (240.2) and consequently to (240.1).

240.4. Prove the equivalence of (240.5) and (240.2).

240.2 The finite element method

Let $\mathcal{T}_h = \{K\}$ be a triangulation of Ω with mesh function $h(x)$ and let V_h be the corresponding finite element space of continuous piecewise linear functions vanishing on Γ. The finite element space V_h is a subspace of the space V defined by (240.3). Let $\mathcal{N}_h = \{N\}$ denote the set of *internal nodes* N and $\mathcal{S}_h = \{S\}$ the set of *internal edges* S of $\mathcal{T}_h$. We exclude the nodes and edges on the boundary because of the homogeneous Dirichlet boundary condition. Let $\{N_1, ..., N_M\}$ be an enumeration of the internal nodes $\mathcal{N}_h$, and $\{\varphi_1, ..., \varphi_M\}$ the corresponding nodal basis for V_h.

The finite element method for (240.1) reads: find $U \in V_h$ such that

$$(\nabla U, \nabla v) = (f, v) \quad \text{for all } v \in V_h. \tag{240.6}$$

As in one dimension, we can interpret this as demanding that U solve the Poisson equation in an "average" sense corresponding to the residual of U being "orthogonal" in a certain sense to V_h. More precisely, using the fact that $(\nabla u, \nabla v) = (f, v)$ for $v \in V_h$ because $V_h \subset V$, (240.6) is equivalent to

$$(\nabla u - \nabla U, \nabla v) = 0 \quad \text{for all } v \in V_h, \tag{240.7}$$

which expresses the Galerkin orthogonality of the finite element approximation.

240.5. Prove that if (240.6) holds with v equal to each of the nodal basis functions V_h, then (240.6) holds for all $v \in V_h$.

240.3 The discrete system of equations

Expanding U in terms of the basis functions $\{\varphi_i\}$ as

$$U = \sum_{j=1}^{M} \xi_j \varphi_j, \quad \text{where } \xi_j = U(N_j),$$

substituting this into (240.6) and choosing $v = \varphi_i$, gives

$$\sum_{j=1}^{M} (\nabla\varphi_j, \nabla\varphi_i)\xi_j = (f, \varphi_i), \quad i = 1, ..., M.$$

This is equivalent to the linear system of equations

$$A\xi = b, \tag{240.8}$$

where $\xi = (\xi_i)$ is the vector of nodal values, $A = (a_{ij})$ is the *stiffness matrix* with elements $a_{ij} = (\nabla\varphi_j, \nabla\varphi_i)$ and $b = (b_i) = (f, \varphi_i)$ is the *load vector*. The stiffness matrix A is obviously symmetric and it is also positive-definite since for any $v = \sum_i \eta_i\varphi_i$ in V_h,

$$\sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j = \sum_{i,j=1}^{M} \eta_i (\nabla\varphi_i, \nabla\varphi_j)\eta_j$$
$$= \left(\nabla \sum_{i=1}^{M} \eta_i\varphi_i, \nabla \sum_{j}^{M} \eta_j\varphi_j \right) = (\nabla v, \nabla v) > 0,$$

unless $\eta_i = 0$ for all i. This means in particular that (240.8) has a unique solution ξ.

Similarly, we determine the linear system determining the L_2 projection $P_h v$ of a function $v \in L_2(\Omega)$ into V_h defined by

$$(P_h v, w) = (v, w) \quad \text{for all } w \in V_h.$$

Substituting $P_h v = \sum_j \eta_j\varphi_j$ and choosing $w = \varphi_i$, $i = 1, ..., M$, we obtain the linear system

$$M\eta = b, \tag{240.9}$$

where the *mass matrix* M has coefficients (φ_j, φ_i) and the data vector b has coefficients (v, φ_i).

240.6. Prove that the mass matrix is symmetric and positive definite.

240.4 The discrete Laplacian

It will be convenient below to use a discrete analog Δ_h of the Laplacian Δ defined as follows: For a given $w \in V$, let $\Delta_h w$ be the unique function in V_h that satisfies

$$-(\Delta_h w, v) = (\nabla w, \nabla v) \quad \text{for all } v \in V_h. \tag{240.10}$$

In particular, if $w \in V_h$, denoting the nodal values of w by the vector η and those of $\Delta_h w$ by ζ, we find that (240.10) is equivalent to the system of equations $-M\zeta = A\eta$, where M is the mass matrix and A the Poisson stiffness matrix. In other words, the nodal values of the *discrete Laplacian* $\Delta_h w$ of the function $w \in V_h$ with nodal values η, are given by $-M^{-1}A\eta$. We may think of Δ_h as a linear operator on V_h corresponding to multiplication of nodal values by the matrix $-M^{-1}A$. Using Δ_h, we may express the finite element problem (240.6) as finding $U \in V_h$ such that

$$-\Delta_h U = P_h f, \tag{240.11}$$

where P_h is the L_2 projection onto V_h. If w is smooth we may write (240.10) also as

$$(\Delta_h w, v) = (\Delta w, v) \quad \text{for all } v \in V_h, \tag{240.12}$$

which is the same as to say that $\Delta_h w = P_h \Delta w$. Usually, we don't actually compute $\Delta_h w$, but we shall see that the notation is handy.

240.7. Verify (240.11).

240.5 An example: uniform triangulation of a square

We compute the stiffness matrix and load vector explicitly on the uniform triangulation of the square $\Omega = [0,1] \times [0,1]$ pictured in Fig. 240.1. We choose an integer $m \geq 1$ and set $h = 1/(m+1)$, then construct the triangles as shown. The diameter of the triangles in $\mathcal{T}_h$ is $\sqrt{2}h$ and there are $M = m^2$ internal nodes. We number the nodes starting from the lower left and moving right, then working up across the rows.

In Fig. 240.2, we show the support of the basis function corresponding to the node N_i along with parts of the basis functions for the neighboring nodes. As in one dimension, the basis functions are "almost" orthogonal in the sense that only basis functions φ_i and φ_j sharing a common triangle in their supports yield a non-zero value in $(\nabla\varphi_i, \nabla\varphi_j)$. We show the nodes neighboring N_i in Fig. 240.3. The support of any two neighboring basis functions overlap on just two triangles, while a basis function "overlaps itself" on six triangles.

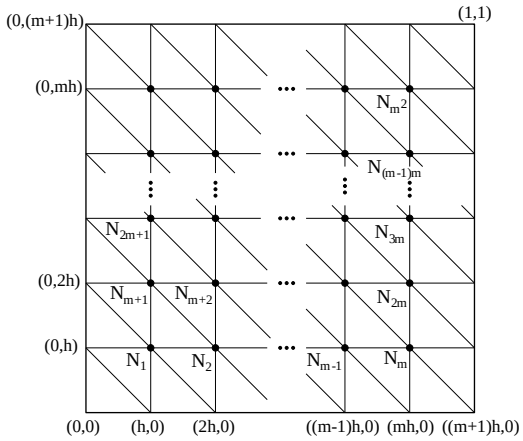


FIGURE 240.1. The standard triangulation of the unit square.

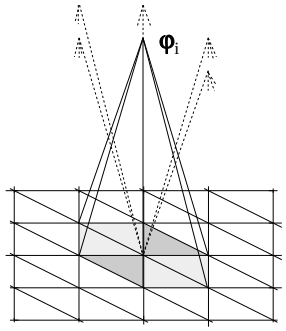


FIGURE 240.2. The support of the basis function φ_i together with parts of the neighboring basis functions.

We first compute

$$(\nabla\varphi_i, \nabla\varphi_i) = \int_\Omega |\nabla\varphi_i|^2\, dx = \int_{\text{support of } \varphi_i} |\nabla\varphi_i|^2\, dx,$$

for $i = 1, ..., m^2$. As noted, we only have to consider the integral over the domain pictured in Fig. 240.3, which is written as a sum of integrals over the six triangles making up the domain. Examining φ_i on these triangles, see Fig. 240.3, we see that there are only two different integrals to be computed since φ_i looks the same, except for orientation, on two of the six triangles and similarly the same on the other four triangles. We shade the corresponding triangles in Fig. 240.2. The orientation affects the direction of $\nabla\varphi_i$ of course, but does not affect $|\nabla\varphi_i|^2$.

We compute $(\nabla\varphi_i, \nabla\varphi_i)$ on the triangle shown in Fig. 240.4. In this case,

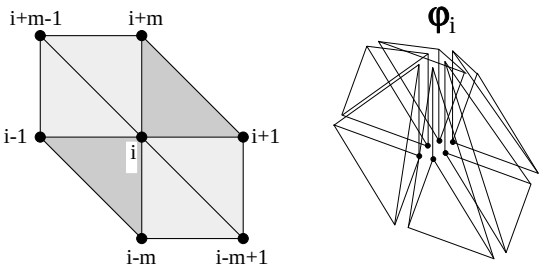


FIGURE 240.3. The indices of the nodes neighboring N_i and an "exploded" view of φ_i.

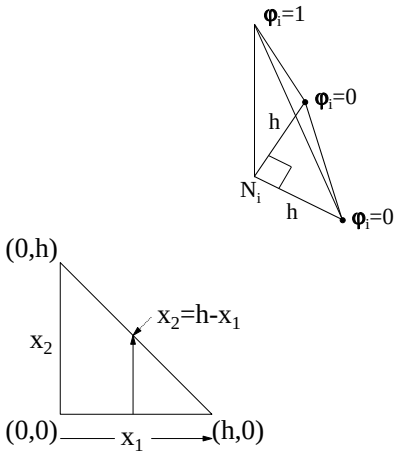


FIGURE 240.4. First case showing φ_i on the left together with the variables used in the reference triangle.

φ_i is one at the node located at the right angle in the triangle and zero at the other two nodes. We change coordinates to compute $(\nabla\varphi_i, \nabla\varphi_i)$ on the *reference triangle* shown in Fig. 240.4. Again, changing to these coordinates does not affect the value of $(\nabla\varphi_i, \nabla\varphi_i)$ since $\nabla\varphi_i$ is constant on the triangle. On the triangle, φ_i can be written $\varphi_i = ax_1 + bx_2 + c$ for some constants a, b, c. Since $\varphi_i(0,0) = 1$, we get $c = 1$. Similarly, we compute a and b to find that $\varphi_i = 1 - x_1/h - x_2/h$ on this triangle. Therefore, $\nabla\varphi_i = \left(-h^{-1}, -h^{-1}\right)$ and the integral is

$$\int_{\triangleright} |\nabla\varphi_i|^2 \, dx = \int_0^h \int_0^{h-x_1} \frac{2}{h^2} \, dx_2 \, dx_1 = 1.$$

In the second case, φ_i is one at a node located at an acute angle of the triangle and is zero at the other nodes. We illustrate this in Fig. 240.5. We use the coordinate system shown in Fig. 240.5 to write $\varphi_i = 1 - x_1/h$.

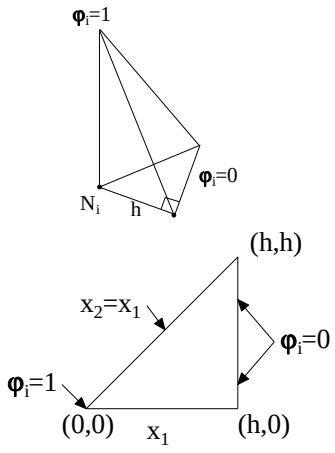


FIGURE 240.5. Second case showing φ_i and the reference triangle.

When we integrate over the triangle, we get $1/2$.

240.8. Verify this.

Summing the contributions from all the triangles gives

$$(\nabla\varphi_i, \nabla\varphi_i) = 1 + 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 4.$$

Next, we compute $(\nabla\varphi_i, \nabla\varphi_j)$ for indices corresponding to neighboring nodes. For a general node N_i, there are two cases of inner products (see Fig. 240.2 and Fig. 240.3):

$$(\nabla\varphi_i, \nabla\varphi_{i-1}) = (\nabla\varphi_i, \nabla\varphi_{i+1}) = (\nabla\varphi_i, \nabla\varphi_{i-m}) = (\nabla\varphi_i, \nabla\varphi_{i+m}),$$

and

$$(\nabla\varphi_i, \nabla\varphi_{i-m+1}) = (\nabla\varphi_i, \nabla\varphi_{i+m-1}).$$

The orientation of the triangles in each of the two cases are different, but the inner product of the gradients of the respective basis functions is not affected by the orientation. Note that the the equations corresponding to nodes next to the boundary are special, because the nodal values on the boundary are zero, see Fig. 240.1. For example, the equation corresponding to N_1 only involves N_1, N_2 and N_{m+1}.

For the first case, we next compute $(\nabla\varphi_i, \nabla\varphi_{i+1})$. Plotting the intersection of the respective supports shown in Fig. 240.6, we conclude that there are equal contributions from each of the two triangles in the intersection. We choose one of the triangles and construct a reference triangle as above. Choosing suitable variables, we find that

$$\nabla\varphi_i \cdot \nabla\varphi_{i+1} = \left(-\frac{1}{h}, -\frac{1}{h}\right) \cdot \left(\frac{1}{h}, 0\right) = -\frac{1}{h^2},$$

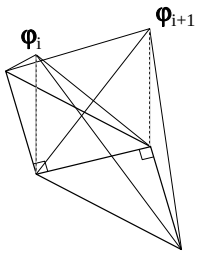


FIGURE 240.6. The overlap of φ_i and φ_{i+1}.

and integrating over the triangle gives $-1/2$.

240.9. Carry out this computation in detail.

Since there are two such triangles, we conclude that $(\nabla\varphi_i, \nabla\varphi_{i+1}) = -1$.

240.10. Prove that $(\nabla\varphi_i, \nabla\varphi_{i-m+1}) = (\nabla\varphi_i, \nabla\varphi_{i+m-1}) = 0$.

We can now determine the stiffness matrix A using the information above. We start by considering the first row. The first entry is $(\nabla\varphi_1, \nabla\varphi_1) = 4$ since N_1 has no neighbors to the left or below. The next entry is $(\nabla\varphi_1, \nabla\varphi_2) = -1$. The next entry after that is zero, because the supports of φ_1 and φ_3 do not overlap. This is true in fact of all the entries up to and including φ_m. However, $(\nabla\varphi_1, \nabla\varphi_{m+1}) = -1$, since these neighboring basis functions do share two supporting triangles. Finally, all the rest of the entries in that row are zero because the supports of the corresponding basis functions do not overlap. We continue in this fashion working row by row. The result is pictured in Fig. 240.7. We see that A has a *block structure* consisting of banded $m \times m$ submatrices, most of which consit only of zeros. Note the pattern of entries around corners of the diagonal block matrices; it is a common mistake to program these values incorrectly.

240.11. Compute the stiffness matrix for the Poisson equation with homogeneous Dirichlet boundary conditions for (a) the *union jack* triangulation of a square shown in Fig. 240.8 and (b) the triangulation of triangular domain shown in Fig. 240.8.

240.12. Compute the coefficients of the mass matrix M on the standard triangulation of the square of mesh size h. Hint: it is possible to use quadrature based on the midpoints of the sides of the triangle because this is exact for quadratic functions. The diagonal terms are $h^2/2$ and the off-diagonal terms are all equal to $h^2/12$. The sum of the elements in a row is equal to h^2.

240.13. Compute the stiffness matrix A for the continuous piecewise quadratic finite element method for the Poisson equation with homogeneous boundary conditions on the unit square using the standard triangulation.

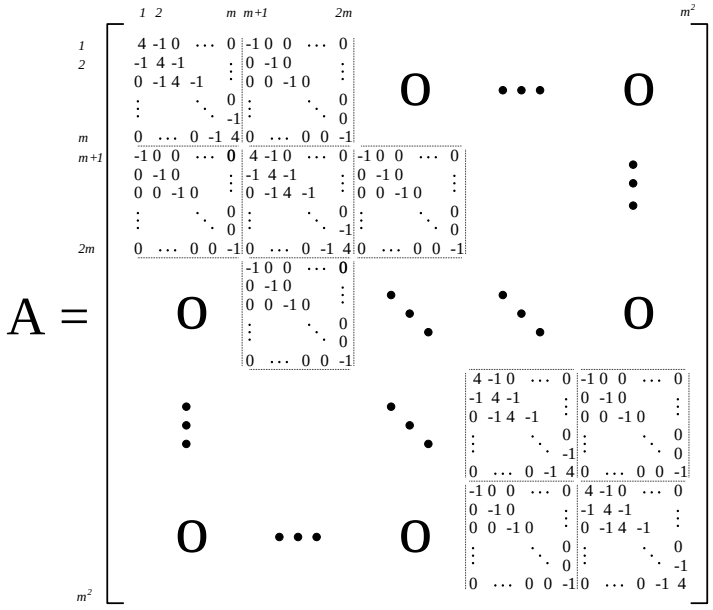


FIGURE 240.7. The stiffness matrix.

240.14. Compute the matrix $-\hat{M}^{-1}A$ on the standard triangulation, where $\hat{M}$ is the lumped mass matrix obtained computing the mass matrix using nodal quadrature. Give an interpretation of $-\hat{M}^{-1}A$ related to Δ_h.

The storage of a sparse matrix and the solution of a sparse system are both affected by the *structure* or *sparsity pattern* of the matrix. The sparsity pattern is affected in turn by the enumeration scheme used to mark the nodes.

240.15. Describe the sparsity pattern of the stiffness matrices A for the Poisson equation with homogeneous Dirichlet data on the unit square corresponding to the continuous piecewise linear finite element method on the standard triangulation using the three numbering schemes pictured in Fig. 240.9.

There are several algorithms for reordering the coefficients of a sparse matrix to form a matrix with a smaller bandwidth. Reordering the coefficients is equivalent to computing a new basis for the vector space.

The load vector b is computed in the same fashion, separating each integral

$$\int_\Omega f\varphi_i\,dx = \int_{\text{support of } \varphi_i} f(x)\varphi_i(x)\,dx$$

into integrals over the triangles making up the support of φ_i. To compute the elements (f,φ_i) of the load vector, we often use one of the quadrature formulas presented in Chapter **??**.

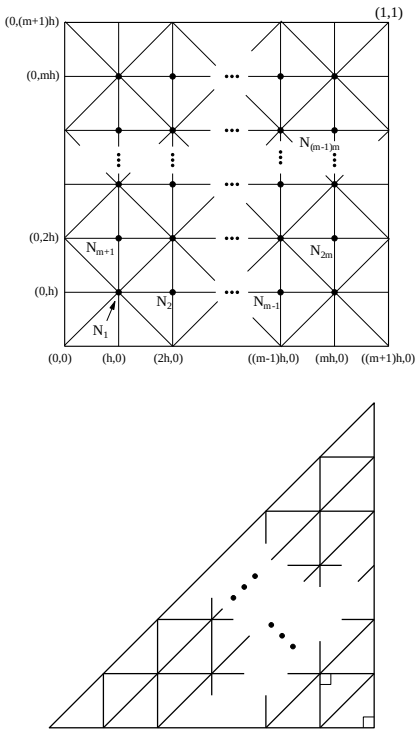


FIGURE 240.8. The "union jack" triangulation of the unit square and a uniform triangulation of a right triangle.

240.16. Compute the load vector b for $f(x) = x_1 + x_2^2$ on the standard triangulation of the unit square using exact integration and the lumped mass (trapezoidal rule) quadrature.

240.6 General remarks on computing the stiffness matrix and load vector

To compute the finite element approximation U, we have to compute the coefficients of the stiffness matrix A and load vector b and solve the linear system of equations (240.8). This is relatively easy to do on a uniform mesh, but it is a considerable programming problem in general because of the complexity of the geometry involved.

The first task is to compute the non-zero elements $a_{ij} = (\nabla\varphi_j, \nabla\varphi_i)$ of the stiffness matrix A. As we saw above, $a_{ij} = 0$ unless both N_i and N_j are nodes of the same triangle K because this is the only way that the support

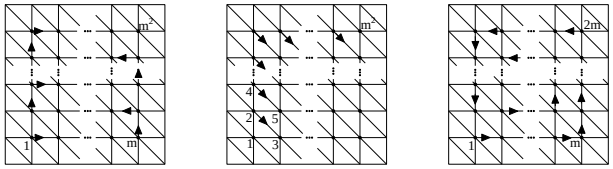


FIGURE 240.9. Three node numbering schemes for the standard triangulation of the unit square.

of different basis functions overlap. The common support corresponding to a non-zero coefficient is equal to the support of φ_j if $i = j$ and equal to the two triangles with the common edge connecting N_j and N_i if $i \neq j$. In each case a_{ij} is the sum of contributions

$$a_{ij}^K = \int_K \nabla\varphi_j \cdot \nabla\varphi_i \, dx \tag{240.13}$$

over the triangles K in the common support. The process of adding up the contributions a_{ij}^K from the relevant triangles K to get the a_{ij}, is called *assembling* the stiffness matrix A. Arranging for a given triangle K the numbers a_{ij}^K, where N_i and N_j are nodes of K, into a 3×3 matrix (renumbering locally the nodes 1, 2 and 3 in some order), we obtain the *element stiffness matrix* for the element K. We refer to the assembled matrix A as the *global stiffness matrix*. The element stiffness matrices were originally introduced as a way to organize the computation of A. They are also useful in iterative methods where the assembly (and storage) of A may be avoided completely, and the coefficients a_{ij} are assembled as they are required for the computation of discrete residuals.

240.17. (a) Show that the element stiffness matrix (240.13) for the linear polynomials on a triangle K with vertices at $(0,0)$, $(h,0)$, and $(0,h)$ numbered 1, 2 and 3, is given by

$$\begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1/2 & 0 \\ -1/2 & 0 & 1/2 \end{pmatrix}.$$

(b) Use this result to verify the formula computed for the stiffness matrix A for the continuous piecewise linear finite element method for the Poisson equation with homogeneous boundary conditions on the unit square using the standard triangulation. (c) Compute the element stiffness matrix for a triangle K with nodes $\{a^i\}$.

240.18. (a) Compute the element stiffness matrix for Poisson's equation for the quadratic polynomials on the reference triangle with vertices at $(0,0)$, $(h,0)$ and $(0,h)$. (b) Use the result to compute the corresponding global stiffness matrix for

the standard triangulation of the unit square assuming homogeneous boundary conditions; cf. Problem 240.13.

240.19. (a) Compute the element stiffness matrix A^K for the continuous bilinear finite element method for the Poisson equation with homogeneous boundary conditions on the unit square using a triangulation into small squares. (b) Use the result to compute the global stiffness matrix.

240.20. There are speculations that the coupling of two nodes N_i and N_j corresponding to a non-zero coefficient a_{ij} in the stiffness matrix $A = (a_{ij})$, is established through the exchange of particles referred to as *femions*. The nature of these hypothetical particles is unknown. It is conjectured that a femion has zero mass and charge but nevertheless a certain "stiffness". Give your opinion on this question.

240.7 Basic data structures

To compute the element stiffness matrices a_{ij}^K, we need the physical coordinates of the nodes of K, and to perform the assembly of A we need the global numbering of the nodes. Similar information is needed to compute the load vector. This information is arranged in a *data structure*, or data base, containing a list of the coordinates of the nodes and a list of the global numbers of the nodes of each triangles. Additional information such as a list of the neighboring elements of each element and a list of the edges, may also be needed for example in adaptive algorithms. It is desirable to organize the data structure so that mesh modification can be handled easily. We discuss this further in the advanced companion book.

240.8 Solving the discrete system

Once we have assembled the stiffness matrix, we solve the linear system $A\xi = b$ to obtain the finite element approximation. We discuss this briefly based on the material presented in Chapter 94. The stiffness matrix resulting from discretizing the Laplacian is symmetric and positive-definite and therefore invertible. These properties also mean that there is a wide choice in the methods used to solve the linear system for ξ, which take advantage of the fact that A is sparse.

In the case of the standard uniform discretization of a square, we saw that A is a banded matrix with five non-zero diagonals and bandwidth $m + 1$, where m is the number of nodes on a side. The dimension of A is m^2 and the asymptotic operations count for using a direct method to solve the system is $O(m^4) = O(h^{-4})$. Note that even though A has mostly zero diagonals inside the band, fill-in occurs as the elimination is performed, so we may as well treat A as if it has non-zero diagonals throughout the

band. Clever rearrangement of A to reduce the amount of fill-in leads to a solution algorithm with an operations count on the order of $O(m^3) = O(h^{-3})$. In contrast, if we treat A as a full matrix, we get an asymptotic operations count of $O\big(h^{-6}\big)$, which is considerably larger for a large number of elements.

240.21. Compute the asymptotic operations count for the direct solution of the system $A\xi = b$ using the three A computed in Problem 240.15.

240.22. Write a code to solve the system $A\xi = b$ that uses the band structure of A.

In general, we get a sparse stiffness matrix, though there may not be a band structure. If we want to use direct methods efficiently in general, then it is necessary to first reorder the system to bring the matrix into banded form.

We can also apply both the Jacobi and Gauss-Seidel methods to solve the linear system arising from discretizing the Poisson equation. In the case of the uniform standard discretization of a square for example, the operations count is $O\big(5M\big)$ per iteration for both methods if we make use of the sparsity of A. Therefore a single step of either method is much cheaper than a direct solve. The question is: How many iterations do we need to compute in order to obtain an accurate solution?

It is not to difficult to show that the spectral radius of the iteration matrix of the Jacobi method M_J is $\rho(M_J) = 1 - h^2\pi^2/2 + O(h^4)$, which means that the convergence rate is $R_J = h^2\pi^2/2 + O(h^4)$. The Gauss-Seidel method is more difficult to analyze, see Isaacson and Keller ([**?**]), but it can be shown that $\rho(M_{GS}) = 1 - h^2\pi^2 + O(h^4)$ yielding a convergence rate of $R_{GS} = h^2\pi^2 + O(h^4)$, which is twice the rate of the Jacobi method. Therefore, the Gauss-Seidel method is preferable to the Jacobi method. On the other hand, the convergence rate of either method decreases like h^2 so as we refine the mesh, both methods become very slow. The number of operations to achieve an error of $10^{-\sigma}$ is of order $5\sigma/(\pi^2 h^4)$. This is the same order as using a direct banded solver.

There has been a lot of activity in developing iterative methods that converge more quickly. For example, a classic approach is based on modifying M_{GS} in order to decrease the spectral radius, and the resulting method is called an accelerated or over-relaxed Gauss-Seidel iteration. In recent years, very efficient *multi-grid methods* have been developed and are now becoming a standard tool. A multi-grid method is based on a sequence of Gauss-Seidel or Jacobi steps performed on a hierarchy of successively coarser meshes and are optimal in the sense that the solution work is proportional to the total number of unknowns. We discuss multigrid methods in detail in the advanced companion volume.

240.23. Program codes to solve $A\xi = b$ using both the Jacobi and Gauss-Seidel iteration methods, making use of the sparsity of A in storage and operations.

Compare the convergence rate of the two methods using the result from a direct solver as a reference value.

240.9 Energy norm error estimates

In this section, we derive a priori and a posteriori error bounds in the *energy norm* for the finite element method for Poisson's equation with homogeneous Dirichlet boundary conditions. The energy norm, which is the L_2 norm of the gradient of a function in this problem, arises naturally in the error analysis of the finite element method because it is closely tied to the variational problem. The gradient of the solution, representing heat flow, electric field, flow velocity, or stress for example, can be a variable of physical interest as much as the solution itself, representing temperature, potential or displacement for example, and in this case, the energy norm is the relevant error measure. We also prove optimal order error estimates in the L_2 norm of the solution itself. We discuss analysis in other norms in the advanced companion book.

240.9.1 A priori error estimate

We first prove that the Galerkin finite element approximation is the best approximation of the true solution in V_h with respect to the energy norm.

Theorem 240.1 *Assume that u satisfies the Poisson equation (240.1) and U is the Galerkin finite element approximation satisfying (240.6). Then*

$$\|\nabla(u-U)\| \leq \|\nabla(u-v)\| \quad \text{for all } v \in V_h. \tag{240.14}$$

Proof: Using the Galerkin orthogonality (240.7) with $U - v \in V_h$, we can write

$$\|\nabla e\|^2 = (\nabla e, \nabla(u-U)) = (\nabla e, \nabla(u-U)) + (\nabla e, \nabla(U-v)).$$

Adding the terms involving U on the right, whereby U drops out, and using Cauchy's inequality, we get

$$\|\nabla e\|^2 = (\nabla e, \nabla(u-v)) \leq \|\nabla e\|\,\|\nabla(u-v)\|,$$

which proves the theorem after dividing by $\|\nabla e\|$. ■

Using the interpolation results of Theorem **??** choosing $v = \pi_h u$, we get the following concrete quantitative a priori error estimate:

Corollary 240.2 *There exists a constant C_i depending only on the minimal angle τ in $\mathcal{T}_h$, such that*

$$\|\nabla(u-U)\| \leq C_i \|hD^2 u\|. \tag{240.15}$$

240.10 A posteriori error estimate

We now prove an a posteriori error estimate following the strategy used for the two-point boundary value problem in Chapter **??**. A new feature occurring in higher dimensions is the appearance of integrals over the internal edges S in $\mathcal{S}_h$. We start by writing an equation for the error $e = u - U$ using (240.2) and (240.6) to get

$$\begin{aligned}\|\nabla e\|^2 &= (\nabla(u-U),\nabla e) = (\nabla u,\nabla e) - (\nabla U,\nabla e)\\ &= (f,e) - (\nabla U,\nabla e) = (f, e-\tilde{\pi}_h e) - (\nabla U,\nabla(e-\tilde{\pi}_h e)),\end{aligned}$$

where $\tilde{\pi}_h e \in V_h$ is an interpolant of e chosen as in (**??**). We may think of $\tilde{\pi}_h e$ as the usual nodal interpolant of e, although from a technical mathematical point of view, $\tilde{\pi}_h e$ will have to be defined slightly differently. We now break up the integrals over Ω into sums of integrals over the triangles K in $\mathcal{T}_h$ and integrate by parts over each triangle in the last term to get

$$\|\nabla e\|^2 = \sum_K \int_K (f+\Delta U)(e-\tilde{\pi}_h e)\,dx - \sum_K \int_{\partial K} \frac{\partial U}{\partial n_K}(e-\tilde{\pi}_h e)\,ds, \quad (240.16)$$

where $\partial U/\partial n_K$ denotes the derivative of U in the outward normal direction n_K of the boundary ∂K of K. In the boundary integral sum in (240.16), each internal edge $S \in \mathcal{S}_h$ occurs twice as a part of each of the boundaries ∂K of the two triangles K that have S as a common side. Of course the outward normals n_K from each of the two triangles K sharing S point in opposite directions. For each side S, we choose one of these normal directions and denote by $\partial_S v$ the derivative of a function v in that direction on S. We note that if $v \in V_h$, then in general $\partial_S v$ is different on the two triangles sharing S; see Fig. 233.8, which indicates the "kink" over S in the graph of v. We can express the sum of the boundary integrals in (240.16) as a sum of integrals over edges of the form

$$\int_S [\partial_S U](e-\tilde{\pi}_h e)\,ds,$$

where $[\partial_S U]$ is the difference, or jump, in the derivative $\partial_S U$ computed from the two triangles sharing S. The jump appears because the outward normal directions of the two triangles sharing S are opposite. We further note that $e-\tilde{\pi}_h e$ is continuous across S, but in general does not vanish on S, even if it does so at the end-points of S if $\tilde{\pi}_h$ is the nodal interpolant. This makes a difference with the one-dimensional case, where the corresponding sum over nodes does indeed vanish, because $e - \pi_h e$ vanishes at the nodes. We may thus rewrite (240.16) as follows with the second sum replaced by a sum over internal edges S:

$$\|\nabla e\|^2 = \sum_K \int_K (f+\Delta U)(e-\tilde{\pi}_h e)\,dx + \sum_{S\in\mathcal{S}_h} \int_S [\partial_S U](e-\tilde{\pi}_h e)\,ds.$$

Next, we return to a sum over element edges ∂K by just distributing each jump equally to the two triangles sharing it, to obtain an *error representation* of the energy norm of the error in terms of the residual error:

$$\|\nabla e\|^2 = \sum_K \int_K (f + \Delta U)(e - \tilde{\pi}_h e)\, dx \\ + \sum_K \frac{1}{2} \int_{\partial K} h_K^{-1} [\partial_S U](e - \tilde{\pi}_h e) h_K \, ds,$$

where we prepared to estimate the second sum by inserting a factor h_K and compensating. In crude terms, the residual error results from substituting U into the differential equation $-\Delta u - f = 0$, but in reality, straightforward substitution is not possible because U is not twice differentiable in Ω. The integral on the right over K is the remainder from substituting U into the differential equation inside each triangle K, while the integral over ∂K arises because $\partial_S U$ in general is different when computed from the two triangles sharing S.

We estimate the first term in the error representation by inserting a factor h, compensating and using the estimate $\|h^{-1}(e - \tilde{\pi}_h e)\| \leq C_i \|\nabla e\|$ of Theorem **??**, to obtain

$$|\sum_K \int_K h(f + \Delta U) h^{-1}(e - \tilde{\pi}_h e)\, dx| \\ \leq \|hR_1(U)\| \|h^{-1}(e - \tilde{\pi}_h e)\| \leq C_i \|hR_1(U)\| \|\nabla e\|,$$

where $R_1(U)$ is the function defined on Ω by setting $R_1(U) = |f + \Delta U|$ on each triangle $K \in \mathcal{T}_h$. We estimate the contribution from the jumps on the edges similarly. Formally, the estimate results from replacing $h_K \, ds$ by dx corresponding to replacing the integrals over element boundaries ∂K by integrals over elements K. Dividing by $\|\nabla e\|$, we obtain the following a posteriori error estimate:

Theorem 240.3 *There is an interpolation constant C_i only depending on the minimal angle τ such that the error of the Galerkin finite element approximation U of the solution u of the Poisson equation satisfies*

$$\|\nabla u - \nabla U\| \leq C_i \|hR(U)\|, \tag{240.17}$$

where $R(U) = R_1(U) + R_2(U)$ with

$$R_1(U) = |f + \Delta U| \quad \text{on } K \in \mathcal{T}_h,$$
$$R_2(U) = \frac{1}{2} \max_{S \subset \partial K} h_K^{-1} \big|[\partial_S U]\big| \quad \text{on } K \in \mathcal{T}_h.$$

As we mentioned, $R_1(U)$ is the contribution to the total residual from the interior of the elements K. Note that in the case of piecewise linear

approximation, $R_1(U) = |f|$. Further, $R_2(U)$ is the contribution to the residual from the jump of the normal derivative of U across edges. In the one dimensional problem considered in Chapter **??**, this contribution does not appear because the interpolation error may be chosen to be zero at the node points. We observe that the presence of the factor of h in front of the residual error $R(U)$ in (240.17) originates from the Galerkin orthogonality and the estimate $\|h^{-1}(e - \tilde{\pi}_h e)\| \leq C_i\|\nabla e\|$.

240.24. Derive a priori and a posteriori error bound in the energy norm for the finite element approximation of the solution of the Poisson equation in which the integrals involving the data f are approximated using the one point Gauss quadrature on each triangle or the "lumped mass" nodal quadrature. Hint: recall the modeling error estimate in Chapter **??**.

240.25. Give a more precise proof of the estimate for the jump term in Theorem 240.3 using Theorem **??** starting from the error representation.

240.26. Implement an "error estimation" routine for a code that approximates the Poisson problem using the continuous piecewise linear finite element method. Construct a test problem with a known solution by choosing a function $u(x)$ that is zero on the boundary of the unit square and setting $f = -\Delta u$, then compare the error estimate to the true error on a few meshes.

240.11 Adaptive error control

An immediate use of an a posteriori error bound is to estimate the error of a computed solution which gives important information to the user. We may also base an adaptive algorithm on the a posteriori error estimate seeking to optimize the computational work needed to reach a certain accuracy.

More precisely, we formulate the basic goal of adaptive error control as: for a given tolerance TOL, find a triangulation $\mathcal{T}_h$ that requires the least amount of computational work to achieve

$$\|\nabla u - \nabla U\| \leq \text{TOL}, \tag{240.18}$$

where $U \in V_h$ is the finite element approximation corresponding to $\mathcal{T}_h$. Measuring the computational work in terms of the number of nodes of the triangulation $\mathcal{T}_h$ and estimating the unknown error by the computable a posteriori error bound, we are led to the problem of finding the triangulation $\mathcal{T}_h$ with the least number of nodes such that the corresponding finite element approximation U satisfies the stopping criterion

$$C_i\|hR(U)\| \leq \text{TOL}. \tag{240.19}$$

This is a nonlinear constrained minimization problem with U depending on $\mathcal{T}_h$. If (240.17) is a reasonably sharp estimate of the error, then a solution of this optimization problem will meet our original goal.

We cannot expect to be able to solve this minimization problem analytically. Instead, a solution has to be sought by an iterative process in which we start with a coarse initial mesh and then successively modify the mesh by seeking to satisfy the stopping criterion (240.19) with a minimal number of elements. More precisely, we follow the following *adaptive algorithm*:

1. Choose an initial triangulation $\mathcal{T}_h^{(0)}$.
2. Given the j^{th} triangulation $\mathcal{T}_{h^{(j)}}$ with mesh function $h^{(j)}$, compute the corresponding finite element approximation $U^{(j)}$.
3. Compute the corresponding residuals $R_1(U^{(j)})$ and $R_2(U^{(j)})$ and check whether or not (240.19) holds. If it does, stop.
4. Find a new triangulation $\mathcal{T}_{h^{(j+1)}}$ with mesh function $h^{(j+1)}$ and with a minimal number of nodes such that $C_i\|h^{(j+1)}R(U^{(j)})\| \leq \text{TOL}$, and then proceed to #2.

The success of this iteration hinges on the mesh modification strategy used to perform step #4. A natural strategy for error control based on the L_2 norm uses the *principle of equidistribution* of the error in which we try to equalize the contribution from each element to the integral defining the L_2 norm. The rationale is that refining an element with large contribution to the error norm gives a large pay-off in terms of error reduction per new degree of freedom.

In other words, the approximation computed on the optimal mesh $\mathcal{T}_h$ in terms of computational work satisfies

$$\|\nabla e\|_{L_2(K)}^2 \approx \frac{\text{TOL}^2}{M} \quad \text{for all } K \in \mathcal{T}_h,$$

where M is the number of elements in $\mathcal{T}_h$. Based on (240.17), we would therefore like to compute the triangulation at step #4 so that

$$C_i^2\big(\big\|h^{(j+1)}R\big(U^{(j+1)}\big)\big\|\big)_{L_2(K)}^2 \approx \frac{\text{TOL}^2}{M^{(j+1)}} \quad \text{for all } K \in \mathcal{T}_{h^{(j+1)}}, \tag{240.20}$$

where $M^{(j+1)}$ is the number of elements in $\mathcal{T}_{h^{(j+1)}}$. However, (240.20) is a nonlinear equation, since we don't know $M^{(j+1)}$ and $U^{(j+1)}$ until we have chosen the triangulation. Hence, we replace (240.20) by

$$C_i^2\big(\big\|h^{(j+1)}R\big(U^{(j)}\big)\big\|\big)_{L_2(K)}^2 \approx \frac{\text{TOL}^2}{M^{(j)}} \quad \text{for all } K \in \mathcal{T}_h{}^{(j+1)}, \tag{240.21}$$

and use this formula to compute the new mesh size $h^{(j+1)}$.

There are several questions that need to be answered about the process described here, including: how much efficiency is lost by replacing (240.18) by (240.19)? In other words, how much bigger is the right-hand side of

(240.17) than the left-hand? Does the iterative process #1–#4 converge to a solution of the minimization problem? How should the initial triangulation $\mathcal{T}_{h^{(0)}}$ be chosen and how does this affect the convergence of the adaptive procedure? Is the approximation (240.21) justified? We address these issues in the advanced companion volume.

We conclude this section with an example that illustrates the behavior of this adaptive algorithm in a situation in which the forcing function is highly localized. We use Femlab to approximate the solution

$$u(x) = \frac{a}{\pi}\,\exp\bigl(-a(x_1^2 + x_2^2)\bigr), \quad a = 400,$$

of Poisson's equation $-\Delta u = f$ on the square $(-.5, .5) \times (-.5, .5)$ with $f(x)$ the following "approximate delta function":

$$f(x) = \frac{4}{\pi}\,a^2\,\bigl(1 - ax_1^2 - ax_2^2\bigr)\,\exp\bigl(-a(x_1^2 + x_2^2)\bigr),$$

We plot f in Fig. 240.10 (note the vertical scale), together with the initial mesh with 224 elements. The adaptive algorithm took 5 steps to achieve

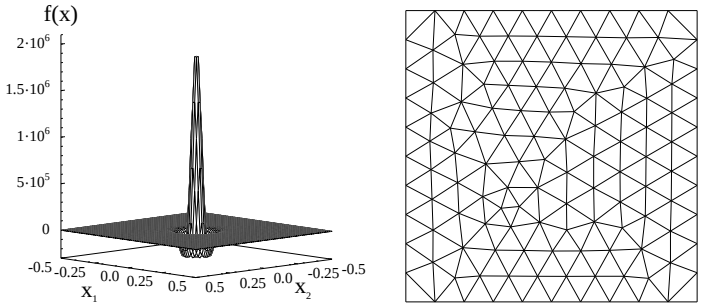


FIGURE 240.10. The approximate delta forcing function f and the initial mesh used for the finite element approximation.

an estimated .5% relative error. We plot the final mesh together with the associated finite element approximation in Fig. 240.11. The algorithm produced meshes with 224, 256, 336, 564, 992, and 3000 elements respectively.

240.27. Let $\omega(x)$ be a positive weight function defined on the domain $\Omega \subset \mathbb{R}^2$. Assume that the mesh function $h(x)$ minimizes the integral $\int_\Omega h^2(x)\omega(x)\,dx$ under the constraint $\int_\Omega h^{-1}(x)\,dx = N$, where N is a given positive integer. Prove that $h^3(x)\omega(x)$ is constant. Interpret the result as equidistribution in the context of error control. Hint: use the Lagrange multiplier method with the Lagrange function $L(h, \lambda) = \int_\Omega h^2(x)\omega\,dx + \lambda(\int_\Omega h^{-1}(x)\,dx - N)$.

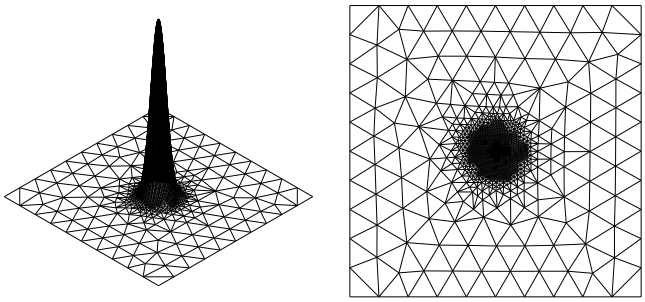


FIGURE 240.11. The finite element approximation with a relative error of .5% and the final mesh used to compute the approximation. The approximation has a maximum height of roughly 5.

240.12 Dealing with different boundary conditions

The variational problem has to be modified when the boundary conditions are changed from homogeneous Dirichlet conditions.

240.12.1 Non-homogeneous Dirichlet boundary conditions

We first discuss the Poisson's equation with non-homogeneous Dirichlet boundary conditions:

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = g & \text{on } \Gamma, \end{cases} \tag{240.22}$$

where g is the given boundary data. The variational formulation takes the following form: find $u \in V_g$, where

$$V_g = \left\{ v : \ v = g \text{ on } \Gamma \text{ and } \int_\Omega (|\nabla v|^2 + v^2) dx < \infty \right\},$$

such that

$$(\nabla u, \nabla v) = (f, v) \quad \text{for all } v \in V_0, \tag{240.23}$$

with

$$V_0 = \left\{ v : \ v = 0 \text{ on } \Gamma \text{ and } \int_\Omega (|\nabla v|^2 + v^2) dx < \infty \right\}.$$

Recall that V_g, where we look for u, is called the *trial space*, while V_0, from which we choose test functions, is called the *test space*. In this case, the trial and test spaces satisfy different boundary conditions, namely, the trial functions satisfy the given non-homogeneous Dirichlet condition $u = g$ on Γ while the test functions satisfy the homogeneous Dirichlet boundary condition. This is important in the construction of the variational formulation

(240.23) because when we multiply the differential equation by a test function $v \in V_0$ and use integration by parts, the boundary integral vanishes because $v = 0$ on Γ. The need to choose test functions satisfying homogeneous Dirichlet boundary conditions can also be understood by considering the minimization problem that is equivalent to (240.23): find $u \in V_g$ such that $F(u) \leq F(w)$ for all $w \in V_g$, where $F(w) = \frac{1}{2}(\nabla w, \nabla w) - (f, w)$. The variational formulation (240.23) results from setting the derivative $\frac{d}{d\epsilon}F(u + \epsilon v)$ equal to zero, where $v \in V_0$ is a perturbation satisfying zero boundary conditions so that $u + \epsilon v \in V_g$.

We compute a finite element approximation on a triangulation $\mathcal{T}_h$, where we now also include the nodes on the boundary, denoting the internal nodes by $\mathcal{N}_h$ as above and the set of nodes on the boundary by $\mathcal{N}_b$. We compute an approximation U of the form

$$U = \sum_{N_j \in \mathcal{N}_b} \xi_j \varphi_j + \sum_{N_j \in \mathcal{N}_h} \xi_j \varphi_j, \tag{240.24}$$

where φ_j denotes the basis function corresponding to node N_j in an enumeration $\{N_j\}$ of all the nodes, and, because of the boundary conditions, $\xi_j = g(N_j)$ for $N_j \in \mathcal{N}_b$. Thus the boundary values of U are given by g on Γ and only the coefficients of U corresponding to the interior nodes remain to be found. To this end, we substitute (240.24) into (240.2) and compute inner products with all the basis functions corresponding to the interior nodes, which yields a square system of linear equations for the unknown coefficients of U:

$$\sum_{N_j \in \mathcal{N}_h} \xi_j (\nabla \varphi_j, \nabla \varphi_i) = (f, \varphi_i) - \sum_{N_j \in \mathcal{N}_b} g(N_j)(\nabla \varphi_j, \nabla \varphi_i), \quad N_i \in \mathcal{N}_h.$$

Note that the terms corresponding to the boundary values of U become data on the right-hand side of the system.

240.28. Show that V_g is not a vector space. Prove that the solution of the weak problem is unique.

240.29. Compute the discrete equations for the finite element approximation for $-\Delta u = 1$ on $\Omega = (0,1) \times (0,1)$ with boundary conditions $u = 0$ for $x_1 = 0$, $u = x_1$ for $x_2 = 0$, $u = 1$ for $x_1 = 1$ and $u = x_1$ for $x_2 = 1$ using the standard triangulation (Fig. 240.1).

240.13 Laplace's equation on a wedge-shaped domain

We consider Laplace's equation with Dirichlet boundary conditions in a wedge-shaped domain making an angle ω:

$$\begin{cases} -\Delta u = 0 & \text{in } \Omega = \{(r,\theta) : 0 \le r < 1,\ 0 < \theta < \omega\} \\ u(r,0) = u(r,\omega) = 0, & 0 \le r < 1, \\ u(1,\theta) = \sin(\gamma\theta), & 0 \le \theta \le \omega, \end{cases} \tag{240.25}$$

where $\gamma = \pi/\omega$, see Fig. 240.12. The boundary conditions are chosen so

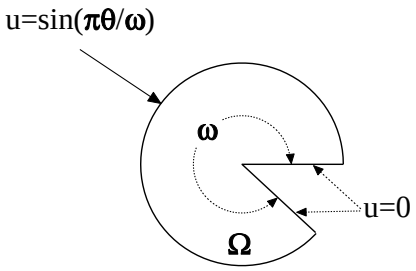


FIGURE 240.12. A domain with an interior corner.

that the exact solution u is given by the following simple explicit formula

$$u(r,\theta) = r^\gamma \sin(\gamma\theta). \tag{240.26}$$

Note that the solution satisfies homogenous Dirichlet boundary conditions on the straight sides joining the corner.

240.30. Verify the formula (240.26) by direct computation using the equation for the Laplacian in polar coordinates.

We noted in Section 239.6 that a solution of Laplace's equation in a domain (a harmonic function) is smooth inside the domain. We now show using the above example that a harmonic function may have a singularity at a corner of the boundary of the domain. Denoting the derivative with respect to r by D_r, we have from (240.26)

$$D_r u(r,\theta) = \gamma r^{\gamma-1}\sin(\gamma\theta), \quad D_r^2 u(r,\theta) = \gamma(\gamma-1)r^{\gamma-2}\sin(\gamma\theta),$$

and so on, which shows that sufficiently high derivatives of u become singular at $r = 0$, with the number depending on γ or ω. For example if $\omega = 3\pi/2$, then $u(r,\theta) \approx r^{2/3}$ and $D_r u(r,\theta) \approx r^{-1/3}$ with a singularity at

$r = 0$. The gradient ∇u corresponds to e.g. stresses in an elastic membrane or to an electric field. The analysis shows that these quantities become infinite at corners of angle $\omega > \pi$, which thus indicates extreme conditions at concave corners. If the boundary conditions change from Dirichlet to Neumann at the corner, then singularities may occur also at convex corners; see Problem 240.31.

More generally, a solution of Poisson's equation with smooth right hand side in a domain with corners, e.g. a polygonal domain, is a sum of terms of the form (240.26) plus a smooth function.

240.31. Solve the wedge problem with the Dirichlet condition replaced by a Neumann condition on one of the straight parts of the boundary.

240.14 An example: an L-shaped membrane

We present an example that shows the performance of the adaptive algorithm on a problem with a corner singularity. We consider the Laplace equation in an L-shaped domain that has a non-convex corner at the origin satisfying homogeneous Dirichlet boundary conditions at the sides meeting at the origin and non-homogeneous conditions on the other sides, see Fig. 240.13. We choose the boundary conditions so that the exact solution is $u(r,\theta) = r^{2/3}\sin(2\theta/3)$ in polar coordinates (r,θ) centered at the origin, which has the typical singularity of a corner problem. We use the knowledge of the exact solution to evaluate the performance of the adaptive algorithm.

We compute using FEniCS with energy norm control based on (240.17) to achieve an error tolerance of TOL $= .005$ using h refinement mesh modification. In Fig. 240.13, we show the initial mesh $\mathcal{T}_{h^{(0)}}$ with 112 nodes and 182 elements. In Fig. 240.14, we show the level curves of the solution and

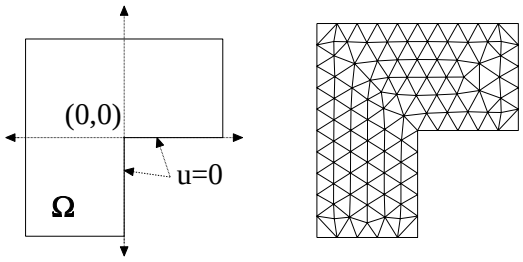


FIGURE 240.13. The L-shaped domain and the initial mesh.

the final mesh with 295 nodes and 538 elements that achieves the desired error bound. The interpolation constant was set to $C_i = 1/8$. The quotient between the estimated and true error on the final mesh was 1.5.

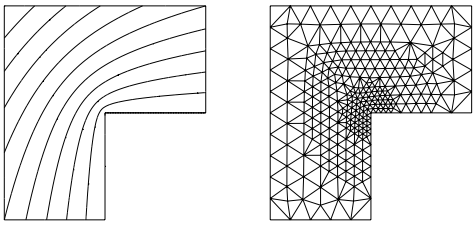


FIGURE 240.14. Level curves of the solution and final adapted mesh on the L-shaped domain.

Since the exact solution is known in this example, we can also use the a priori error estimate to determine a mesh that gives the desired accuracy. We do this by combining the a priori error estimate (240.15) and the principle of equidistribution of error to determine $h(r)$ so that $C_i\|hD^2u\| = \text{TOL}$ while keeping h as large as possible (and keeping the number of elements at a minimum). Since $D^2u(r) \approx r^{-4/3}$, as long as $h \leq r$, that is up to the elements touching the corner, we determine that

$$\left(hr^{-4/3}\right)^2 h^2 \approx \frac{\text{TOL}^2}{M} \qquad \text{or} \quad h^2 = \text{TOL}\, M^{-1/2} r^{4/3},$$

where M is the number of elements and h^2 measures the element area. To compute M from this relation, we note that $M \approx \int_\Omega h^{-2}\, dx$, since the number of elements per unit area is $O(h^{-2})$, which gives

$$M \approx M^{1/2}\text{TOL}^{-1} \int_\Omega r^{-4/3}\, dx.$$

Since the integral is convergent (prove this), it follows that $M \propto \text{TOL}^{-2}$, which implies that $h(r) \propto r^{1/3}\,\text{TOL}$. Note that the total number of unknowns, up to a constant, is the same as that required for a smooth solution without a singularity, namely TOL^{-2}. This depends on the very local nature of the singularity in the present case. In general, of course solutions with singularies may require a much larger number of elements than smooth solutions do.

240.32. Use Femlab to approximate the solution of the Poisson equation on the L-shaped domain using the stated boundary conditions. Start with a coarse triangulation and use a smallish error tolerance. Print out the final mesh and use a ruler to measure the value of h versus r roughly, and then plot the points on a log-log plot. Compute a line through the data and compare the slope of this to the relation $h \approx r^{1/3}\,\text{TOL}$ based on the a priori result.

240.15 Robin and Neumann boundary conditions

Next, we consider Poisson's equation with homogeneous Dirichlet conditions on part of the boundary and non-homogeneous Robin conditions on the remainder:

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_1, \\ \partial_n u + \kappa u = g & \text{on } \Gamma_2, \end{cases} \tag{240.27}$$

where $\Gamma = \Gamma_1 \cup \Gamma_2$ is a partition of Γ into two parts and $\kappa \geq 0$. The natural trial space is

$$V = \left\{ v : \ v = 0 \text{ on } \Gamma_1 \text{ and } \int_\Omega \left(|\nabla v|^2 + v^2 \right) dx < \infty \right\},$$

where the trial functions satisfy the homogeneous Dirichlet condition but the Robin condition is left out. The test space is equal to the trial space, because of the homogeneous Dirichlet condition.

To find the variational formulation, we multiply the Poisson equation by a test function $v \in V$, integrate over Ω, and use Green's formula to move derivatives from u to v:

$$\begin{aligned} (f, v) &= -\int_\Omega \Delta u \, v \, dx = \int_\Omega \nabla u \cdot \nabla v \, dx - \int_\Gamma \partial_n u v \, ds \\ &= \int_\Omega \nabla u \cdot \nabla v \, dx + \int_{\Gamma_2} \kappa u v \, ds - \int_{\Gamma_2} g v \, ds, \end{aligned}$$

where we use the boundary conditions to rewrite the boundary integral. We are thus led to the following variational formulation: find $u \in V$ such that

$$(\nabla u, \nabla v) + \int_{\Gamma_2} \kappa u v \, ds = (f, v) + \int_{\Gamma_2} g v \, ds \quad \text{for all } v \in V. \tag{240.28}$$

It is clear that a solution of (240.27) satisfies (240.28). Conversely, we show that a solution of (240.28) that has two continuous derivatives also satisfies the differential equation (240.27). We integrate (240.28) by parts using Green's formula to put all the derivatives onto u to get

$$-\int_\Omega \Delta u \, v \, dx + \int_\Gamma \partial_n u v \, ds + \int_{\Gamma_2} \kappa u v \, ds = \int_\Omega f v \, dx + \int_{\Gamma_2} g v \, ds \quad \text{for all } v \in V$$

or

$$\int_\Omega (-\Delta u - f) v \, dx + \int_{\Gamma_2} (\partial_n u + \lambda u - g) v \, ds = 0 \quad \text{for all } v \in V. \tag{240.29}$$

By first varying v inside Ω as above while keeping $v = 0$ on the whole of the boundary Γ, it follows that u solves the differential equation $-\Delta u = f$ in Ω. Thus (240.29) reduces to

$$\int_{\Gamma_2} (\partial_n u + \kappa u - g)v \, ds = 0 \quad \text{for all } v \in V.$$

The same argument works here; if $\partial_n u + \kappa u - g$ is non-zero, say positive, at some point of Γ, then it is positive in some small neighborhood of the point in Γ and choosing v to be a positive "hill" centered at the point and zero outside the neighborhood, gives a contradiction. Thus by varying v on Γ_2, we see that the Robin boundary condition $\partial_n u + \lambda u = g$ on Γ_2 must be satisfied (provided $\partial_n u + \kappa u - g$ is continuous).

We recall that boundary conditions like the Dirichlet condition that are enforced explicitely in the choice of the space V are called *essential boundary conditions*. Boundary conditions like the Robin condition that are implicitly contained in the weak formulation are called *natural boundary conditions*. (To remember that we must assume essential conditions: there are two "ss" in assume and essential.)

To discretize the Poisson equation with Robin boundary conditions on part of the boundary (240.27), we triangulate Ω as usual, but we number both the internal nodes and the nodes on Γ_2, where the Robin boundary conditions are posed. We do not number the nodes on Γ_1 where the homogeneous Dirichlet conditions are imposed. Nodes located where Γ_1 and Γ_2 meet should then be considered Dirichlet nodes. We then write U as in (240.24) with $\mathcal{N}_b$ denoting the nodes on Γ_2. In this problem, however, the coefficients of U corresponding to nodes in $\mathcal{N}_b$ are unknown. We substitute (240.24) into the weak form (240.28) and compute the inner products with all the basis functions corresponding to nodes in $\mathcal{N}_h \cup \mathcal{N}_b$ to get a square system. The boundary value g enters into the discrete equations as data on the right-hand side of the linear system for U.

Note that the stiffness matrix and load vector related to (240.28) contain contributions from both integrals over Ω and Γ_2 related to the basis functions corresponding to the nodes on the boundary Γ_2.

To illustrate, we compute the solution of Laplace's equation with a combination of Dirichlet, Neumann and Robin boundary conditions on the domain shown in Fig. 240.15 using Femlab. We show the boundary conditions in the illustration. The problem models e.g. stationary heat flow around a hot water pipe in the ground. We show the mesh that Femlab used to compute the approximation so that the error in the L_2 norm is smaller than .0013 together with a contour plot of the approximation in Fig. 240.16. We notice that the level curves are parallel to a boundary with a homogeneous Dirichlet condition, and orthogonal to a boundary with a homogeneous Neumann condition.

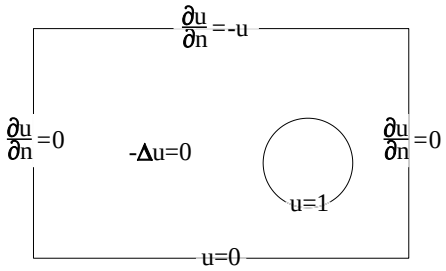


FIGURE 240.15. A problem with Robin boundary conditions.

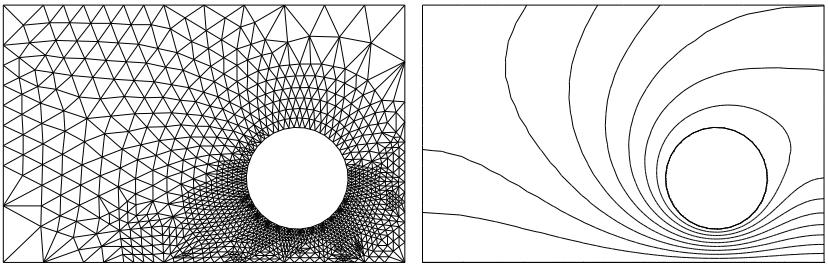


FIGURE 240.16. The adaptive mesh and contour lines of the approximate solution of the problem shown in Fig. 240.15 computed with error tolerance .0013.

240.33. Compute the discrete system of equations for the finite element approximation of the problem $-\Delta u = 1$ in $\Omega = (0,1) \times (0,1)$ with $u = 0$ on the side with $x_2 = 0$ and $\partial_n u + u = 1$ on the other three sides of Ω using the standard triangulation. Note the contribution to the stiffness matrix from the nodes on the boundary.

240.34. (a) Show that the variational formulation of the Neumann problem

$$\begin{cases} -\nabla \cdot (a\nabla u) + u = f & \text{in } \Omega, \\ a\partial_n u = g & \text{on } \Gamma, \end{cases} \tag{240.30}$$

where $a(x)$ is a positive coefficient, is to find $u \in V$ such that

$$\int_\Omega a\nabla u \cdot \nabla v \, dx + \int_\Omega uv \, dx = \int_\Omega fv \, dx + \int_\Gamma gv \, ds \quad \text{for all } v \in V, \tag{240.31}$$

where

$$V = \left\{ v : \int_\Omega a|\nabla v|^2 dx + \int_\Omega v^2 dx < \infty \right\}.$$

(b) Apply the finite element method to this problem and prove a priori and a posteriori error estimates. (c) Derive the discrete equations in the case of a uniform triangulation of a square and $a = 1$.

240.35. Apply the finite element method with piecewise linear approximation to the Poisson equation in three dimensions with a variety of boundary conditions. Compute the stiffness matrix and load vector in some simple case.

240.16 Error estimates in the L_2 norm

> Major scientific progress in different directions can only be gained through extended observation in a prolonged stay in a specific region, while observations during a balloon expedition cannot escape being of a superficial nature. (Nansen, in *Farthest North*, 1897)

So far in this chapter we have used the energy norm to measure the error. The main reason is that the energy norm arises naturally from the variational problem. However, it is often desirable to measure the error in different norms. In fact, specifying the quantities to be approximated and the norm in which to measure the error is a fundamentally important part of modeling, and directly affects the choice of approximation and error control algorithm.

As an example, we develop an error analysis in the L_2 norm. Actually, it is possible to derive an L_2 error estimate directly from the energy norm error estimates. In the two-point boundary value model problem (216.9) with $a = 1$, this follows by first expressing a function v defined on $[0, 1]$ and satisfying $v(0) = 0$ as the integral of its derivative:

$$v(y) = v(0) + \int_0^y v'(x)\,dx = \int_0^y v'(x)\,dx \quad \text{for } 0 \le y \le 1,$$

and then using Cauchy's inequality to estimate

$$|v(y)| \le \int_0^1 |v'(x)|\,dx \le \left(\int_0^1 1^2\,dx\right)^{1/2} \|v'\| = \|v'\|,$$

where $\|\cdot\|$ denotes the L_2 norm on $(0, 1)$. Squaring this inequality and integrating from 0 to 1 in x, we find

$$\|v\| \le \|v'\|.$$

Applying this estimate with $v = U - u$ and recalling the a priori energy norm error estimate 216.1, we thus obtain the following L_2 error estimate for the two-point boundary value problem (216.9) with $a = 1$:

$$\|u - U\| \le C_i \|hu''\|.$$

However, this estimate is not *optimal* because we expect the L_2 error of a good approximation of u in V_h, like for example the piecewise linear interpolant, to decrease quadratically in the mesh size h and not linearly

as in the estimate. We now improve the estimate and show that the error of the finite element approximation indeed is optimal in order with respect to the L_2 norm. This is remarkable, because it requires the error in the derivative to be "in average" better than first order.

240.36. Prove that if e is zero on the boundary of the unit square Ω, then

$$\left(\int_\Omega |e|^2\,dx\right)^{1/2} \leq \left(\int_\Omega |\nabla e|^2\,dx\right)^{1/2}.$$

Hint: extend the proof of the corresponding result in one dimension. Use the result to obtain an error estimate in the L_2-norm for the finite element method for Poisson's equation with homogeneous Dirichlet boundary conditions.

240.17 Error analysis based on duality

An approach to error analysis in a general norm is to use the idea of *duality* to compute the norm of a function by maximizing weighted average values, or inner products, of the function over a set of weights. For example,

$$\|u\|_{L_2(\Omega)} = \max_{v\in L_2(\Omega), v\neq 0} \frac{\int_\Omega u\,v\,dx}{\|v\|_{L_2(\Omega)}},$$

which follows because Cauchy's inequality shows that the right-hand side is bounded by the left-hand side, while choosing $v = u$ shows the equality. The fact that the norm of a function can be measured by computing a sufficient number of average values is both fundamentally important and widely applicable in a variety of situations. In fact, we already used this technique in the analysis of the parabolic model problem discussed in Chapter **??**, though without much background. We now give a more careful development.

We illustrate the idea behind a duality argument by first estimating the error of a numerical solution of a linear $n \times n$ system of equations $A\xi = b$. Recall that we discussed this previously in Chapter 94. We let $\bar{\xi}$ denote a numerical solution obtained for instance using an iterative method and estimate the Euclidean norm $|e|$ of the error $e = \xi - \bar{\xi}$. We start by posing the *dual problem* $A^\top \eta = e$, where e is considered to be the data. Of course, we don't know e but we will get around this. Using the dual problem, we get the following *error representation* by using the definition of the transpose,

$$|e|^2 = (e, A^\top\eta) = (Ae, \eta) = (A\xi - A\bar{\xi}, \eta) = (b - A\bar{\xi}, \eta) = (r, \eta)$$

where $r = b - A\bar{\xi}$ is the *residual error*. Suppose that it is possible to estimate the solution η of the equation $A^\top\eta = e$ in terms of the data e as

$$|\eta| \leq S|e|, \tag{240.32}$$

where S is a *stability factor*. It follows by Cauchy's inequality that

$$|e|^2 \le |r||\eta| \le S|r||e|,$$

or

$$|e| \le S|r|.$$

This is an a posteriori error estimate for the error e in terms of the residual r and the stability factor S.

We can guarantee that (240.32) holds by defining the stability factor by

$$S = \max_{\theta\in\mathbb{R}^n,\,\theta\neq 0} \frac{|\zeta|}{|\theta|}$$

where ζ solves $A^\top\zeta = \theta$.

The point of this example is to show how duality can be used to get an error representation in terms of the residual and the dual solution, from which the error can be estimated in terms of the residual and a stability factor. We use this approach repeatedly in this book, and also take advantage of the Galerkin orthogonality.

240.18 An a posteriori estimate for a two-point boundary value problem

We first prove an a posteriori error estimate in the $L_2(0,1)$ norm , denoted by $\|\cdot\|$, for the problem

$$\begin{cases} -(au')' + cu = f, & \text{in } (0,1), \\ u(0) = 0, \quad u(1) = 0, \end{cases} \tag{240.33}$$

where $a(x) > 0$ and $c(x) \ge 0$. We denote by U the cG(1) solution to the problem using the usual finite element space V_h of continuous piecewise linear functions. The *dual problem* takes just the same form as (240.33) because the given problem is symmetric:

$$\begin{cases} -(a\varphi')' + c\varphi = e, & \text{in } (0,1), \\ \varphi(0) = 0, \quad \varphi(1) = 0, \end{cases} \tag{240.34}$$

where $e = u - U$. We now use (240.33), (240.34), and the Galerkin orthogonality with the test function $v = \pi_h e \in V_h$, to obtain

$$\begin{aligned}\|e\|^2 &= \int_0^1 e(-(a\varphi')' + c\varphi)\,dx = \int_0^1 (ae'\varphi' + ce\varphi)\,dx \\ &= \int_0^1 (au'\varphi' + cu\varphi)\,dx - \int_0^1 (aU'\varphi' + cU\varphi)\,dx \\ &= \int_0^1 f\varphi\,dx - \int_0^1 (aU'\varphi' + cU\varphi)\,dx \\ &= \int_0^1 f(\varphi - \pi_h\varphi)\,dx - \sum_{j=1}^{M+1}\int_{I_j} (aU'(\varphi - \pi_h\varphi)' + cU(\varphi - \pi_h\varphi))\,dx.\end{aligned}$$

Integrating by parts over each sub-interval I_j, we find that all the boundary terms disappear, and we end up with

$$\|e\|^2 \le \|h^2 R(U)\|\|h^{-2}(\varphi - \pi_h\varphi)\|,$$

where $R(U) = f + (aU')' - cU$ on each sub-interval. Using an interpolation error estimate of the form $\|h^{-2}(\varphi - \pi_h\varphi)\| \le C_i\|\varphi''\|$, and defining the strong stability factor by

$$S = \max_{\xi \in L_2(I)} \frac{\|\varphi''\|}{\|\xi\|} \tag{240.35}$$

where φ satisfies (240.34) with e replaced by ξ, we obtain the following a posteriori error estimate:

Theorem 240.4 *The finite element approximation U of* (**??**) *satisfies*

$$\|u - U\| \le SC_i\|h^2 R(U)\|.$$

Note that the size of the stability factor S varies with the choice of the coefficients $a(x)$ and $c(x)$.

240.37. Prove that if $a > 0$ and $c \ge 0$ are constant, then $S \le a^{-1}$.

The implementation of an adaptive error control based on Theorem 240.34 is the same as for error control based on the energy norm. For an example, we choose $a = 0.01$, $c = 1$ and $f(x) = 1/x$ and compute using Femlab1d with the L_2 norm of the error bounded by $TOL = .01$. We plot the finite element approximation, the residual, and the mesh size in Fig. 240.17. In this example, there are two sources of singularities in the solution. First, because the diffusion coefficient a is small, the solution may become steep near the boundaries, forming what are called *boundary layers*. Secondly, the source term f itself is large near $x = 0$ and undefined at 0. The singularity in the data f affects the residual, while the size of a affects both the residual and the stability factor S. The adaptive algorithm approximates the stability factor S by solving the dual problem (240.34) with e replaced by an approximation. In this example, $S \approx 37$.

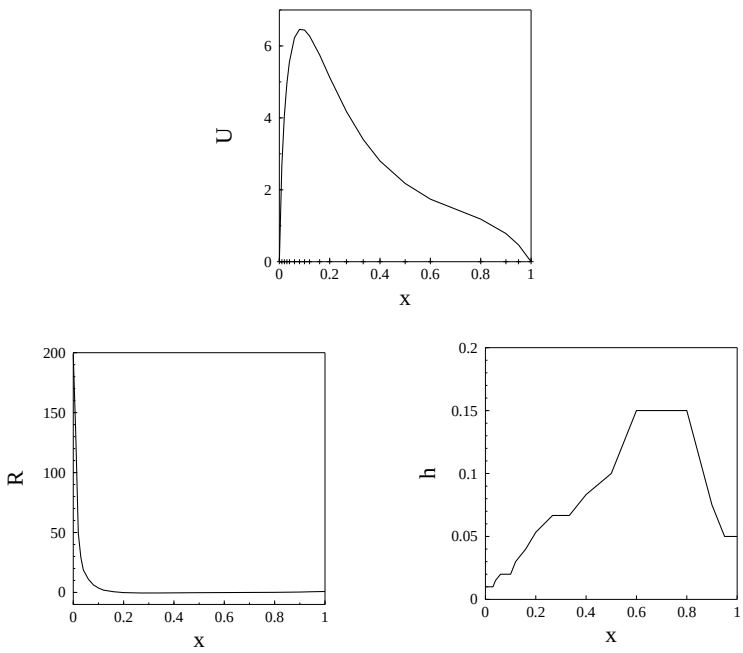


FIGURE 240.17. Finite element approximation, residual error, and meshsize computed with adaptive error control based on the L_2 norm.

240.19 A priori error estimate for a two-point boundary value problem

We also prove an a priori error estimate in the L_2 norm assuming for simplicity that the mesh size h is constant, and $c = 0$. Note the presence of the weighted norm $\|\cdot\|_a$.

Theorem 240.5 *The finite element approximation U of* (**??**) *satisfies*

$$\|u - U\| \le C_i S_a \|h(u - U)'\|_a \le C_i^2 S_a \|h^2 u''\|_a,$$

where $S_a = \max_{\xi \neq 0} \|\varphi''\|_a / \|\xi\|$ with φ satisfying (240.34) *with e replaced by ξ.*

Proof: Assuming that φ satisfies (240.34) with $c = 0$, and using the Galerkin orthogonality (**??**) and an L_2 estimate for the interpolation error, we obtain

$$\begin{aligned}\|e\|^2 &= \int_0^1 ae'\varphi'\,dx = \int_0^1 ae'(\varphi - \pi_h\varphi)'\,dx \\ &\le \|he'\|_a \|h^{-1}(\varphi - \pi_h\varphi)'\|_a \le C_i \|he'\|_a \|\varphi''\|_a,\end{aligned}$$

where $C_i = C_i(a)$. The proof is completed by using the definition of S_a and noting that multiplying the energy norm error estimate by h gives

$$\|he'\|_a \leq C_i \|h^2 u''\|_a. \tag{240.36}$$

■

This estimate generalizes to the case of variable h assuming that the mesh size h does not change too rapidly from one element to the next.

240.38. Prove that if $a > 0$ then $S_a \leq 1/\sqrt{a}$. Note that S and S_a involve somewhat different norms, which is compensated by the presence of the factor a in $R(U)$.

240.20 A priori and a posteriori error estimates for the Poisson equation

We now carry through the same program for the Poisson equation in two dimensions. We here assume that the mesh function $h(x)$ is differentiable and there is a constant $\tau_1 > 0$ such that $\tau_1 h_K \leq h(x) \leq h_K$ for $x \in K$ for each K in $\mathcal{T}_h$. This may be realized by smoothing of the original piecewise constant mesh function.

The proofs are based on a basic strong stability (or elliptic regularity) estimate for the solution of the Poisson equation (240.1) giving an estimate of the strong stability factor S. In Chapter **??**, we give the proof in the case of a convex domain with smooth boundary. In this case $S = 1$, and the stability estimate states that all second derivatives of a function u vanishing on the boundary of Ω can be bounded by the particular combination of second derivatives given by Δu.

Theorem 240.6 *If Ω is convex with polygonal boundary, or if Ω is a general domain with smooth boundary, then there is a constant S independent of f, such that the solution u of (240.1) satisfies*

$$\|D^2 u\| \leq S\|\Delta u\| = S\|f\|. \tag{240.37}$$

If Ω is convex, then $S = 1$.

The a priori error estimate is

Theorem 240.7 *Let Ω be convex with polygonal boundary or a general domain with smooth boundary. Then there exists a constant C_i only depending on τ and τ_1, such that the finite element approximation U of the Poisson problem (240.1) satisfies*

$$\|u - U\| \leq SC_i \|h\nabla(u - U)\|, \tag{240.38}$$

where S is defined in Theorem 240.6. Furthermore, if $|\nabla h(x)|_\infty \leq \mu$ for $x \in \Omega$ for some sufficiently small positive constant μ, then

$$\|h\nabla(u-U)\| \leq C_i\|h^2D^2u\|, \tag{240.39}$$

where C_i also depends on μ. In particular, if Ω is convex then

$$\|u-U\| \leq C_i\|h^2D^2u\|. \tag{240.40}$$

Proof: Letting φ solve the dual problem $-\Delta\varphi = e$ in Ω together with $\varphi = 0$ on Γ, we obtain by integration by parts, using the Galerkin orthogonality and the interpolation estimate Theorem **??**

$$\begin{aligned}\|e\|^2 &= (u-U,u-U) = (\nabla(u-U),\nabla\varphi)\\ &= (\nabla(u-U),\nabla(\varphi-\pi_h\varphi)) \leq C_i\|h\nabla(u-U)\|\|D^2\varphi\|,\end{aligned}$$

from which the first estimate follows using the strong stability result. The second estimate (240.39) follows directly from the energy norm error estimate if h is constant and we discuss the general result in the advanced companion volume. The final result (240.40) is obtained using the regularity estimate (240.37). ■

The a posteriori error estimate is

Theorem 240.8 *There are constants C_i and S such that, if U is the finite element approximation of (240.1), then with the residual R defined as in Theorem 240.3,*

$$\|u-U\| \leq SC_i\|h^2R(U)\|. \tag{240.41}$$

If Ω is convex, then $S = 1$.

Proof: With φ defined as in the previous proof, we have

$$\begin{aligned}\|e\|^2 &= (\nabla(u-U),\nabla\varphi) = (f,\varphi) - (\nabla U,\nabla\varphi)\\ &= (f,\varphi-\pi_h\varphi) - (\nabla U,\nabla(\varphi-\pi_h\varphi)).\end{aligned}$$

The desired result follows by an argument similar to that used in the a posteriori energy norm estimate by estimating $\|h^{-2}(\varphi-\pi_h\varphi)\|$ in terms of $C_i\|D^2\varphi\|$ and using the strong stability estimate to close the loop. ■

> It is like an attempt, over and over again, to reveal the heart of things. (K. Jarret)

> A poem should be equal to:
> Not true ...
> A poem should not mean
> But be. (Archibald MacLeish)

241
The Power of Abstraction

> Maybe in order to understand mankind, we have to look at the word itself. *Mankind*. Basically, it's made up of two separate words - "mank" and "ind". What do these word mean? It's a mystery, and that's why, so is mankind. (Jack Handley)

> The use of mathematical symbolism eliminates the waste of mental energy on trivialities, and liberates this energy for deployment where it is needed, to wit, on the chaotic frontiers of theory and practice. It also facilitates reasoning where it is easy, and restrains it where it is complicated. (Whitehead)

241.1 Introduction

Up until now we have considered a set of specific examples spanning the fundamental models in science. In this chapter, we consider an "abstract" linear elliptic problem, concentrating on the basic questions of existence, uniqueness, and stability of solutions together with the basic approximation properties of the Galerkin method. After that, we apply the abstract theory to specific problems including Poisson's equation with various boundary conditions, a model of linear elasticity, and Stoke's equations for creeping fluid flow. The abstract framework we describe is the result of a long development of variational methods initiated by Euler and Lagrange, continued by Dirichlet, Riemann, Hilbert, Rayleigh, Ritz, Galerkin, and continuing at the present time partly because of the modern interest in the finite el-

ement method. The advantage of considering a problem in abstract form is that we can emphasize the essential ingredients and moreover we can apply results for the abstract problem to specific applications as soon as the assumptions of the abstract problem are satisfied without having go through the same type of argument over and over again. This is the real "power" of abstraction. We focus on linear elliptic problems, since setting up an abstract framework is easist in this case. The framework may be extended naturally to a class of nonlinear elliptic problems related to convex minimization problem and to the related parabolic problems. An abstract framework for hyperbolic problems is less developed; see the advanced companion book for details. We keep the presentation in this chapter short, and give more details in the advanced companion volume. The idea is to indicate a framework, not to develop it in detail.

We recall from Chapters **??** and **??** that we started by rewriting a given boundary value problem in variational form. We then applied Galerkin's method to compute an approximate solution in a subspace of piecewise polynomials and we proved energy norm error estimates using the Galerkin orthogonality. The abstract elliptic problem we consider is formulated in variational terms and has stability and continuity properties directly related to the energy norm. The basic theorem on the the existence, uniqueness, and stability of the solution of the abstract elliptic problem is the *Lax-Milgram theorem*. We also give a related result stating that Galerkin's method is optimal in the energy norm. These results guarantee that some of the basic models of science including Poisson's equation and the equations for linear elasticity and Stokes flow have a satisfactory mathematical form and may be solved approximately using Galerkin's method. This is a cornerstone in science and engineering.

241.2 The abstract formulation

The ingredients of the abstract formulation are

(i) a Hilbert space V where we look for the solution, with norm $\|\cdot\|_V$ and scalar product $(\cdot,\cdot)_V$,

(ii) a bilinear form $a : V \times V \to \mathbb{R}$ that is determined by the underlying differential equation,

(iii) a linear form $L : V \to \mathbb{R}$ that is determined by the data.

A *Hilbert space* is a vector space with a scalar product that is *complete*, which means that any Cauchy sequence in the space converges to a limit in the space. Recall that we discussed the importance of using a space with this property in Chapter 199, where we used the completeness of the continuous functions on an interval to prove the Fundamental Theorem of

Calculus. A *bilinear form* $a(\cdot,\cdot)$ is a function that takes $V \times V$ into the real numbers, i.e. $a(v,w) \in \mathbb{R}$ for all $v, w \in V$, such that $a(v,w)$ is linear in each argument v and w, that is $a(\alpha_1 v_1 + \alpha_2 v_2, w_1) = \alpha_1 a(v_1, w_1) + \alpha_2 a(v_2, w_1)$ and $a(v_1, \alpha_1 w_1 + \alpha_2 w_2) = \alpha_1 a(v_1, w_1) + \alpha_2 a(v_1, w_2)$ for $\alpha_i \in \mathbb{R}$, v_i, $w_i \in V$. Finally, a *linear form* $L(\cdot)$ is a function on V such that $L(v) \in \mathbb{R}$ for all $v \in V$ and $L(v)$ is linear in v.

The abstract problem reads: find $u \in V$ such that

$$a(u,v) = L(v) \quad \text{for all } v \in V. \tag{241.1}$$

We make some assumptions on $a(\cdot,\cdot)$ and $L(\cdot)$, which gives an abstract definition of a linear "elliptic" problem. We first assume that $a(\cdot,\cdot)$ is *V-elliptic* or *coercive*, which means that there is a positive constant κ_1 such that for all $v \in V$,

$$a(v,v) \geq \kappa_1 \|v\|_V^2. \tag{241.2}$$

We also require that $a(\cdot,\cdot)$ is *continuous* in the sense that there is a constant κ_2 such that for all $v, w \in V$

$$|a(v,w)| \leq \kappa_2 \|v\|_V \|w\|_V. \tag{241.3}$$

We finally require that the linear form $L(\cdot)$ is *continuous* in the sense that there is a constant κ_3 such that for all $v \in V$,

$$|L(v)| \leq \kappa_3 \|v\|_V. \tag{241.4}$$

The reason that we say that L is continuous if (241.4) holds is because by linearity $|L(v) - L(w)| \leq \kappa_3 \|v - w\|_V$, which shows that $L(v) \to L(w)$ if $\|v - w\|_V \to 0$, i.e., if $v \to w$ in V. Assumption (241.3) similarly implies that $a(\cdot,\cdot)$ is continuous in each variable. Further, we define the *energy norm* $\|\cdot\|_a$ by $\|v\|_a = \sqrt{a(v,v)}$, noting that (241.2) in particular guarantees that $a(v,v) \geq 0$. By (241.2) and (241.3), we have $\kappa_1 \|v\|_V^2 \leq \|v\|_a^2 \leq \kappa_2 \|v\|_V^2$. In other words, if a quantity is small in the energy norm $\|\cdot\|_a$ then it is small in the norm $\|\cdot\|_V$ and vica versa. We refer to this situation by saying that $\|\cdot\|_a$ and $\|\cdot\|_V$ are *equivalent norms*. Thus, without changing anything qualitatively, we could choose the norm in the Hibert space V to be the energy norm $\|\cdot\|_a$ related to the bilinear form a, in which case $\kappa_1 = \kappa_2 = 1$. In this sense, the energy norm is a natural choice to use to analyze the bilinear form a. In applications, the energy norm fits with the notion of energy in mechanics and physics.

241.1. Determine a and L for (216.9), (**??**), and (240.1).

241.3 The Lax-Milgram theorem

We now state and prove the basic Lax-Milgram theorem.

Theorem 241.1 *Suppose $a(\cdot,\cdot)$ is a continuous, V-elliptic bilinear form on the Hilbert space V and L is a continuous linear functional on V. Then there is a unique element $u \in V$ satisfying (241.1). Moreover, the following stability estimate holds*

$$\|u\|_V \leq \frac{\kappa_3}{\kappa_1}. \tag{241.5}$$

Recall that the bilinear forms a associated to the two-point boundary value problem (216.9) and to Poisson's equation (240.1) are *symmetric*, i.e.

$$a(v,w) = a(w,v) \quad \text{for all } v, w \in V.$$

Symmetric problems have additional structure that make the proof of the Lax-Milgram theorem easier, and this is the case we consider now. We treat the non-symmetric case in the companion volume, see also Renardy and Rogers ([**?**]).

If a is symmetric, then the variational problem (241.1) is equivalent to the minimization problem: find $u \in V$ such that

$$F(u) \leq F(v) \quad \text{for all } v \in V, \tag{241.6}$$

where $F(v) = a(v,v)/2 - L(v)$. We state and prove this equivalence in the following theorem.

Theorem 241.2 *An element $u \in V$ satisfies (241.1) if and only if u satisfies (241.6).*

Proof: Assume first that $u \in V$ satisfies (241.6). Choose $v \in V$ and consider the function $g(\epsilon) = F(u + \epsilon v)$ for $\epsilon \in \mathbb{R}$. By (241.6) we know that $g(\epsilon) \geq g(0)$ for $\epsilon \in \mathbb{R}$, so that $g'(0) = 0$ if $g'(0)$ exists. But, differentiating the expression $g(\epsilon) = (a(u,u) + \epsilon(a(u,v) + a(v,u)) + \epsilon^2 a(v,v))/2 - L(u) - \epsilon L(v)$ with respect to ϵ and setting $\epsilon = 0$, gives $a(u,v) - L(v) = 0$, and (241.1) follows. Note that the symmetry of $a(\cdot,\cdot)$ is crucial to this argument.

Conversely, if (241.1) is satisfied, then for all $w \in V$,

$$\begin{aligned} F(u+w) &= \frac{1}{2}a(u,u) + a(u,w) + \frac{1}{2}a(w,w) - L(u) - L(w) \\ &= F(u) + \frac{1}{2}a(w,w) \geq F(u), \end{aligned}$$

with equality only if $w = 0$, which proves (241.6). ■

We now prove the Lax-Milgram theorem for symmetric $a(\cdot,\cdot)$ by using the equivalence of (241.1) and (241.6).

Proof: Since we have assumed that the energy norm and the norm of V are equivalent in (241.2) and (241.3), without loss of generality, we can take $(\cdot,\cdot)_V$ to be $a(\cdot,\cdot)$, so that $a(v,v) = \|v\|_V^2$, and $\kappa_1 = \kappa_2 = 1$.

We consider the set of real numbers that can be obtained as the limit of sequences $\{F(u_j)\}$ with $u_j \in V$. We observe that this set is bounded below

by $-1/2$ since $F(v) \geq \|v\|_V^2/2 - \|v\|_V \geq -1/2$ for all $v \in V$. We claim that the set of limits of $\{F(u_j)\}$ contains a smallest real number and we denote this number by β. Clearly, $\beta \geq -1/2$ if β exists. Now, the existence of β follows from the basic property of the real numbers that a set of real numbers that is bounded below has a greatest lower bound. In other words, there is a largest real number that is smaller or equal to all numbers in the set, which in our case is the number β. This property is equivalent to the property of convergence of a Cauchy sequence of real numbers. As another example, the set of positive real numbers ξ such that $\xi^2 > 2$ is clearly bounded below and its largest lower bound is nothing but $\sqrt{2}$. See Rudin ([**?**]) for more details.

Accepting the existence of β, we also know that $\beta \leq F(0) = 0$, and thus $-1/2 \leq \beta \leq 0$. Now let $\{u_j\}$ be a *minimizing sequence* for the minimization problem (241.6), i.e. a sequence such that

$$F(u_j) \to \beta \quad \text{as } j \to \infty. \tag{241.7}$$

We prove that $\{u_j\}$ is a *Cauchy sequence* in the sense that for any given $\epsilon > 0$ there is a natural number N_ϵ such that

$$\|u_i - u_j\|_V < \epsilon \quad \text{if } i, j \geq N_\epsilon. \tag{241.8}$$

Since V is complete, there is a $u \in V$ such that $\|u - u_j\| \to 0$ as $j \to \infty$. By the continuity properties of F, it follows that $F(u) = \beta$. and thus $u \in V$ is a solution of (241.6) and therefore (241.1). The uniqueness follows from the last inequality of the proof of Theorem 241.2 above.

To prove that the minimizing sequence is a Cauchy sequence, we note that (241.7) implies that for any $\epsilon > 0$ there is a N_ϵ such that

$$F(u_j) \leq \beta + \frac{\epsilon^2}{8} \quad \text{if } j \geq N_\epsilon. \tag{241.9}$$

We use the parallelogram law

$$\|u_i - u_j\|_V^2 = 2\|u_i\|_V^2 + 2\|u_j\|_V^2 - \|u_i + u_j\|_V^2,$$

together with the definition of $F(v)$, the definition of β, and (241.9), to argue

$$\begin{aligned} \frac{1}{4}\|u_i - u_j\|_V^2 &= F(u_i) + F(u_j) - 2F\left(\frac{1}{2}(u_i + u_j)\right) \\ &\leq F(u_i) + F(u_j) - 2\beta \leq \frac{\epsilon^2}{4}, \end{aligned}$$

proving (241.8).

Finally, the stability estimate follows immediately after taking $v = u$ in (241.1) and using the V-ellipticity and the continuity of L. ■.

241.4 The abstract Galerkin method

We consider Galerkin's method in abstract form applied to the problem (241.1): given a finite dimensional space $V_h \subset V$, find $U \in V_h$ such that

$$a(U, v) = L(v) \quad \text{for all } v \in V_h. \tag{241.10}$$

This leads to a linear system of equations whose size is determined by the dimension of V_h. We could for example choose V_h to be the space of polynomials of a fixed degree or less, the space of trigonometric functions with integer frequencies up to a fixed maximum, or in the case of the finite element method, the space of piecewise polynomial functions. We note that since $V_h \subset V$, we have the familiar Galerkin orthogonality:

$$a(u - U, v) = 0 \quad \text{for all } v \in V_h. \tag{241.11}$$

The basic a priori error estimate reads:

Theorem 241.3 *If u and U satisfy (241.1) and (241.10) then for all $v \in V_h$,*

$$\|u - U\|_V \leq \frac{\kappa_2}{\kappa_1}\|u - v\|_V.$$

If the norm $\|\cdot\|_V$ is equal to the energy norm $\|\cdot\|_a$, then

$$\|u - U\|_a \leq \|u - v\|_a, \tag{241.12}$$

which expresses the optimality of Galerkin's method in the energy norm.

Proof: The V-ellipticity and continuity of a together with Galerkin orthogonality implies that for all $v \in V_h$,

$$\begin{aligned}\kappa_1\|u - U\|_V^2 &\leq a(u - U, u - U) = a(u - U, u - U) + a(u - U, U - v)\\ &= a(u - U, u - v) \leq \kappa_2\|u - U\|_V\|u - v\|_V,\end{aligned}$$

which proves the desired result. ■

241.2. Prove (241.12). Prove that the solution U of (241.10) satisfies the following analog of (241.5): $\|U\|_V \leq \kappa_3/\kappa_1$.

241.5 Applications

We now present some basic applications of the Lax-Milgram theorem. In each case, we need to specify a, L and V and show that the assumptions of the Lax-Milgram theorem are satsified. Usually, the main issue is to verify the V-ellipticity of the bilinear form a. We illustrate some tools for this purpose in a series of examples.

241.5.1 A problem with Neumann boundary conditions

As a first example, we consider Poisson's equation with an absorption term together with Neumann boundary conditions as given in Problem 240.34,

$$\begin{cases} -\Delta u + u = f & \text{in } \Omega, \\ \partial_n u = 0 & \text{on } \Gamma, \end{cases} \tag{241.13}$$

where Ω is a bounded domain in $\mathbb{R}^d$ with boundary Γ. This problem takes the variational form (241.1) with

$$a(v,w) = \int_\Omega \big(\nabla v \cdot \nabla w + vw\big)\, dx, \quad L(v) = \int_\Omega fv\, dx, \tag{241.14}$$

and

$$V = \left\{ v : \int_\Omega \big(|\nabla v|^2 + v^2\big)\, dx < \infty \right\}. \tag{241.15}$$

The issue is to verify that the assumptions of the Lax-Milgram theorem are satisfied with these choices.

Clearly, V has natural scalar product and norm

$$(v,w)_V = \int_\Omega \big(\nabla v \cdot \nabla w + vw\big)\, dx, \quad \|v\|_V = \left(\int_\Omega \big(|\nabla v|^2 + v^2\big)\, dx \right)^{1/2}.$$

It turns out that V is complete, a fact ultimately based the completeness of $\mathbb{R}$, and therefore V is a Hilbert space. Further, we note that (241.2) and (241.3) trivially hold with $\kappa_1 = \kappa_2 = 1$. Finally, to show (241.4), we note that

$$|L(v)| \le \|f\|_{L_2(\Omega)} \|u\|_{L_2(\Omega)} \le \|f\|_{L_2(\Omega)} \|u\|_V,$$

which means that we may take $\kappa_3 = \|f\|_{L_2(\Omega)}$ provided we assume that $f \in L_2(\Omega)$. We conclude that the Lax-Milgram theorem applies to (241.13).

241.5.2 The spaces $H^1(\Omega)$ and $H^1_0(\Omega)$

The space V defined in (241.15) naturally occurs in variational formulations of second order elliptic differential equations and it has a special notation:

$$H^1(\Omega) = \left\{ v : \int_\Omega \big(|\nabla v|^2 + v^2\big)\, dx < \infty \right\}, \tag{241.16}$$

while the scalar product and norm are denoted by

$$(v,w)_{H^1(\Omega)} = \int_\Omega \big(\nabla v \cdot \nabla w + vw\big)\, dx,$$

and the associated norm

$$\|v\|_{H^1(\Omega)} = \big(\int_\Omega \big(|\nabla v|^2 + v^2\big)\, dx \big)^{1/2}.$$

The space $H^1(\Omega)$ is the *Sobolev space* of functions on Ω that are square integrable together with their gradients, named after the Russian mathematician Sobolev (1908-1994). The index one refers to the fact that we require first derivatives to be square integrable.

We also use the subspace $H_0^1(\Omega)$ of $H^1(\Omega)$ consisting of the functions in $H^1(\Omega)$ that vanish on the boundary Γ of Ω:

$$H_0^1(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma\}.$$

We motivate below why this is a Hilbert space with the same norm and scalar product as $H^1(\Omega)$.

241.3. (a) Find r such that $x^r \in H^1(0,1)$ but $x^s \notin H^1(0,1)$ for any $s < r$. (b) With $\Omega = \{x : |x| \leq 1\}$ denoting the unit disk, find conditions on r such that $|x|^r \in H^1(\Omega)$ but $|x|^s \notin H^1(\Omega)$ for any $s < r$.

241.4. Define $H^2(\Omega)$ and find a function that is in $H^1(\Omega)$ but not in $H^2(\Omega)$ where Ω is the unit disk.

241.5.3 Poisson's equation with Dirichlet boundary conditions

The first elliptic problem in several dimensions we studied was Poisson's equation with homogeneous Dirichlet boundary conditions posed on a bounded domain $\Omega \subset \mathbb{R}^2$ with boundary Γ:

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma. \end{cases}$$

This problem has the variational formulation (241.1) with $V = H_0^1(\Omega)$ and

$$a(v,w) = \int_\Omega \nabla v \cdot \nabla w \, dx, \quad L(v) = \int_\Omega f v \, dx.$$

In this case the V-ellipticity of a does not follow automatically from the definition of the norm in $V = H_0^1(\Omega)$ as above, because the bilinear form $a(v,v)$ in this case does not contain the term $\int_\Omega v^2 \, dx$ contained in the squared V norm. Further, we need to show that it makes sense to impose the boundary condition $v = 0$ on Γ for functions v in $V = H^1(\Omega)$, which is the essential issue in proving that $H_0^1(\Omega)$ is a Hilbert space.

To verify the V-ellipticity we use the *Poincaré-Friedrichs inequality*, which states that the $L_2(\Omega)$-norm of a function $v \in H^1(\Omega)$ can be estimated in terms of the $L_2(\Omega)$-norm of the gradient ∇v plus the $L_2(\Gamma)$-norm of the restriction of v to the boundary Γ. The corresponding theorem in one dimension for an interval $(0,1)$ states that

$$\|v\|_{L_2(0,1)}^2 \leq 2\big(v(0)^2 + \|v'\|_{L_2(0,1)}^2\big). \tag{241.17}$$

This inequality is proved easily by integrating the inequality

$$v^2(x) = \big(v(0) + \int_0^x v'(y)\,dy\big)^2 \le 2\big(v^2(0) + \int_0^1 (v'(y))^2\,dy\big)$$

for $0 \le x \le 1$, which is obtained by using Cauchy's inequality and the fact that $(a+b)^2 \le 2(a^2+b^2)$. The result for higher dimensions is

Theorem 241.4 *There is a constant C depending on Ω such that for all $v \in H^1(\Omega)$,*

$$\|v\|^2_{L_2(\Omega)} \le C\big(\|v\|^2_{L_2(\Gamma)} + \|\nabla v\|^2_{L_2(\Omega)}\big). \tag{241.18}$$

241.5. (a) Prove (241.18). Hint: Take $\varphi = |x|^2/(2d)$ where $\Omega \subset \mathbb{R}^d$, so $\Delta\varphi = 1$ and use the fact that

$$\int_\Omega v^2 \Delta\varphi\,dx = \int_\Gamma v^2 \partial_n\varphi\,ds - \int_\Omega 2v\nabla v\cdot\nabla\varphi\,dx. \tag{241.19}$$

(b) Give a different proof for square domains of the form $\{x \in \mathbb{R}^2 : |x_i| \le 1\}$ analogous to the proof in one dimension by directly representing u in Ω through line integrals starting at Γ.

For functions $v \in H^1(\Omega)$ with $v = 0$ on Γ, i.e., $\|v\|_{L_2(\Gamma)} = 0$, Poincaré-Friedrichs' inequality implies

$$\|v\|^2_{H^1(\Omega)} = \|\nabla v\|^2_{L_2(\Omega)} + \|v\|^2_{L_2(\Omega)} \le (1+C)\|\nabla v\|^2_{L_2(\Omega)} = (1+C)a(v,v),$$

which proves the V-ellipticity (241.2) with $\kappa_1 = (1+C)^{-1} > 0$.

Since (241.3) and (241.4) follow exactly as in the case of Neumann boundary conditions considered above, it now remains to show that the space $H^1_0(\Omega)$ is a well defined Hilbert space, that is, we need to show that a function in $H^1_0(\Omega)$ has well defined values on the boundary Γ. We start noting that it is in general impossible to uniquely define the boundary values of a function v in $L_2(\Omega)$. This is because by changing a function $v \in L^2(\Omega)$ only very close to the boundary, we can significantly change the boundary values of v without much changing the $L_2(\Omega)$ norm. This is reflected by the fact that there is no constant C such that $\|v\|_{L_2(\Gamma)} \le C\|v\|_{L_2(\Omega)}$ for all functions $v \in L_2(\Omega)$. However, if we change $L_2(\Omega)$ to $H^1(\Omega)$, such an equality holds, and therefore a function v in $H^1(\Omega)$ has well defined boundary values, i.e., the *trace* of $v \in H^1(\Omega)$ on the boundary Γ is well defined. This is expressed in the following *trace inequality:*

Theorem 241.5 *If Ω is a bounded domain with boundary Γ, then there is a constant C such that for all $v \in H^1(\Omega)$,*

$$\|v\|_{L_2(\Gamma)} \le C\|v\|_{H^1(\Omega)}. \tag{241.20}$$

241.6. Prove this. Hint: choose φ such that $\partial\varphi = 1$ on Γ and use (241.19).

241.7. Prove that there is no constant C such that $\|v\|_{L_2(\Gamma)} \leq C\|v\|_{L_2(\Omega)}$ for all $v \in L_2(\Omega)$.

The trace inequality shows that a function v in $H^1(\Omega)$ has well defined boundary values and in particular the boundary condition $v = 0$ on Γ makes sense, and it follows that $H_0^1(\Omega)$ is a Hilbert space.

Note that (241.18) implies that we may use the energy norm $\|\nabla v\|_{L_2(\Omega)} = \sqrt{a(v,v)}$ as an equivalent norm on $V = H_0^1(\Omega)$. As we said, choosing this norm, (241.2) and (241.3) hold with $\kappa_1 = \kappa_2 = 1$.

241.8. Verify that the assumptions of the Lax-Milgram theorem are satisfied for the following problems with appropriate assumptions on α and f:

$$\text{(a)} \begin{cases} -u'' + \alpha u = f & \text{in } (0,1), \\ u(0) = u'(1) = 0, & \alpha = 0 \text{ and } 1. \end{cases}$$

$$\text{(b)} \begin{cases} -u'' = f & \text{in } (0,1), \\ u(0) - u'(0) = u(1) + u'(1) = 0. \end{cases}$$

241.9. Verify that the assumptions of the Lax-Milgram theorem are satisfied for the beam problem:

$$\frac{d^4u}{dx^4} = f \quad \text{in } (0,1),$$

with the boundary conditions; (a) $u(0) = u'(0) = u(1) = u'(1) = 0$, (b) $u(0) = u''(0) = u'(1) = u'''(1) = 0$, (c) $u(0) = -u''(0) + u'(0) = 0, \quad u(1) = u''(1) + u'(1) = 0$; under appropriate assumptions on f. Give mechanical interpretations of the boundary conditions.

241.6 Remark

We saw earlier that if $f \in L_2(\Omega)$ then (241.4) holds with $V = H^1(\Omega)$ and $\kappa_3 = \|f\|_{L_2(\Omega)}$. We may ask what is the weakest assumption on the right-hand side f that allows (241.4) to hold with $\kappa_3 < \infty$. In true mathematical style, we answer this by defining a weak $H^{-1}(\Omega)$ norm of f,

$$\|f\|_{H^{-1}(\Omega)} = \sup_{v \in H_0^1(\Omega)} \frac{(f,v)}{\|v\|_V},$$

where $V = H_0^1(\Omega)$ using the equivalent norm $\|v\|_{H_0^1(\Omega)} = \|\nabla v\|_{L_2(\Omega)}$. By definition, (241.4) holds with $\kappa_3 = \|f\|_{H^{-1}(\Omega)}$. By (241.18), the norm $\|\cdot\|_{H^{-1}(\Omega)}$ may be dominated by the $L_2(\Omega)$ norm:

$$\|f\|_{H^{-1}(\Omega)} \leq \frac{\|f\|_{L_2(\Omega)}\,\|v\|_{L_2(\Omega)}}{\|\nabla v\|_{L_2(\Omega)}} \leq \sqrt{C}\|f\|_{L_2(\Omega)};$$

In fact, the $H^{-1}(\Omega)$ norm is weaker than the $L_2(\Omega)$ norm, which allows us to use right-hand sides $f(x)$ in Poisson's equation that do not belong to

$L_2(\Omega)$, such as the "near point load" used in the tent problem considered in Chapter **??**.

241.10. Show that Lax-Milgram applies to problem (241.13) with $\Omega = \{x \in \mathbb{R}^2 : |x| < 1\}$ and $f(x) = |x|^{-1}$, although in this case $f \notin L_2(\Omega)$.

241.6.1 Non-homogeneous boundary data

Generally, nonhomogeneous boundary data is incorporated into the linear form L along with the right-hand side f. For example, recalling the discussion on Neumann/Robin boundary conditions in Chapter **??**, we see that the problem $-\Delta u + u = f$ in Ω posed with nonhomogeneous Neumann conditions $\partial_n u = g$ on Γ takes the variational form (241.1) with $V = H^1(\Omega)$, $a(u, v)$ defined as in (241.14) and

$$L(v) = \int_\Omega f v \, dx + \int_\Gamma g v \, ds.$$

The continuity of $L(\cdot)$ follows assuming $f \in H^{-1}(\Omega)$ and $g \in L_2(\Gamma)$.

241.11. Prove the last claim.

241.12. Formulate the variational problem associated to Poisson's equation with non-homogeneous Dirichlet boundary conditions given by g on Γ.

241.13. Show that the Lax-Milgram theorem applies to the problem $-\Delta u + \alpha u = f$ in Ω, $\partial_n u + \sigma u = g$ on Γ, for (a) $\alpha = 1$ and $\sigma = 0$, (b) $\alpha = 0$ and $\sigma = 1$. What can be said in the case $\alpha = \sigma = 0$.

241.6.2 A diffusion dominated convection-diffusion problem

The convection-diffusion problem

$$\begin{cases} -\epsilon \Delta u + \beta \cdot \nabla u + \alpha u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma, \end{cases} \tag{241.21}$$

where Ω is domain in $\mathbb{R}^d$ with boundary Γ, $\epsilon > 0$ is constant, and $\beta(x)$ and $\alpha(x)$ are given coefficients, takes the variational form (241.1) with $V = H_0^1(\Omega)$ and

$$a(u, v) = \int_\Omega \big(\epsilon \nabla u \cdot \nabla v + \beta \cdot \nabla u \, v + \alpha u \, v\big) \, dx, \quad L(v) = \int_\Omega f v \, dx.$$

In this case, $a(\cdot, \cdot)$ is not symmetric because of the convection term. To guarantee ellipticity we assume recalling (247.9) that $-\frac{1}{2}\nabla \cdot \beta + \alpha \geq 0$ in Ω, which by (247.13) guarantees that for all $v \in H_0^1(\Omega)$,

$$\int_\Omega \big(\beta \cdot \nabla v \, v + \alpha v^2\big) \, dx \geq 0.$$

It follows that $a(v,v) \geq 0$ and the assumptions of the Lax-Milgram theorem hold, but the stability estimate degrades with decreasing ϵ so that the theorem is mostly relevant for diffusion-dominated problems.

241.14. Prove the preceding statement with specific focus on the dependence of the constants on ϵ.

241.6.3 Linear elasticity in $\mathbb{R}^3$

No body is so small that it is without elasticity. (Leibniz)

As an example of a problem in $\mathbb{R}^3$, we let Ω be a bounded domain in $\mathbb{R}^3$ with boundary Γ split into two parts Γ_1 and Γ_2 and consider the basic problem of linear elasticity modeled by *Cauchy-Navier's elasticity equations*: find the *displacement* $u = (u_i)_{i=1}^3$ and the *stress tensor* $\sigma = (\sigma_{ij})_{i,j=1}^3$ satisfying

$$\begin{cases} \sigma = \lambda \,\mathrm{div}\, @, u\, I + 2\mu\epsilon(u) & \text{in } \Omega, \\ -\mathrm{div}\, @, \sigma = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma_1, \\ \sigma \cdot n = g & \text{on } \Gamma_2, \end{cases} \tag{241.22}$$

where λ and μ are positive constants called the *Lamé coefficients*, $\epsilon(u) = (\epsilon_{ij}(u))_{i,j=1}^3$ is the *strain tensor* with components

$$\epsilon_{ij}(u) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right),$$

$$\mathrm{div}\, @, \sigma = \left(\sum_{j=1}^3 \frac{\partial \sigma_{ij}}{\partial x_j}\right)_{i=1}^3 \quad \text{and} \quad \mathrm{div}\, @, u = \sum_{i=1}^3 \frac{\partial u_i}{\partial x_i},$$

$I = (\delta_{ij})_{i,j=1}^3$ with $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$, $f \in [L_2(\Omega)]^3$ and $g \in [L_2(\Gamma_1)]^3$ are given loads, $n = (n_j)$ is the outward unit normal to Γ_1, and $(\sigma \cdot n)_i = \sum_{j=1}^3 \sigma_{ij} n_j$. For simplicity, we assume that λ and μ are constant. The equations (241.22) express *Hooke's law* connecting stresses and strains and the *equilibrium equation* stating equilibrium of external and internal forces.

The problem has the variational form (241.1) with the choices:

$$V = \left\{ v \in \left[H^1(\Omega)\right]^3 : v = 0 \text{ on } \Gamma_1 \right\},$$

$$a(u,v) = \int_\Omega \left(\lambda \,\mathrm{div}\, @, u \,\mathrm{div}\, @, v + 2\mu\epsilon(u) : \epsilon(v)\right) dx,$$

$$L(v) = \int_\Omega f \cdot v \, dx + \int_{\Gamma_1} g \cdot v \, ds,$$

where $\epsilon(u) : \epsilon(v) = \sum_{i,j=1}^{3} \epsilon_{ij}(u)\epsilon_{ij}(v)$. We note that the bilinear form a has the form of "virtual work",

$$a\,(u, v) = \int_{\Omega} \sigma(u) : \epsilon(v)\, dx,$$

where $\sigma(u) = \lambda \text{div } @, u\, I + 2\mu\epsilon(u)$. To prove V-ellipticity, we use *Korn's inequality*. For simplicity, we assume that $\Gamma_1 = \Gamma$.

Theorem 241.6 *There is a constant c such that for all $v \in [H_0^1(\Omega)]^3$,*

$$\sum_{i,j=1}^{3} \int_{\Omega} \epsilon_{ij}(v)\epsilon_{ij}(v) dx \geq c \sum_{i=1}^{3} \|v_i\|^2_{H^1(\Omega)}.$$

Proof Using the notation $v_{i,j} = \partial v_i/\partial x_j$, $v_{i,jl} = \partial^2 v_i/\partial x_j \partial x_l$, etc.,

$$\sum_{i,j=1}^{3} \epsilon_{ij}(v)\epsilon_{ij}(v) = \sum_{i,j=1}^{3} \frac{1}{2} v_{i,j} v_{i,j} + \sum_{i,j=1}^{3} \frac{1}{2} v_{i,j} v_{j,i}.$$

Integrating the second term on the right and then using integration by parts, we get

$$\begin{aligned}
\sum_{i,j=1}^{3} \int_{\Omega} v_{i,j} v_{j,i}\, dx &= \int_{\Gamma} v_{i,j} v_j n_i\, ds - \int_{\Omega} v_{i,ji} v_j\, dx \\
&= \sum_{i,j=1}^{3} \int_{\Gamma} v_{i,j} v_j n_i\, ds - \int_{\Gamma} v_{i,i} v_j n_j\, ds + \int_{\Omega} v_{i,i} v_{j,j}\, dx \\
&= \sum_{i,j=1}^{3} \int_{\Omega} v_{i,i} v_{j,j}\, dx,
\end{aligned}$$

since $v = 0$ on Γ. We conclude that

$$\sum_{i,j=1}^{3} \int_{\Omega} \epsilon_{ij}(v)\epsilon_{ij}(v)\, dx = \frac{1}{2} \sum_{i,j=1}^{3} \int_{\Omega} \big(v_{i,j}\big)^2\, dx + \frac{1}{2} \int_{\Omega} \big(\sum_{i=1}^{3} v_{i,i}\big)^2\, dx.$$

The desired inequality follows using Poincaré's inequality to bound the L_2 norm of v_i in terms of the L_2 norm of ∇v_i. ■

241.15. Provide the last details.

241.16. Solve the Cauchy-Navier elasticity equations for the cantilever beam in two dimensions using Femlab. Compare with analytic solutions of the beam equation.

241.6.4 The Stokes equations

The Stokes equations for stationary incompressible creeping fluid flow with zero velocity boundary conditions read: find the *velocity* $u = (u_i)_{i=1}^3$, total *stress* $\sigma = (\sigma_{ij})_{i,j=1}^3$, and the *pressure* p such that

$$\begin{cases} \sigma = -pI + 2\mu\epsilon(u) & \text{in } \Omega, \\ -\text{div } @, \sigma = f & \text{in } \Omega, \\ \text{div } @, u = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma, \end{cases}$$

Eliminating the stress σ gives

$$\begin{cases} -\mu\Delta u + \nabla p = f & \text{in } \Omega, \\ \text{div } @, u = 0 & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma. \end{cases} \tag{241.23}$$

This can be formulated in variational form (241.1) with

$$V = \big\{v \in [H^1(\Omega)]^3 : \text{div}@, u = 0 \text{ in } \Omega\big\},$$

$$a(u,v) = \int_\Omega \sum_{i=1}^3 \nabla u_i \cdot \nabla v_i \, dx, \text{ and } L(v) = \int_\Omega f \cdot v \, dx.$$

The picture on the cover of the book shows streamlines of Stokes flow around a sphere.

241.17. Prove that the assumptions of the Lax-Milgram theorem hold in this case. this.

241.18. Extend the mechanical models of Section **??** to several dimensions.

Note that the stationary Navier-Stokes equations are obtained by adding the term $(\nabla \cdot u)u$ to the first equation in (241.23).

241.7 A strong stability estimate for Poisson's equation

We conclude this chapter by proving the strong stability estimate (240.6) for solutions to Poisson's equation that we used in the proofs of the L_2 error estimates for elliptic and parabolic problems. The estimate shows that the $L_2(\Omega)$ norm of all second derivatives of a function v vanishing on the boundary of a convex domain are bounded by the $L_2(\Omega)$ norm of the particular combination of second derivatives given by the Laplacian. For simplicity, we consider the case of a convex domain in the plane with smooth boundary.

Theorem 241.7 *If Ω is a bounded domain in $\mathbb{R}^2$ with smooth boundary Γ then for all smooth functions v with $v = 0$ on Γ,*

$$\sum_{i,j=1}^{3} \int_\Omega \left(D^2 v\right)^2 dx + \int_\Gamma \frac{1}{R} \left(\frac{\partial v}{\partial n}\right)^2 ds = \int_\Omega (\Delta v)^2 \, dx,$$

where $R(x)$ is the radius of curvature of Γ at $x \in \Gamma$ with $R(x) \geq 0$ if Ω is convex, see Fig. 241.1.

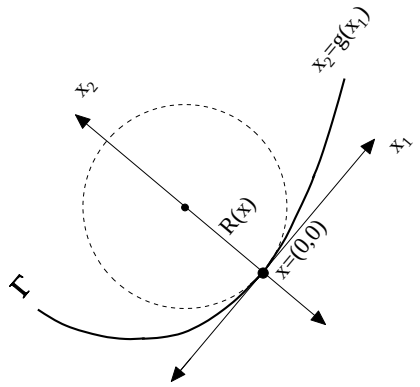


FIGURE 241.1. The radius of curvature and the local coordinate system near a point x on Γ.

Proof: We use the notation $v_{(i)} = \partial v / \partial x_i$, $v_{(ij)} = \partial^2 v / \partial x_i \partial x_j$, etc.. Assuming that v is smooth with $v = 0$ on Γ, integration by parts gives

$$\begin{aligned}
\int_\Omega \Delta v \Delta v \, dx &= \sum_{i,j=1}^{3} \int_\Omega v_{(ii)} v_{(jj)} \, dx \\
&= \sum_{i,j=1}^{3} \int_\Gamma v_{(i)} v_{(jj)} n_i \, ds - \sum_{i,j=1}^{3} \int_\Omega v_{(i)} v_{(ijj)} \, dx \\
&= \sum_{i,j=1}^{3} \int_\Gamma \left(v_{(i)} v_{(jj)} n_i - v_{(i)} v_{(ij)} n_j \right) ds + \sum_{i,j=1}^{3} \int_\Omega v_{(ij)} v_{(ij)} \, dx.
\end{aligned}$$

Recalling the definition of $D^2 v$ from Chapter **??**

$$\int_\Omega \left((\Delta v)^2 - (D^2 v)^2 \right) dx = \sum_{i,j=1}^{3} \int_\Gamma \left(v_{(i)} v_{(jj)} n_i - v_{(i)} v_{(ij)} n_j \right) ds.$$

To evaluate the integrand on the right at a point $x \in \Gamma$, we use the fact that the integrand is invariant under orthogonal coordinate transformations. We

may assume that $x = (0,0)$ and that in a neighborhood of x, the graph of Γ is described by the equation $x_2 = g(x_1)$ in a local coordinate system, see Fig. 241.1. Since $v = 0$ on Γ, we have $v(x_1, g(x_1)) = 0$ for x_1 in some neighborhood of 0 and thus by differentiation with respect to x_1, we find that

$$v_{(1)} + v_{(2)}g'(x_1) = 0,$$
$$v_{(11)} + 2v_{(12)}g'(x_1) + v_{(22)}(g'(x_1))^2 + v_{(2)}g''(x_1) = 0.$$

Since $g'(0) = 0$ and, by the definition of the radius of curvature, $g''(0) = 1/R(0)$, we conclude that

$$v_{(1)}(0,0) = 0$$
$$v_{(11)}(0,0) = -v_{(2)}(0,0)/R(0).$$

At $x = (0,0)$, since $n = (0,-1)^\top$ at that point,

$$\sum_{i,j=1}^{3} \big(v_{(i)}v_{(jj)}n_i - v_{(i)}v_{(ij)}n_j\big) = -v_{(2)}\big(v_{(1)1} + v_{(22)}\big) + v_{(2)}v_{(22)}$$
$$= -v_{(2)}v_{(11)} = \big(v_{(2)}\big)^2/R = \big(\partial v/\partial n\big)^2/R.$$

and the statement of the theorem follows. ■

241.19. (A maximum principle). Prove that if u is continuous in $\Omega \cup \Gamma$, where Ω is a domain with boundary Γ, and $\Delta u(x) \geq 0$ for $x \in \Omega$, then u attains its maximum on the boundary Γ. Hint: consider first the case that $\Delta u(x) > 0$ for $x \in \Omega$ and arrive at a contradiction by assuming a maximum is attained in Ω that is not on Γ by using the fact that at such a point, the second derivatives with respect to x_i cannot be positive. Extend this to the case $\Delta u(x) \geq 0$ by considering the function $u_\epsilon(x) = u(x) + \epsilon|x - \bar{x}|^2$, which for $\epsilon > 0$ sufficiently small also has an interior maximum.

241.20. Consider the problem

$$\begin{cases} -(u_{(11)} - u_{(12)} + 2u_{(22)}) + u_{(1)} + u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_1, \\ u_{(1)}n_1 - \frac{1}{2}u_{(1)}n_2 - \frac{1}{2}u_{(2)}n_1 + u_{(2)}n_2 + u = g & \text{on } \Gamma_2. \end{cases}$$

Give a variational formulation of this problem and show that the conditions in the Lax-Milgram lemma (except symmetry) are satisfied.

Ein jeder Geist steht vor den ganzen Bau der Dinge,
Als ob die Fernung sich in einen Spiegel bringe,
Nach jeden Augenpunct, verdunckelt oder klar,
Er ist ein Bild, wie er ein Zweck der Schöpfung war
(Leibniz, at the funeral of Queen Sophie Charlotte, 1705)

FIGURE 241.2. Queen Sophie Charlotte von Brandenburg, gifted student of Leibniz's philosophy.

242
Heat Equation Analysis

The simpler a hypothesis is, the better it is. (Leibniz)

Thus then, we are led to the conception of a complicated mechanism capable of a vast variety of motion... Such a mechanism must be subject to the general laws of Dynamics, and we ought to be able to work out all the consequences of its motion, provided we know the form of the relation between the motions of the parts... We now proceed to investigate whether the properties of that which constitutes the electromagnetic field, deduced from electromagnetic phenomena alone, are sufficient to explain the propagation of light through the same substance. (Maxwell)

242.1 Introduction

In this chapter, we consider the numerical solution of the *heat equation*, which is the prototype of a linear parabolic partial differential equation. Recall that we originally derived the heat equation in Chapter **??** to model heat flow in a conducting object. More generally, the same equation may be used to model *diffusion* type processes. From a quite different point of view, we begin this chapter by deriving the heat equation as a consequence of Maxwell's equations under some simplifying assumptions. After that, we recall some of the properties of solutions of the heat equation, focussing on the characteristic parabolic "smoothing" and stability properties. We then proceed to introduce a finite element method for the heat equation, derive

a posteriori and a priori error estimates and discuss adaptive error control. The analysis follows the basic steps used in the analysis of the parabolic model problem in Chapter **??** and of Poisson's equation in Chapter **??**.

242.2 Maxwell's equations

We met in the previous chapter a special case of Maxwell's equations in the form of Poisson's equation for an electric potential in electrostatics. Here, we consider another special case that leads to a parabolic problem for a magnetic potential, which in the simplest terms reduces to the heat equation. Another important special case gives rise to the wave equation studied in Chapter **??**.

It is remarkable that the complex phenomena of interaction between electric and magnetic fields can be described by the relatively small set of Maxwell's equations:

$$\begin{cases} \dfrac{\partial B}{\partial t} + \nabla \times E = 0, \\ -\dfrac{\partial D}{\partial t} + \nabla \times H = J, \\ \nabla \cdot B = 0, \quad \nabla \cdot D = \rho, \\ B = \mu H, \quad D = \epsilon E, \quad J = \sigma E, \end{cases} \tag{242.1}$$

where E is the electric field, H is the magnetic field, D is the electric displacement, B is the magnetic flux, J is the electric current, ρ is the charge, μ is the magnetic permeability, ϵ is the dielectric constant, and σ is the electric conductivity. The first equation is referred to as *Faraday's law*, the second is *Ampère's law*, $\nabla \cdot D = \rho$ is *Coulomb's law*, $\nabla \cdot B = 0$ expresses the absence of "magnetic charge", and $J = \sigma E$ is *Ohm's law*. Maxwell included the term $\partial D/\partial t$ for purely mathematical reasons and then using calculus predicted the existence of electromagnetic waves before these had been observed experimentally. We assume to start with that $\partial D/\partial t$ can be neglected; cf. Problem 242.1 and Problem 245.11.

Because $\nabla \cdot B = 0$, B can be written as $B = \nabla \times A$, where A is a magnetic vector potential. Inserting this into Faraday's law gives

$$\nabla \times \left(\frac{\partial A}{\partial t} + E \right) = 0,$$

from which it follows that

$$\frac{\partial A}{\partial t} + E = \nabla V,$$

for some scalar potential V. Multiplying by σ and using the laws of Ohm and Ampère, we obtain a vector equation for the magnetic potential A:

$$\sigma \frac{\partial A}{\partial t} + \nabla \times \left(\mu^{-1} \nabla \times A \right) = \sigma \nabla V.$$

To obtain a scalar equation in two variables, we assume that $B = (B_1, B_2, 0)$ is independent of x_3. It follows that A has the form $A = (0, 0, u)$ for some scalar function u that depends only on x_1 and x_2, so that $B_1 = \partial u/\partial x_2$ and $B_2 = -\partial u/\partial x_1$, and we get a scalar equation for the scalar magnetic potential u of the form

$$\sigma \frac{\partial u}{\partial t} - \nabla \cdot \left(\mu^{-1} \nabla u\right) = f, \tag{242.2}$$

for some function $f(x_1, x_2)$. This is a parabolic equation with variable coefficients σ and μ. Choosing $\sigma = \mu = 1$ leads to the heat equation:

$$\begin{cases} \frac{\partial}{\partial t} u(x,t) - \Delta u(x,t) = f(x,t) & \text{for } x \in \Omega,\ 0 < t \le T, \\ u(x,t) = 0 & \text{for } x \in \Gamma,\ 0 < t \le T, \\ u(x,0) = u_0(x) & \text{for } x \in \Omega, \end{cases} \tag{242.3}$$

where $\Omega \subset \mathbb{R}^2$ with boundary Γ, and we posed homogeneous Dirichlet boundary conditions.

242.1. What equation is obtained if $\partial D/\partial t$ is not neglected, but the x_3 independence is kept?

242.2. Show that the magnetic field H around a unit current along the x_3-axis is given by $\frac{1}{2\pi|x|}(-x_2, x_1, 0)$, where $|x| = (x_1^2 + x_2^2)^{\frac{1}{2}}$.

242.3 The basic structure of solutions of the heat equation

The structure of solutions of the heat equation is closely related to the properties of solutions of the initial value problems discussed in Chapter 212.1 and the boundary value problems discussed in Chapters **??** and **??**.

For some domains the method of separation of variables can be employed to find analytic solutions of the heat equation in terms of series expansions into eigenfunctions. We illustrate this approach for the one-dimensional, homogeneous heat equation

$$\begin{cases} \dot{u}(x,t) - u''(x,t) = 0 & \text{for } 0 < x < \pi,\ t > 0, \\ u(0,t) = u(\pi,t) = 0 & \text{for } t > 0, \\ u(x,0) = u_0(x) & \text{for } 0 < x < \pi. \end{cases} \tag{242.4}$$

We start by seeking solutions of the differential equation and the boundary conditions in (252.27) of the form $u(x,t) = \varphi(x)\,\psi(t)$ with $\varphi(0) = \varphi(\pi) = 0$. Substituting this into (252.27) and separating the functions depending on x and t, gives

$$\frac{\dot{\psi}(t)}{\psi(t)} = \frac{\varphi''(x)}{\varphi(x)}.$$

Since x and t are independent variables, each fraction must be equal to the same constant $-\lambda \in \mathbb{R}$ and we are led to the eigenvalue problem

$$\begin{cases} -\varphi''(x) = \lambda\varphi(x) & \text{for } 0 < x < \pi, \\ \varphi(0) = \varphi(\pi) = 0, \end{cases} \tag{242.5}$$

and the initial value problem

$$\begin{cases} \dot{\psi}(t) = -\lambda\psi(t) & \text{for } t > 0, \\ \psi(0) = 1, \end{cases} \tag{242.6}$$

where $\psi(0)$ is normalized to 1. Thus, seeking solutions in the form of a product of functions of one variable decouples the partial differential equation into two ordinary differential equations. It is important to this technique that the differential equation is linear, homogeneous, and has constant coefficients.

The problem (242.5) is an eigenvalue problem with eigenfunctions $\varphi_j(x) = \sin(jx)$ and corresponding eigenvalues $\lambda_j = j^2$, $j = 1, 2, ...$ For each eigenvalue, we can solve (242.6) to get the corresponding solution $\psi(t) = \exp(-j^2 t)$. We obtain a set of solutions $\{\exp(-j^2 t)\sin(jx)\}$ of (252.27) with corresponding initial data $\{\sin(jx)\}$ for $j = 1, 2, ..$, which are called the *eigenmodes*. Each eigenmode decays exponentially as time passes and the rate of decay increases with the frequency j. We illustrate this in Fig. 252.8. Any finite linear combination of eigenmodes

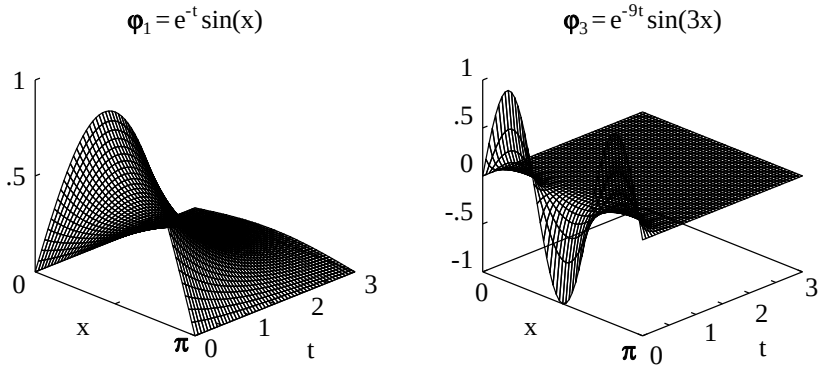


FIGURE 242.1. The solutions of the heat equation corresponding to frequencies $j = 1$ and $j = 3$.

$$\sum_{j=1}^{J} a_j \exp(-j^2 t)\sin(jx),$$

with coefficients $a_j \in \mathbb{R}$, is a solution of the homogeneous heat equation corresponding to the initial data

$$u_0(x) = \sum_{j=1}^{J} a_j \sin(jx). \tag{242.7}$$

More generally, if the initial data u_0 has a convergent Fourier series,

$$u_0(x) = \sum_{j=1}^{\infty} u_{0,j} \sin(jx),$$

with Fourier coefficients given by $u_{0,j} = 2\pi^{-1} \int_0^\pi u_0(x) \sin(jx)\, dx$, then the function defined by

$$u(x,t) = \sum_{j=1}^{\infty} u_{0,j} \exp(-j^2 t) \sin(jx), \tag{242.8}$$

solves $\dot{u} - u'' = 0$. This is seen by differentiating the series term by term, which is possible because the coefficients $u_{0,j} \exp(-j^2 t)$ decrease very quickly with j as long as $t > 0$. Moreover $u(0) = u(\pi) = 0$, so to show that u is a solution of (252.27), we only have to check that $u(x,t)$ equals the initial data u_0 at $t = 0$. If we only require that $u_0 \in L_2(0,\pi)$, then it is possible to show that

$$\lim_{t \to 0} \|u(\cdot, t) - u_0\| = 0. \tag{242.9}$$

If u_0 has additional smoothness and also satisfies the boundary conditions $u_0(0) = u_0(\pi) = 0$ (which is not required if we only assume that $u_0 \in L_2(0,\pi)$), then the initial data is assumed in the stronger pointwise sense, i.e.

$$\lim_{t \to 0} u(x,t) = u_0(x) \quad \text{for } 0 < x < \pi. \tag{242.10}$$

Recalling that the rate at which a function's Fourier coefficients tends to zero reflect the smoothness of the function, we see from the solution formula (252.30) that a solution $u(x,t)$ of the homogeneous heat equation becomes smoother with increasing time. This is known as *parabolic smoothing*. We illustrate the smoothing in Fig. 242.2, where we plot the solution starting with the discontinuous function

$$u_0(x) = \begin{cases} x, & 0 \le x \le \pi/2, \\ x - \pi, & \pi/2 < x \le \pi, \end{cases}$$

at various times (the solution formula is given in Problem 242.3). This corresponds well with intuition about a diffusive process in which sharp features are smoothed out for positive time. Nonsmooth functions have slowly decreasing Fourier coefficients, so that the Fourier coefficients of the

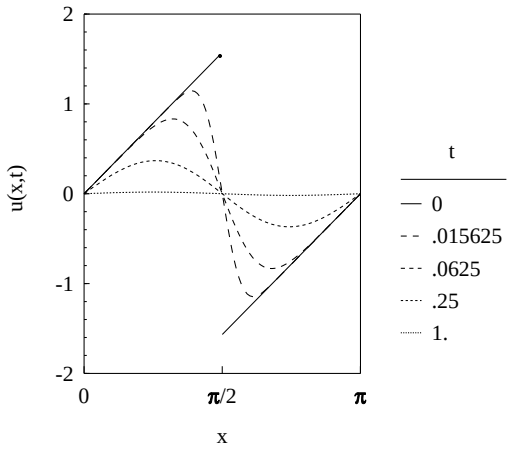


FIGURE 242.2. The evolution of discontinuous initial data for the heat equation.

high modes with j large are relatively large compared to those of smooth functions. As soon as $t > 0$, these high modes are damped rapidly because of the presence of the factor $\exp(-j^2 t)$, and the solution becomes smoother as t increases.

242.3. Verify the following formulas for the solutions of the heat equation corresponding to the indicated initial data:

1. $u_0(x) = x(\pi - x)$,

$$u(x,t) = \sum_{j=1}^{\infty} \frac{8}{(2j-1)^3}\, e^{-(2j-1)^2 t}\, \sin((2j-1)x).$$

2. $u_0(x) = \begin{cases} x, & 0 \le x \le \pi/2 \\ \pi - x, & \pi/2 < x \le \pi \end{cases}$,

$$u(x,t) = \sum_{j=1}^{\infty} \frac{4(-1)^{j+1}}{\pi(2j-1)^2}\, e^{-(2j-1)^2 t}\, \sin((2j-1)x).$$

3. $u_0(x) = \begin{cases} x, & 0 \le x \le \pi/2 \\ x - \pi, & \pi/2 < x \le \pi \end{cases}$,

$$u(x,t) = \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j}\, e^{-4j^2 t}\, \sin(2jx).$$

242.4. Find a formula for the solution of (252.27) with the Dirichlet boundary conditions replaced by the Neumann conditions $u'(0) = 0$ and $u'(\pi) = 0$. Hint: the series expansion is in terms of cosine functions. Do the same with the boundary conditions $u(0) = 0$ and $u'(\pi) = 0$.

242.5. (a) Prove (242.10) assuming that $\sum_{j=1}^{\infty} |u_{0,j}| < \infty$. (b) Prove (242.9) assuming that $u_0 \in L_2(0,\pi)$, that is $\sum_{j=1}^{\infty} |u_{0,j}|^2 < \infty$.

242.6. (Strauss ([**?**])) Waves in a resistant medium are described by the problem

$$\begin{cases} \ddot{u}(x,t) + c\dot{u}(x,t) - u''(x,t) = 0, & 0 < x < \pi,\ t > 0, \\ u(0,t) = u(\pi,t) = 0, & t > 0, \\ u(x,0) = u_0(x), & 0 < x < \pi, \end{cases}$$

where $c > 0$ is a constant. Write down a series expansion for the solution using separation of variables. Can you say something about the behavior of the solution as time passes?

242.7. Give the Fourier series formula for the solution of the homogeneous heat equation (242.3) posed on the unit square $\Omega = (0,1) \times (0,1)$. Hint: first use separation of variables to get an ordinary differential equation in t and an eigenvalue problem for the Laplacian in (x_1, x_2). Then, use separation of variables to decompose the eigenvalue problem for the Laplacian into independent eigenvalue problems for x_1 and x_2. Hint: see Chapter **??**.

242.8. Consider the *backward heat equation*

$$\begin{cases} \dot{u}(x,t) + u''(x,t) = 0 & \text{for } 0 < x < \pi,\ t > 0, \\ u(0,t) = u(\pi,t) = 0 & \text{for } t > 0, \\ u(x,0) = u_0(x) & \text{for } 0 < x < \pi. \end{cases} \tag{242.11}$$

Write down a solution formula in the case u_0 is a finite Fourier series of the form (242.7). Investigate how the different components of u_0 get amplified with time. Why is the equation called the backward heat equation? Can you find a connection to image reconstruction?

242.4 The fundamental solution of the heat equation

The solution of the homogeneous heat equation

$$\begin{cases} \dot{u} - \Delta u = 0 & \text{in } \mathbb{R}^2 \times (0,\infty), \\ u(\cdot,0) = u_0 & \text{in } \mathbb{R}^2, \end{cases} \tag{242.12}$$

with u_0 equal to the delta function at the origin δ_0, is called the *fundamental solution* of the heat equation and is given by

$$u(x,t) = E(x,t) = \frac{1}{4\pi t} \exp\left(-\frac{|x|^2}{4t}\right). \tag{242.13}$$

Direct computation shows that $E(x,t)$ solves $\dot{E} - \Delta E = 0$ for $x \in \mathbb{R}^2$ and $t > 0$. Further $E(\cdot,t)$ approaches the delta function δ_0 as $t \to 0^+$ since $E(x,t) \geq 0$, $\int_{\mathbb{R}^2} E(x,t)\,dx = 1$ for $t > 0$, and $E(x,t)$ rapidly decays as $|x|/\sqrt{t}$ increases, so that the support of $E(x,t)$ becomes more and more concentrated around $x = 0$ as $t \to 0^+$. In terms of a model of heat, $E(x,t)$ corresponds to choosing the initial conditions to be a "hot spot" at the origin. In Fig. 242.3 we plot $E(x,t)$ at three different times.

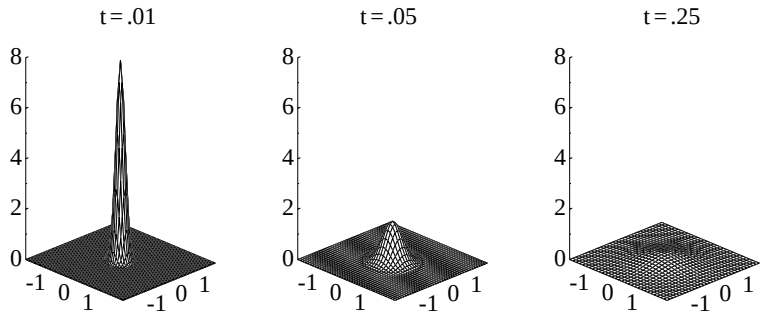


FIGURE 242.3. The fundamental solution $E(x,t)$ at three times.

242.9. Show that E defined by (242.13) solves $\dot{E} - \Delta E = 0$ for $t > 0$, and verify that $\int_{\mathbb{R}} E(x,t)\,dx = 1$.

242.10. Determine the fundamental solution of the heat equation in $\mathbb{R}^d$, d=1,3.

242.11. Give the formula for the fundamental solution $E_\epsilon(x,t)$ for the heat equation $\dot{u} - \epsilon\Delta u = 0$ in two space dimensions, where ϵ is a positive constant. Determine, as a function of ϵ and t, the diameter of the set of points x outside which $E_\epsilon(x,t)$ is essentially zero.

The solution of (242.12) can be expressed in terms of the fundamental solution and the initial data as follows:

$$u(x,t) = \frac{1}{4\pi t}\int_{\mathbb{R}^2} u_0(y)\exp\left(-\frac{|x-y|^2}{4t}\right)dy. \tag{242.14}$$

242.12. Motivate this formula.

From the solution formula we see that the value $u(x,t)$ at a point $x \in \mathbb{R}^2$ and $t > 0$ is a weighted mean value of all the values $u_0(y)$ for $y \in \Omega$. The influence of the value $u_0(y)$ on $u(x,t)$ decreases with increasing distance $|x-y|$ and decreasing time t. In principle, information appears to travel with an *infinite speed of propagation* because even for very small time t there is an influence on $u(x,t)$ from $u_0(y)$ for $|x-y|$ arbitrarily large.

However, the nature of the fundamental solution causes the influence to be extremely small if t is small and $|x-y|$ is large. In particular, the solution formula shows that if $u_0 \geq 0$ is concentrated around $x = 0$, say $u_0(x) \equiv 0$ for $|x| \geq d$ for some small $d > 0$, then $u(x,t)$ "spreads out" over a disk of radius proportional to $\sqrt{t}$ for $t > 0$ and rapidly decays to zero outside this disk.

242.13. (a) Write a code that inputs an x and t and then uses the composite trapezoidal rule to approximate the integrals in (242.14) when $u_0(x)$ is 1 for $|x| \leq 1$ and 0 otherwise and use the code to generate plots of the solution at several different times. (b) *(Harder.)* Verify the claim about the rate of spread of the solution.

242.5 Stability

Throughout the book, we emphasize that the stability properties of parabolic problems are an important characteristic. To tie into the previous stability results for parabolic-type problems, we prove a strong stability estimate for an abstract parabolic problem of the form: find $u(t) \in H$ such that

$$\begin{cases} \dot{u}(t) + Au(t) = 0 \quad \text{for } t > 0, \\ u(0) = u_0, \end{cases} \tag{242.15}$$

where H is a vector space with inner product $(\cdot,\cdot)$ and norm $\|\cdot\|$, A is a positive semi-definite symmetric linear operator defined on a subspace of H, i.e. A is a linear transformation satisfying $(Aw, v) = (w, Av)$ and $(Av, v) \geq 0$ for all v and w in the domain of definition of A, and u_0 is the initial data. In the parabolic model problem of Chapter 212.1, $H = \mathbb{R}^d$ and A is a positive semi-definite symmetric $d \times d$ matrix. In the case of the heat equation (242.3), $A = -\Delta$ is defined on the infinite-dimensional space of functions v in $L^2(\Omega)$ which are square integrable and satisfy homogeneous Dirichlet boundary conditions.

Lemma 242.1 *The solution u of (242.15) satisfies for $T > 0$,*

$$\|u(T)\|^2 + 2\int_0^T (Au(t), u(t))\,dt = \|u_0\|^2, \tag{242.16}$$

$$\int_0^T t\|Au(t)\|^2\,dt \leq \frac{1}{4}\|u_0\|^2, \tag{242.17}$$

$$\|Au(T)\| \leq \frac{1}{\sqrt{2}\,T}\|u_0\|. \tag{242.18}$$

Proof: The proof uses the same ideas used to show (**??**). Taking the inner product of (242.15) with $u(t)$, we obtain

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2 + (Au(t), u(t)) = 0,$$

from which (242.16) follows.

Next, taking the inner product of the first equation of (242.15) with $tAu(t)$ and using the fact that

$$(\dot{u}(t), tAu(t)) = \frac{1}{2}\frac{d}{dt}\big(t(Au(t), u(t))\big) - \frac{1}{2}(Au(t), u(t)),$$

since A is symmetric, we find after integration that

$$\frac{1}{2}T(Au(T), u(T)) + \int_0^T t\|Au(t)\|^2\,dt = \frac{1}{2}\int_0^T (Au(t), u(t))\,dt,$$

from which (242.17) follows using (242.16) and the fact that $(Av, v) \geq 0$.

Finally, taking the inner product in (242.15) with $t^2A^2u(t)$, we obtain

$$\frac{1}{2}\frac{d}{dt}\big(t^2\|Au(t)\|^2\big) + t^2(A^2u(t), Au(t)) = t\|Au(t)\|^2,$$

from which (242.18) follows after integration and using (242.17).■

242.14. Assuming that there is an $a > 0$ such that A is strictly positive-definite, so that $(Av, v) \geq a\|v\|^2$ for all v, show that the solution of $\dot{u} + Au = f$, $u(0) = u_0$, satisfies

$$\|u(T)\|^2 + a\int_0^T \|u(t)\|^2\,dt \leq \|u_0\|^2 + \frac{1}{a}\int_0^T \|f\|^2\,dt.$$

Hint: use that $|(v, w)| \leq (4\epsilon)^{-1}\|v\|^2 + \epsilon\|w\|^2$ for any $\epsilon > 0$.

In the case of a solution of the heat equation (242.3), these estimates read

$$\|u(T)\|^2 + 2\int_0^T (\nabla u(t), \nabla u(t))\,dt \leq \|u_0\|^2, \tag{242.19}$$

$$\int_0^T t\|\Delta u(t)\|^2\,dt \leq \frac{1}{4}\|u_0\|^2, \tag{242.20}$$

$$\|\Delta u(T)\| \leq \frac{1}{\sqrt{2}\,T}\|u_0\|. \tag{242.21}$$

242.15. (a) Consider u and $\tilde{u}$ solving (242.3) with initial data $u_0(x)$ and $\tilde{u}_0(x) = u_0(x) + \epsilon(x)$ respectively. Show that the difference $\tilde{u} - u$ solves (242.3) with initial data $\epsilon(x)$. (b) Give estimates for the difference between u and $\tilde{u}$. (c) Prove that the solution of (242.3) is unique.

Recall that we call these *strong stability* estimates because they provide bounds on derivatives of the solution as well as the solution itself. Such estimates are related to parabolic smoothing. For example, (242.21) implies that the L_2 norm of the derivative $\dot{u}(T) = \Delta u(T)$ decreases (increases) like $1/T$ as T increases (decreases), which means that the solution becomes smoother as time passes.

242.16. Compute (exactly or approximately) the quantities on the left-hand sides of (242.16), (242.17), and (242.18) for the solutions of (252.27) computed in Problem 242.3. Compare to the bounds on the right-hand sides.

242.17. Prove the stability estimates of Lemma 242.1 applied to the one-dimensional heat equation (232.41) using the Fourier series formula for the solution.

243
Heat Equation FEM

Think for yourself, 'cause I won't be with you. (George Harrison)

I see no essential difference between a materialism, which includes a soul as a complicated type of material particle, and a spiritualism that includes particles as a primitive type of soul. (Wiener)

243.1 Space-Time Discretization

The time discretization of the heat equation (242.3) is based on a partition $0 = t_0 < t_1 < \cdots < t_N = T$ of the time interval $I = [0, T]$ into sub-intervals $I_n = (t_{n-1}, t_n)$ of length $k_n = t_n - t_{n-1}$. We divide each *space-time slab* $S_n = \Omega \times I_n$ into space-time prisms $K \times I_n$, where $\mathcal{T}_n = \{K\}$ is a triangulation of Ω with mesh function h_n; see Fig. 243.1. Note that the space mesh may change from one time interval to the next. We construct a finite element method using approximations consisting of continuous piecewise linear functions in space and discontinuous polynomials of degree r in time, which we call the cG(1)dG(r) method. We define the trial space $W_k^{(r)}$ to be the set of functions $v(x,t)$ defined on $\Omega \times I$ such that the restriction $v|_{S_n}$ of v to each space-time slab S_n is continuous and piecewise linear in x and

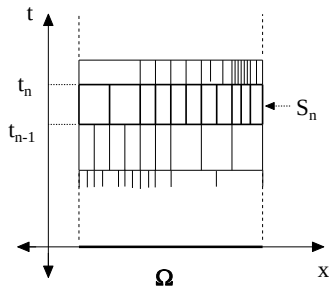


FIGURE 243.1. Space-time discretization for the cG(1)dG(r) method.

a polynomial of degree r in t, that is, $v|_{S_n}$ belongs to the space

$$W_{kn}^{(r)} = \left\{ v(x,t) : \, v(x,t) = \sum_{j=0}^{r} t^j \psi_j(x), \ \psi_j \in V_n, \ (x,t) \in S_n \right\},$$

where $V_n = V_{h_n}$ is the space of continuous piecewise linear functions vanishing on Γ associated to $\mathcal{T}_n$. The "global" trial space $W_k^{(r)}$ is the space of functions v defined on $\Omega \times I$, such that $v|_{S_n} \in W_{kn}^{(r)}$ for $n = 1, 2, ..., N$. The functions in $W_k^{(r)}$ in general are discontinuous across the discrete time levels t_n and we use the usual notation $[w_n] = w_n^+ - w_n^-$ and $w_n^{+(-)} = \lim_{s \to 0^{+(-)}} w(t_n + s)$.

243.1. Describe a set of basis functions for (a) $W_{kn}^{(0)}$ and (b) $W_{kn}^{(1)}$.

The cG(1)dG(r) method is based on a variational formulation of (242.3) as usual and reads: find $U \in W_k^{(r)}$ such that for $n = 1, 2, \ldots, N$,

$$\int_{I_n} \left((\dot{U}, v) + (\nabla U, \nabla v) \right) dt + \left([U_{n-1}], v_{n-1}^+ \right) = \int_{I_n} (f, v) \, dt \quad \text{for all } v \in W_{kn}^{(r)}, \quad (243.1)$$

where $U_0^- = u_0$ and $(\cdot, \cdot)$ is the $L_2(\Omega)$ inner product.

Using the discrete Laplacian Δ_n, se (240.10), we may write (243.1)in the case $r = 0$ as follows: find $U_n \in V_n$:

$$(I - k_n \Delta_n) U_n = P_n U_{n-1} + \int_{I_n} P_n f \, dt, \quad (243.2)$$

where we set $U_n = U_n^- = U|_{I_n} \in V_n$, and P_n is the $L_2(\Omega)$-projection onto V_n. Note that the "initial data" $U_{n-1} \in V_{n-1}$ from the previous time interval I_{n-1} is projected into the space V_n. If $V_{n-1} \subset V_n$, then $P_n U_{n-1} =$

U_{n-1}. In the case $r = 1$, writing $U(t) = \Phi_n + (t - t_{n-1})\Psi_n$ on I_n with Φ_n, $\Psi_n \in V_n$, then (243.1) becomes

$$\begin{cases} (I - k_n\Delta_n)\Phi_n + \big(I - \frac{k_n}{2}\Delta_n\big)\Psi_n = P_nU_{n-1} + \int_{I_n} P_nf\,dt, \\ \big(\frac{1}{2}I - \frac{k_n}{3}\Delta_n\big)\Psi_n - \frac{k_n}{2}\Delta_n\Phi_n = \int_{I_n} \frac{t - t_{n-1}}{k_n} P_nf\,dt, \end{cases} \tag{243.3}$$

which gives a system of equations for Φ_n and Ψ_n.

243.2. Verify (243.2) and (243.3).

243.3. Writing $U(t) = \Phi_n(t_n - t)/k_n + \Psi_n(t - t_{n-1})/k_n$ on I_n with Φ_n, $\Psi_n \in V_n$, formulate equations for the cG(1)dG(1) approximation using the discrete Laplacian.

243.2 Constructing the discrete equations

To construct the matrix equation that determines U_n in the case $r = 0$ according to (243.2), we introduce some notation. We let $\{\varphi_{n,j}\}$ denote the nodal basis of V_n associated to the M_n interior nodes of $\mathcal{T}_n$ numbered in some fashion, so U_n can be written

$$U_n = \sum_{j=1}^{M_n} \xi_{n,j}\varphi_{n,j},$$

where the coefficients $\xi_{n,j}$ are the nodal values of U_n. We abuse notation to let $\xi_n = \big(\xi_{n,j}\big)$ denote the vector of coefficients. We define the $M_n \times M_n$ *mass matrix* B_n, *stiffness matrix* A_n, and again abusing notation, the $M_n \times 1$ data vector b_n with coefficients

$$(B_n)_{ij} = (\varphi_{n,j}, \varphi_{n,i}),\ (A_n)_{ij} = (\nabla\varphi_{n,j}, \nabla\varphi_{n,i}),\ (b_n)_i = (f, \varphi_{n,i}),$$

for $1 \le i, j \le M_n$. Finally, we define the $M_n \times M_{n-1}$ matrix $B_{n-1,n}$ with coefficients

$$(B_{n-1,n})_{ij} = (\varphi_{n,j}, \varphi_{n-1,i}) \quad 1 \le i \le M_n, 1 \le j \le M_{n-1}. \tag{243.4}$$

The discrete equation for the cG(1)dG(0) approximation on I_n is

$$(B_n + k_nA_n)\xi_n = B_{n-1,n}\xi_{n-1} + b_n. \tag{243.5}$$

The coefficient matrix $B_n + k_nA_n$ of this system is sparse, symmetric, and positive-definite and the system can be solved using a direct or an iterative method.

243.4. Prove that $B_{n-1,n}\xi_{n-1} = B_n\hat{\xi}_{n-1}$ where $\hat{\xi}_{n-1}$ are the coefficients of P_nU_{n-1} with respect to $\{\varphi_{n,j}\}$.

243.5. Specify the matrix equations for the cG(1)dG(1) method. Hint: consider (243.3).

243.6. Assume that $\Omega = (0,1)\times(0,1]$ and the standard uniform triangulation is used on each time step. Compute the coefficient matrix in (243.5).

243.7. (a) Formulate the cG(1)dG(r) with $r = 0, 1$, for the heat equation in one dimension with homogeneous Dirichlet boundary conditions. (b) Write out the matrix equations for the coefficients of U_n in the case of a uniform partition and $r = 0$. (c) Assume that $\mathcal{T}_n$ is obtained by dividing each element of $\mathcal{T}_{n-1}$ into two intervals. Compute $B_{n-1,n}$ explicitly. (d) Repeat (c) assuming that $\mathcal{T}_{n-1}$ has an even number of elements and that $\mathcal{T}_n$ is obtained by joining together every other neighboring pair of elements.

243.8. Repeat Problem 243.7 for the modified heat equation $\dot{u} - \Delta u + u = f$ with homogeneous Neumann boundary conditions.

243.3 The use of quadrature

In general it may be difficult to compute the integrals in (243.5) exactly, and therefore quadrature is often used to compute the integrals approximately. If K denotes an element of $\mathcal{T}_n$ with nodes $N_{K,1}$, $N_{K,2}$, and $N_{K,3}$ and area $|K|$, then we use the lumped mass quadrature for a function $g \in V_n$,

$$Q_K(g) = \frac{1}{3}|K|\sum_{j=1}^{3} g(N_{K,j}) \approx \int_K g(x)\,dx.$$

For the integration in time, we use the midpoint rule,

$$g\left(\frac{t_n + t_{n-1}}{2}\right)k_n \approx \int_{t_{n-1}}^{t_n} g(t)\,dt.$$

We define the approximations $\tilde{B}_n$, $\tilde{B}_{n-1,n}$, and $\tilde{b}_n$ by

$$(\tilde{B}_n)_{ij} = \sum_{K\in\mathcal{T}_n} Q_K(\varphi_{n,i}\varphi_{n,j}),\ (\tilde{B}_{n-1,n})_{ij} = \sum_{K\in\mathcal{T}_n} Q_K(\varphi_{n,i}\varphi_{n-1,j}),$$
$$\text{and } (\tilde{b}_n)_i = \sum_{K\in\mathcal{T}_n} Q_K\big(f(\,\cdot\,,(t_n+t_{n-1})/2)\varphi_{n,i}(\cdot)\big)k_n,$$

for indices in the appropriate ranges. Note that the terms in the sums over $K \in \mathcal{T}_n$ for $\tilde{B}_n$ and $\tilde{B}_{n-1,n}$ are mostly zero, corresponding to the near orthogonality of the nodal basis functions. We find that $\tilde{\xi}_n$, the vector of

nodal values of the cG(1)dG(0) approximation computed using quadrature, satisfies

$$(\tilde{B}_n + k_n A_n)\tilde{\xi}_n = \tilde{B}_{n-1,n}\tilde{U}_{n-1} + \tilde{b}_n. \tag{243.6}$$

If we use the rectangle rule with the right-hand end point of I_n instead, the resulting scheme is called the backward Euler-continuous Galerkin approximation.

243.9. Repeat Problem 243.6 using $\tilde{B}_n$, $\tilde{B}_{n-1,n}$, and $\tilde{b}_n$ instead of B_n, $B_{n-1,n}$, b_n respectively.

243.10. Repeat Problem 243.7 using $\tilde{B}_n$, $\tilde{B}_{n-1,n}$, and $\tilde{b}_n$ instead of B_n, $B_{n-1,n}$, b_n respectively.

243.11. Formulate the cG(1)dG(1) finite element method for the heat equation using the lumped mass quadrature rule in space and the two point Gauss quadrature rule for the time integration over I_n.

243.12. (a) Formulate the cG(1)dG(0) finite element method for the non-constant coefficient heat equation

$$\dot{u}(x,t) - (a(x,t)u'(x,t))' = f(x,t), \quad (x,t) \in (0,1) \times (0,\infty),$$

together with homogeneous Dirichlet boundary conditions and initial data u_0, using lumped mass quadrature rule in space and the midpoint rule in time to evaluate B_n, $B_{n-1,n}$, and any integrals involving a and f. (b) Assuming that $a(x,t) \geq a_0 > 0$ for all x and t, prove the modified mass and stiffness matrices are positive definite and symmetric. (c) Write down the matrix equations explicitly. (d) Assuming that the same space mesh is used for every time step, compute explicit formulas for $\tilde{B}_n$, $\tilde{A}_n$, and $\tilde{b}_n$.

243.4 Error estimates and adaptive error control

In this section, we state a posteriori and a priori error estimates for the cG(1)dG(0) method (243.1) and discuss an adaptive algorithm based on the a posteriori estimate. We also illustrate the performance of the algorithm in an example. The proofs of the error estimates are presented in the next section. For simplicity, we assume that Ω is convex so that the strong stability estimate (240.37) of Lemma 240.6 with stability constant $S = 1$ holds. We also assume that $u_0 \in V_1$; otherwise an additional term accounting for an initial approximation of u_0 appears in the estimates. We define $\tau = \min_n \tau_n$, where τ_n is the minimal angle of $\mathcal{T}_n$.

We begin by stating the a posteriori error estimate including residual errors associated to space discretization, time discretization, and mesh changes between space-time slabs. Here $\|\cdot\|$ denotes the $L_2(\Omega)$-norm and $\|v\|_J = \max_{t\in J} \|v(t)\|$.

Theorem 243.1 *There is a constant C_i only depending on τ such that for $N \geq 1$,*

$$\|u(t_N) - U_N\| \leq L_N C_i \max_{1\leq n\leq N} \big(\|h_n^2 R_2(U)\|_{I_n} + \|h_n^2 f\|_{I_n} + \|[U_{n-1}]\| + \|k_n f\|_{I_n} + \big\|\frac{h_n^2}{k_n}[U_{n-1}]\big\|^* \big),$$

where $u(t_N) = u(\cdot, t_N)$,

$$L_N = 2 + \max_{1\leq n\leq N} \max\left\{ \left(\log(\frac{t_n}{k_n}) \right)^{1/2}, \log(\frac{t_n}{k_n}) \right\},$$

$$R_2(U) = \frac{1}{2} \max_{S\subset\partial K} h_K^{-1} \big|[\partial_S U]\big| \quad \text{on } K \in \mathcal{T}_n,$$

and the starred term is present only if $V_{n-1} \not\subseteq V_n$.

The two first terms on the right of (243.1) measure the residual error of the space discretization with f the contribution from the element interiors (there $\Delta U = 0$), and $R_2(U)$ the contribution from the jumps in the normal derivative $[\partial_S U]$ on elements edges S, cf. Chapter **??**. The next two terms measure the residual error of the time discretization and finally the last term reflects the effect of changing from one mesh to the next. The case $V_{n-1} \not\subseteq V_n$ occurs e.g. when $\mathcal{T}_n$ is obtained from $\mathcal{T}_{n-1}$ by removing some nodes, introducing the L_2-projection $P_n U_{n-1} \in V_n$ of $U_{n-1} \in V_{n-1}$. The starred term is of the same order as the time residual term $\|[U_{n-1}]\|$ if h_n^2/k_n is kept bounded by a moderate constant, which usually may be arranged.

243.13. Draw examples in one space dimension that show a mesh coarsening in which $V_{n-1} \not\subseteq V_n$ and a mesh refinement in which $V_{n-1} \subseteq V_n$.

In the proof of the a priori error estimate, we use the following bounds on the change of mesh size on consecutive slabs. We assume there are positive constants γ_i, with γ_2 sufficiently small, such that for $n = 1, ..., N$,

$$\gamma_1 k_n \leq k_{n+1} \leq \gamma_1^{-1} k_n, \tag{243.7}$$

$$\gamma_1 h_n(x) \leq h_{n+1}(x) \leq \gamma_1^{-1} h_n(x) \quad \text{for } x \in \Omega, \tag{243.8}$$

$$\bar{h}_n^2 \leq \gamma_2 k_n, \tag{243.9}$$

where $\bar{h}_n = \max_{x\in\bar{\Omega}} h_n(x)$, and (243.9) only enters if $V_{n-1} \not\subseteq V_n$. The a priori error estimate reads as follows:

Theorem 243.2 *If Ω is convex and γ_2 sufficiently small, there is a constant C_i depending only on τ and $\gamma_i, i = 1, 2$, such that for $N \geq 1$,*

$$\|u(t_N) - U_N\| \leq C_i L_N \max_{1\leq n\leq N} \left(k_n \|\dot{u}\|_{I_n} + \|h_n^2 D^2 u\|_{I_n} \right). \tag{243.10}$$

243.5 Adaptive error control

The a posteriori error bound can be used to estimate the error of a particular computation and also as the basis of an adaptive algorithm.

Suppose we seek an approximation $U(t)$ satisfying

$$\max_{0\le t\le T} \|u(t) - U(t)\| \le \text{TOL},$$

for a given error tolerance TOL, while using the least amount of computational work. We try to achieve this goal by computing a sequence of triangulations $\{\mathcal{T}_n\}$ and time steps $\{k_n\}$ so that for $n = 1, ..., N$, with $t_N = T$,

$$C_i L_N \max_{1\le n\le N} \big(\|h_n^2 R_2(U_n)\| + \|[U_{n-1}]\| + \|(k_n + h_n^2)f\|_{I_n} + \|h_n^2 k_n^{-1}[U_{n-1}]\|^*\big) = \text{TOL}, \quad (243.11)$$

while the total number of degrees of freedom is minimal. This is a nonlinear constrained minimization problem that we try to solve approximately using an iterative process based on the L_2 equidistribution strategy for elliptic problems described in Chapter **??** and the time step control described in Chapter **??**. From the current time level t_{n-1}, we compute U_n using a predicted time step k_n and predicted mesh size h_n and then we check whether (243.11) holds or not. If not, we compute a new time step k_n and mesh size h_n using (243.11) seeking to balance the error contributions from space and time. It is relatively rare for the error control to require more than a few iterations.

We illustrate the adaptive error control using Femlab. We choose $\Omega = (-1, 1) \times (-1, 1)$ and approximate the solution of (242.3) with forcing

$$f(x,t) = \begin{cases} 10^3, & (x_1 + .5 - t)^2 + (x_2 + .5 - t)^2 < .1, \\ 0, & \text{otherwise}, \end{cases}$$

which in the context of a model of heat flow, amounts to swiping a hot blowtorch diagonally across a square plate. We compute the approximation using TOL=.05 and plot the results at the second, sixth, and tenth time steps in Fig. 243.2-Fig. 243.4. The time steps used are $k_1 \approx .017$, $k_2 \approx .62$, and $k_n \approx .1$ for $n \ge 3$. In Fig. 243.2, we can see the refined region centered around the heated region. At later times, we can see further refinement in the direction that the hot region moves and mesh coarsening in regions which have been passed by. Notice the shape of the refined region and the solution at later times indicating residual heat.

243.14. Implement an error estimation block in a code for the heat equation using the cG(1)dG(0) method. Construct several test problems with known solutions and compare the error bound to the true error.

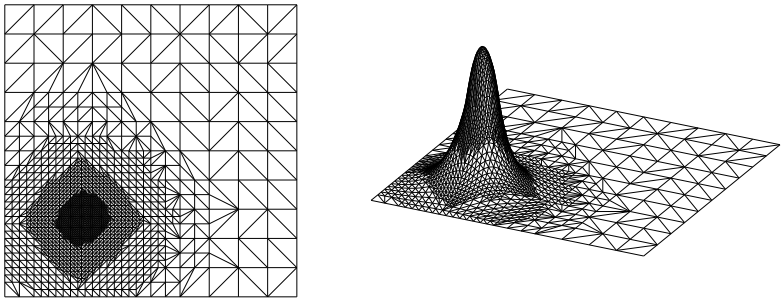


FIGURE 243.2. The approximation and mesh at $t \approx .64$.

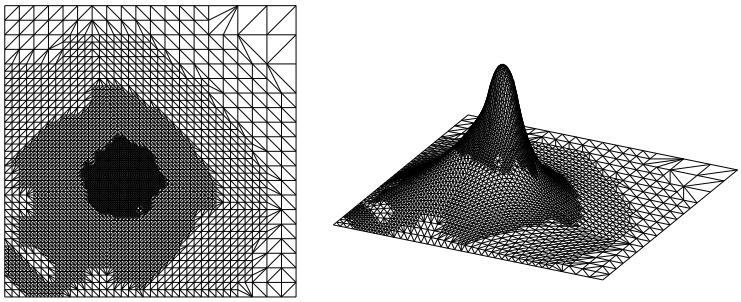


FIGURE 243.3. The approximation and mesh at $t \approx 1.04$.

243.6 A Posteriori Error Analysis

The proofs to follow are based on a combination of the techniques used to prove the error estimates for the parabolic model problem in Chapter **??** and the Poisson problem of Chapter **??**.

Let P_n be the L_2 projection into V_n, and π_k the L_2 projection into the piecewise constants on the time partition $\{t_n\}$, that is, $\pi_k v$ on I_n is the average of v on I_n. We use the following error estimate for P_n which is analogous to the interpolation error estimates discussed in Chapter **??**.

Lemma 243.3 *There is a constant C_i only depending on τ such that if $\varphi = 0$ on Γ, then for all $w \in V_n$,*

$$|(\nabla w, \nabla(\varphi - P_n\varphi))| \leq C_i \|h_n^2 R_2(w)\| \|D^2\varphi\|.$$

In particular, if Ω is convex, then for all $w \in V_n$,

$$|(\nabla w, \nabla(\varphi - P_n\varphi))| \leq C_i \|h^2 R_2(w)\| \|\Delta\varphi\|. \tag{243.12}$$

The proof of this lemma is a little technical, so we put it off until the advanced book. Note that the second estimate follows from the first using (240.37).

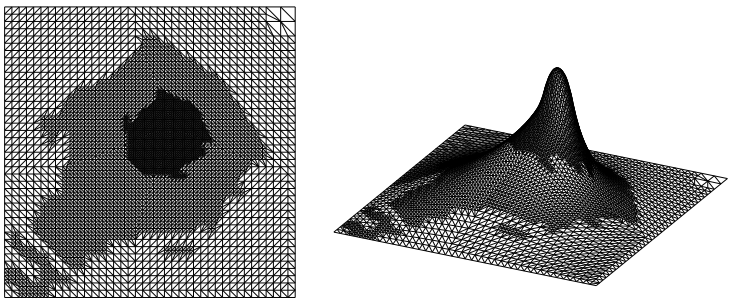


FIGURE 243.4. The approximation and mesh at $t \approx 1.44$.

We introduce the continuous dual problem

$$\begin{cases} -\dot{\varphi} - \Delta\varphi = 0 & \text{in } \Omega \times (0, t_N), \\ \varphi = 0 & \text{on } \Gamma \times (0, t_N), \\ \varphi_N(\cdot, t_N) = e_N & \text{in } \Omega, \end{cases} \tag{243.13}$$

where $e_N = u(t_N) - U_N$. By the definition,

$$\|e_N\|^2 = (e_N, \varphi_N) + \sum_{n=1}^{N} \int_{I_n} (e, -\dot{\varphi} - \Delta\varphi)\, dt,$$

with $e = u - U$ and $\varphi_N = \varphi(\cdot, t_N)$. After integrating by parts in t over each interval I_n and using Green's formula in space, we get

$$\|e_N\|^2 = \sum_{n=1}^{N} \int_{I_n} (\dot{e}, \varphi)\, dt + \sum_{n=1}^{N} \int_{I_n} (\nabla e, \nabla\varphi)\, dt + \sum_{n=1}^{N} ([e_{n-1}], \varphi_{n-1}).$$

Using the facts that $\dot{u} - \Delta u = f$, $[u_n] = 0$, $\dot{U} \equiv 0$ on each I_n, and $U_0^- = u_0$ together with (243.1) with $v = \pi_k P_h \varphi \in W_k^{(0)}$, we obtain the *error representation*:

$$\begin{aligned} \|e_N\|^2 = \sum_{n=1}^{N} \int_{I_n} \big(\nabla U, \nabla(\pi_k P_h \varphi - \varphi)\big)\, dt \\ + \sum_{n=1}^{N} \big([U_{n-1}], (\pi_k P_h \varphi)_{n-1}^+ - \varphi_{n-1}\big) \\ + \int_0^T (f, \varphi - \pi_k P_h \varphi)\, dt = T_1 + T_2 + T_3. \end{aligned}$$

This formula is analogous to the error representation for the model problem studied in Chapter **??**. We now estimate the terms T_1, T_2 and T_3 by

repeatedly using the splitting $\pi_k P_h\varphi - \varphi = (\pi_k - I)P_h\varphi + (P_h - I)\varphi$, where I is the identity, which is a way to split the time and space approximations. First, noting that

$$\int_{I_n} \big(\nabla U, \nabla(\pi_k P_h\varphi - P_h\varphi)\big)\,dt = \int_{I_n} (-\Delta_h U, \pi_k\varphi - \varphi)\,dt = 0, \quad 1 \le n \le N,$$

because U is constant on I_n, the term T_1 reduces to

$$T_1 = \sum_{n=1}^{N} \int_{I_n} \big(\nabla U, \nabla(P_h - I)\varphi\big)\,dt = \sum_{n=1}^{N} \Big(\nabla U_n, \nabla(P_n - I)\int_{I_n} \varphi\,dt\Big).$$

Recalling (243.12), we find that

$$\begin{aligned} |T_1| &\le C_i \sum_{n=1}^{N} \|h_n^2 R_2(U_n)\| \left\|\Delta \int_{I_n} \varphi\,dt\right\| \\ &\le C_i \max_{1\le n\le N} \|h_n^2 R_2(U_n)\| \left(\int_0^{t_{N-1}} \|\Delta\varphi\|\,dt + 2\|\varphi\|_{I_N}\right), \end{aligned}$$

where on the interval I_N, we used the fact that

$$\Delta \int_{I_N} \varphi\,dt = \int_{I_N} \Delta\varphi\,dt = \int_{I_N} \dot\varphi\,dt = \varphi(t_N) - \varphi(t_{N-1}).$$

To estimate T_2, we again use (243.12) to get

$$|([U_{n-1}], (P_n - I)\varphi_{n-1})| \le C_i \|h_n^2 [U_{n-1}]\|^* \|\Delta\varphi_{n-1}\|,$$

where the star is introduced since the left-hand side is zero if $V_{n-1} \subset V_n$. Using the interpolation estimate $\|\varphi_{n-1} - (\pi_k\varphi)_{n-1}^+\| \le \min\{\int_{I_n} \|\dot\varphi\|\,dt, \|\varphi\|_{I_n}\}$ combined with the stability estimate $\|P_n v\| \le \|v\|$, we further have

$$\big|([U_{n-1}], ((\pi_k - I)P_h\varphi)_{n-1}^+)\big| \le \|[U_{n-1}]\| \min\Big\{\int_{I_n} \|\dot\varphi\|\,dt, \|\varphi\|_{I_n}\Big\},$$

and we conclude that

$$\begin{aligned} |T_2| \le C_i \max_{1\le n\le N} \|h_n^2 [U_{n-1}]/k_n\|^* \sum_{n=1}^{N} k_n \|\Delta\varphi_{n-1}\| \\ + \max_{1\le n\le N} \|[U_{n-1}]\| \left(\int_0^{t_{N-1}} \|\dot\varphi\|\,dt + \|\varphi\|_{I_N}\right). \end{aligned}$$

Finally to estimate T_3, we have arguing as in the previous estimates

$$\begin{aligned} \left|\sum_{n=1}^{N} \int_{I_n} (f, P_h\varphi - \pi_k P_h\varphi)\,dt\right| \\ \le \max_{1\le n\le N} \|k_n f\|_{I_n} \left(\int_0^{t_{N-1}} \|\dot\varphi\|\,dt + \|\varphi\|_{I_N}\right), \end{aligned}$$

$$\left|\sum_{n=1}^{N-1}\int_{I_n}(f,(I-P_h)\varphi)\,dt\right| \leq C_i \max_{1\leq n\leq N-1}\|h_n^2 f\|_{I_n}\left(\int_0^{t_{N-1}}\|\Delta\varphi\|\,dt\right),$$

and

$$\left|\int_{I_N}(f,(I-P_h)\varphi)\,dt\right| \leq \|k_N f\|_{I_N}\|\varphi\|_{I_N}.$$

To complete the proof, we bound the different factors involving φ in the estimates above in terms of $\|e_N\|$ using the strong stability estimates (242.19)-(242.21) applied to the dual problem (243.13) with time reversed. We obtain with $w=\dot\varphi=\Delta\varphi$,

$$\begin{aligned}\int_0^{t_{N-1}}\|w\|\,dt &\leq \left(\int_0^{t_{N-1}}(t_N-t)^{-1}\,dt\right)^{1/2}\left(\int_0^{t_N}(t_N-t)\|w\|^2\,dt\right)^{1/2}\\ &\leq \left(\log(\frac{t_N}{k_N})\right)^{1/2}\|e_N\|,\end{aligned}$$

$$\begin{aligned}\sum_{n=1}^{N-1}k_n\|w_{n-1}\| &\leq \sum_{n=1}^{N-1}\frac{k_n}{t_N-t_{n-1}}\|e_N\|\\ &\leq \int_0^{t_{N-1}}(t_N-t)^{-1}\,dt\|e_N\|,\end{aligned}$$

and

$$k_N\|\Delta\varphi_{N-1}\|\leq\|e_N\|.$$

Together, the above estimates prove the a posteriori error estimate.

243.15. (a) Write out the details of the proof in the case of the heat equation in one dimension with $\Omega=(0,1)$ and $r=0$. (b) *(Hard.)* Do the same for $r=1$.

243.16. *(Ambitious.)* Formulate and prove an a posteriori error estimate for the cG(1)dG(0) method that uses the lumped mass and midpoint quadrature rules as described above. Less ambitious is to do the same for the method that uses quadrature only to evaluate integrals involving f.

243.7 A Priori Error Analysis

The a priori analysis follows the same line as the a posteriori analysis, after we introduce a discrete dual problem. The proof of the stability estimate on the solution of the discrete dual problem simplifies if $V_n\subset V_{n-1}$, and in particular, only assumption (243.7) is needed. We present this case below, and leave the general case to a later time.

The discrete strong stability estimate reads.

Lemma 243.4 *Assume that $V_{n-1} \subset V_n$ and that (243.7) holds. Then there is a constant C depending on γ_1 such that the solution U of (243.1) with $f \equiv 0$ satisfies for $N = 1, 2, ...,$*

$$\|U_N\|^2 + 2\sum_{n=1}^{N} \|\nabla U_n\|^2 k_n + \sum_{n=0}^{N-1} \|[U_n]\|^2 = \|U_0\|^2, \tag{243.14}$$

$$\sum_{n=1}^{N} t_n \|\Delta_n U_n\|^2 k_n \leq C\|U_0\|^2, \tag{243.15}$$

and

$$\sum_{n=1}^{N} \|[U_{n-1}]\| \leq C\big(2 + \big(\log(\frac{t_N}{k_1})\big)^{1/2}\big)\|U_0\|. \tag{243.16}$$

Proof: We recall the equation satisfied by U:

$$(I - k_n\Delta_n)U_n = U_{n-1}, \tag{243.17}$$

where we used that $V_{n-1} \subset V_n$. Multiplying by U_n gives

$$\|U_n\|^2 + k_n\|\nabla U_n\|^2 = (U_{n-1}, U_n)$$

or

$$\frac{1}{2}\|U_n\|^2 + \|U_n - U_{n-1}\|^2 + k_n\|\nabla U_n\|^2 = \frac{1}{2}\|U_{n-1}\|^2,$$

which upon summation proves (243.14).

Next, multiplying (243.17) by $-t_n\Delta_n U_n$ gives

$$t_n\|\nabla U_n\|^2 + t_n\|\Delta_n U_n\|^2 k_n = t_n(\nabla U_{n-1}, \nabla U_n),$$

that is

$$\frac{1}{2}t_n\|\nabla U_n\|^2 + t_n\|\nabla(U_n - U_{n-1})\|^2 + t_n\|\Delta_n U_n\|^2 k_n = \frac{1}{2}t_{n-1}\|\nabla U_{n-1}\|^2 + \frac{1}{2}\|\nabla U_{n-1}\|^2 k_n.$$

Summing over $n = 2, ..., N$ using that $k_n \leq \gamma_1 k_{n-1}$ and (243.14) proves (243.15) with the summation starting at $n = 2$. Finally, we note that

$$\sum_{n=2}^{N} \|[U_{n-1}]\| = \sum_{n=2}^{N} \|\Delta_n U_n\| k_n \leq \big(\sum_{n=2}^{N} t_n\|\Delta_n U_n\|^2 k_n\big)^{1/2} \big(\sum_{n=2}^{N} \frac{k_n}{t_n}\big)^{1/2} \leq C\big(\log(\frac{t_N}{k_1})\big)^{1/2} \|U_0\|.$$

The term corresponding to $n = 1$ in (243.15) and (243.16) is estimated using the equation (243.17) with $n = 1$ and the fact that $\|U_1\| \leq \|U_0\|$. This concludes the proof.■.

We can now complete the proof of the a priori error estimate. We first estimate $\|U_n - \tilde{U}_n\|$ where $\tilde{U}_n \in W_{kn}^{(0)}$ is the *average elliptic projection* of u defined for $n = 1, \ldots, N$, by

$$\int_{I_n} (\nabla(u - \tilde{U}_n), \nabla v)\, dt = 0 \quad \text{for all } v \in W_{kn}^{(0)},$$

Using the estimate $\|\tilde{u}_n - \tilde{U}_n\| \leq C_i \|h_n^2 D^2 u\|_{I_n}$ (see Chapter **??**), where $\tilde{u}_n = \pi_k u|_{I_n}$ is the average of u on I_n, together with the obvious estimate $\|u(t_n) - \tilde{u}_n\| \leq k_n \|\dot{u}\|_{I_n}$, we obtain the desired estimate for $\|u_n - U_n\|$.

We let $\Phi \in W_k^{(0)}$ be the solution of the discrete dual problem

$$-(\Phi_{n+1} - \Phi_n) - k_n \Delta_n \Phi_n = 0 \quad \text{for } n = N, ..., 1,$$

where $\Phi_{N+1} = U_N - \tilde{U}_N$. Multiplying by $\tilde{e}_n = U_n - \tilde{U}_n$ and summing over n gives the *error representation*

$$\begin{aligned}\|\tilde{e}_N\|^2 &= (\tilde{e}_N, \Phi_{N+1}) - \sum_{n=1}^{N} (\tilde{e}_n, \Phi_{n+1} - \Phi_n) + \sum_{n=1}^{N} (\nabla \tilde{e}_n, \nabla \Phi_n) k_n \\ &= \sum_{n=1}^{N} (\tilde{e}_n - \tilde{e}_{n-1}, \Phi_n) + \sum_{n=1}^{N} \int_{I_n} (\nabla \tilde{e}_n, \nabla \Phi_n)\, dt,\end{aligned}$$

where we used a summation by parts formula and the assumption that $\tilde{e}_0 = 0$.

243.17. Show that the last formula holds.

Using the fact that for all $v \in W_{kn}^{(0)}$

$$(u_n - U_n - (u_{n-1} - U_{n-1}), v) + \int_{I_n} (\nabla(u - U), \nabla v)\, dt = 0,$$

the error representation takes the form

$$\begin{aligned}\|\tilde{e}_N\|^2 &= \sum_{n=1}^{N} (\rho_n - \rho_{n-1}, \Phi_n) + \sum_{n=1}^{N} \int_{I_n} (\nabla \rho_n, \nabla \Phi_n)\, dt \\ &= \sum_{n=1}^{N} (\rho_n - \rho_{n-1}, \Phi_n)\end{aligned}$$

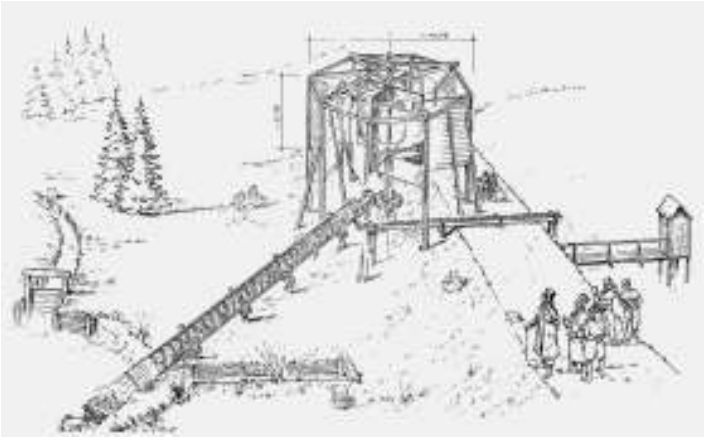
where $\rho = u - \tilde{U}$ and in the last step we used the definition of $\tilde{U}$. Summing by parts again, we get

$$\|\tilde{e}_N\|^2 = -\sum_{n=2}^{N} (\rho_{n-1}, \Phi_n - \Phi_{n-1}) + (\rho_N, \Phi_N),$$

using the assumption that $\rho_0 = 0$. Applying Lemma 243.4 to Φ (after reversing time) proves the desired result. Note that the assumption $V_n \subset V_{n-1}$ of the a priori error estimate corresponds to the assumption $V_{n-1} \subset V_n$ in the stability lemma, because time is reversed.

243.18. Consider the cG(1)dG(1) method for the homogeneous heat equation, i.e. (243.1) with $f \equiv 0$, under the assumption that $k_n \leq Ck_{n-1}$ for some constant C. (a) Show that $\|U_n^-\| \leq \|U_0^-\|$ for all $1 \leq n \leq N$. (b) Show that $\|U\|_{I_n} \leq 5\|U_0^-\|$ for all $1 \leq n \leq N$.

243.19. *(Hard.)* Referring to Problem 243.16, prove the corresponding a priori error estimate.



Proposal for a wind-driven pump by Leibniz

244
Wave Equation Analysis

244.1 Introduction

The wave equation is a basic prototype of a hyperbolic partial differential equation, and models propagation of different types of waves such as elastic waves in an elastic string, membrane, or solid, sound waves in a gas or fluid, or electromagnetic waves. The simplest model of wave propagation is an equation for transport in one direction, which we derive in the next section. After that, we derive the wave equation by examining the familiar model of the motion of a discrete system of masses and springs in the limit as the number of masses increases. We then recall some of the properties of solutions of the wave equation; contrasting their behavior to that of solutions of the heat equation, which is the other basic example of a time dependent partial differential equation. We continue with a discussion of the wave equation in higher dimensions, emphasizing the important fact that the behavior of solutions of the wave equation depends on the dimension. Finally, we discuss the approximate solution of the wave equation using a Galerkin finite element method.

244.2 Transport in 1D

The simplest model for wave propagation is in fact the simplest of all partial differential equations. We model the convective transport of a pollutant suspended in water that is flowing at constant speed c through a pipe of

uniform cross section assuming that there is no diffusion of the pollutant. We illustrate this in Fig. 244.1. Letting $u(x,t)$ denote the concentration of

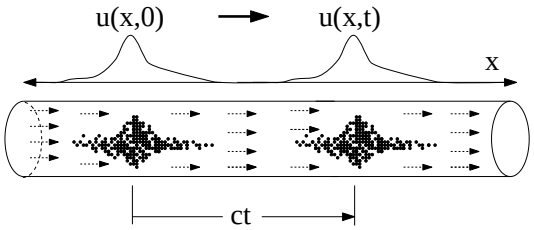


FIGURE 244.1. The transport of a pollutant suspended in a fluid flowing in a pipe.

the pollutant at the point x in the pipe at time t, the conservation of mass can be formulated in terms of integrals as

$$\int_0^{\bar{x}} u(x,t)\,dx = \int_{c(\bar{t}-t)}^{\bar{x}+c(\bar{t}-t)} u(x,\bar{t})\,dx \quad \text{for } \bar{x} > 0,\ \bar{t} \geq t.$$

This equation states that the amount of pollutant in the portion of the fluid occupying $[0,\bar{x}]$ at time t and $[c(\bar{t}-t), \bar{x}+c(\bar{t}-t)]$ at time $\bar{t}$ is the same. To obtain a differential equation expressing the conservation of mass, we first differentiate with respect to $\bar{x}$ to get $u(\bar{x},t) = u(\bar{x}+c(\bar{t}-t),\bar{t})$ and then differentiate with respect to $\bar{t}$ (or t) to get $0 = cu'(x,t) + \dot{u}(x,t)$, after letting $\bar{t} \to t$ and $\bar{x} \to x$.

Assuming that the pipe is infinitely long in order to avoid having to deal with what happens at the ends, we obtain the initial value problem: Find $u(x,t)$ such that

$$\begin{cases} \dot{u}(x,t) + cu'(x,t) = 0 & \text{for } x \in \mathbb{R},\ t > 0, \\ u(x,0) = u_0(x) & \text{for } x \in \mathbb{R}, \end{cases} \tag{244.1}$$

where c is a constant. The solution is $u(x,t) = u_0(x-ct)$, which simply says that the solution at time t is the initial data u_0 translated a distance ct. The line $x - ct = \xi$ is called a *characteristic* line and c is called the *speed*. Since the value of the solution is constant, namely $u_0(\xi)$, at all points along the characteristic, we say that information travels along characteristics.

244.1. (a) Verify this formula. (b) Plot the solution corresponding to $u_0(x) = \sin(x)$ at times $t = 0, \pi/4, \pi/3, \pi/2$, and $23\pi/2$.

The transport problem (244.1) is the basic model of wave propagation. Below, we will see that the wave equation, which describes the propagation of vibrations in an elastic string, can be written as a system of transport

equations. We will also meet the scalar transport model in the context of convection-diffusion problems in Chapter **??**, where we consider the additional effect of diffusion.

We point out an interesting fact: the solution formula $u(x,t) = u_0(x-ct)$ is defined even if u_0 is discontinuous, though in this case, u obviously doesn't satisfy the differential equation at every point. Such initial data corresponds to a sharp signal, for example turning a light switch on and off. We can use the variational formulation of (244.1) to make sense of the solution formula when the data is nonsmooth, and we pick this up again later.

244.2. Plot the solution corresponding to $u_0(x) = \sin(x)$ for $0 \le x \le \pi$ and 0 otherwise at times $t = 0, \pi/4, \pi/3, \pi/2$, and $23\pi/2$.

One important difference between parabolic equations like the heat equation and hyperbolic equations like the transport and wave equations lies in the treatment of boundaries. It is natural to consider the transport equation with a boundary condition posed on the *inflow* boundary. If $c > 0$, then the inflow boundary is on the left. Choosing the boundary to be at $x = 0$ arbitrarily, we obtain

$$\begin{cases} \dot{u}(x,t) + cu'(x,t) = 0 & \text{for } x > 0,\ t > 0, \\ u(0,t) = g(t) & \text{for } t > 0, \\ u(x,0) = u_0(x) & \text{for } x > 0, \end{cases} \tag{244.2}$$

where c is constant and $g(t)$ gives the inflow of material. By direct computation, we can verify that the solution satisfies

$$u(x,t) = \begin{cases} g(t - x/c), & x - ct \le 0, \\ u_0(x - ct), & x - ct > 0 \end{cases}$$

and we illustrate this in Fig. 244.2.

244.3. (a) Plot the solution of (244.2) for $u_0(x) = \sin(x)$ for $0 < x < \pi$ and 0 otherwise and $g(t) = t$ at $t = 0, \pi/6, \pi/4, \pi/3$, and $\pi/2$. (b) What does such boundary conditions mean interpreted in terms of the transport of a pollutant down a pipe?

244.4. Show that the solution of (244.2) for g given for $t \ge 0$ and u_0 given for $x > 0$ agrees with the solution of (244.1) corresponding to initial data $\bar{u}_0$ defined so that $\bar{u}_0(x) = u_0(x)$ for $x > 0$ and $\bar{u}_0(x) = g(-x/c)$ for $x \le 0$ in the region $x \ge 0$, $t \ge 0$.

244.5. Find a formula for the solution of the homogeneous wave equation posed with a boundary condition on the left at a point x_0.

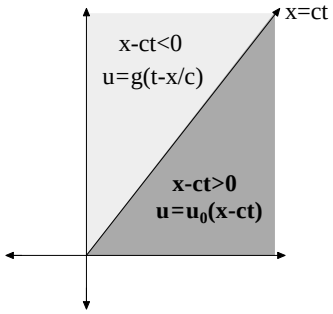


FIGURE 244.2. Solving the transport equation with a boundary condition on the inflow boundary when $c > 0$.

Note that once again the solution formula holds even though it may imply that u is discontinuous across the line $x = ct$. We can resolve this difficulty using the variational formulation as well.

The value of the solution at any *outflow boundary*, which is located on the right when $c > 0$, is determined from the initial data and therefore we cannot impose arbitrary values for the solution on an outflow boundary. In general, a hyperbolic problem posed on a finite domain may have inflow, outflow, or both kinds of boundaries and this is an important consideration in the design of numerical methods. This is a sharp contrast to the situation with the heat equation.

244.3 Wave Equation in 1D

We begin by describing a physical system consisting of N weights each of mass m joined by $N+1$ springs with equal length and spring constant. We choose coordinates so that the system occupies the interval $(0, 1)$ and assume that the springs at the ends are fixed and the masses are constrained to move horizontally along the x axis without friction. The rest position of the n'th weight is nh with $h = 1/(N+1)$. We let $u_n(t)$ denote the displacement of the n'th weight from the rest position with $u_n > 0$ representing a displacement to the right. We illustrate this in Fig. 244.3. Below, we

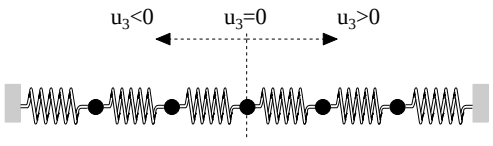


FIGURE 244.3. The coordinate system for a system of masses and springs.

want to compare the motion of systems with different numbers of weights but totalling the same mass. Hence, we assume that $m = h$, so that as N increases, the total mass of the system tends to one.

Hamilton's principle states that the Lagrangian of the system, which is equal to the difference between the kinetic and potential energies integrated over an arbitrary time interval (t_1, t_2),

$$\int_{t_1}^{t_2} \left(\sum_{n=1}^{N} \frac{m}{2} (\dot{u}_n)^2 - \sum_{n=1}^{N+1} \frac{1}{2} h^{-1} (u_n - u_{n-1})^2 \right) dt,$$

where we set $u_0 = 0$ and $u_{N+1} = 0$, is stationary at the trajectory followed by the system. We assume that the spring constant is $1/h$, since it should scale with the length of the springs.

To obtain the differential equation for $u = (u_n(t))$, we add an arbitrary small perturbation to u_n in the direction of $v = (v_n)$, with $v_0 = v_{N+1} = 0$, to get $u_n + \epsilon v_n$ for $\epsilon \in \mathbb{R}$. Differentiating with respect to ϵ and setting the derivative equal to zero for $\epsilon = 0$, which corresponds to the Lagrangian being stationary at the solution u, and then varying v gives the following system

$$\ddot{u}_n - h^{-2}(u_{n-1} - 2u_n + u_{n+1}) = 0, \quad t > 0,\ n = 1, ..., N, \tag{244.3}$$

where $u_0 = 0$ and $u_{N+1} = 0$. The differential equation (244.3) is supplemented by initial conditions specifying the initial position and velocity of each weight.

We present an example with $N = 5$ in which the n'th weight is displaced a distance $.5h \sin(nh)$ to the right of the rest position and the initial speed is zero. We solve the system (244.3) using the Cards code keeping the error below .06. We show the position of the weights for a few times in Fig. 244.4.

244.6. (a) Derive (244.3). (b) Change the system of equations (244.3) into a first order system by introducing new unknowns $v_n = \dot{u}_n$. (c) Solve the system keeping the error below .05 for $N = 5, 10, 15, ..., 55$ and compare the solutions. (d) Compute the solution for $N = 5$ where the masses start at the rest position with initial velocities $\{\sin(nh)\}$ and plot the results for $t = 0, .25, .5, .75, 1.0$ and 1.25.

Letting the number of weights N tend to infinity (with a corresponding decrease in the mass of each weight since $m = h$) in the discrete equation (244.3), we formally obtain the wave equation in one dimension:

$$\begin{cases} \ddot{u}(x,t) - u''(x,t) = 0 & \text{for } 0 < x < 1 \text{ and } t > 0, \\ u(0,t) = u(1,t) = 0 & \text{for } t > 0, \\ u(x,0) = u_0(x), \dot{u}(x,0) = \dot{u}_0(x) & \text{for } 0 < x < 1, \end{cases} \tag{244.4}$$

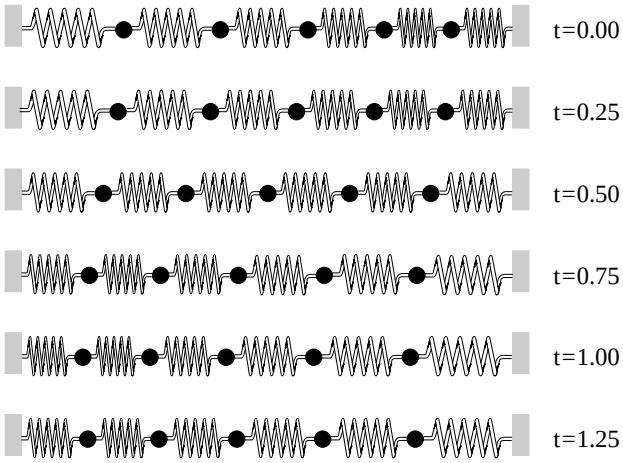


FIGURE 244.4. The evolution of the discrete system of masses and springs.

where now with abuse of notation u_0 and $\dot{u}_0$ are given initial data. This is the initial value problem describing the longitudinal vibrations in an elastic string. It turns out that the same equation describes also the transversal vibration of an elastic string, like a string on a guitar.

244.4 Sound Waves in a Tube

The wave equation (244.4) is also used to model the propagation of sound waves. We consider a long thin tube, represented by $\mathbb{R}$, filled with gas of density ρ, pressure p, and velocity u. The behavior of the gas is described by a set of nonlinear equations that result from the conservation of mass and Newton's law relating the rate of change of momentum to the pressure:

$$\begin{cases} \dot{\rho} + (u\rho)' = 0 & \text{in } \mathbb{R} \times (0,\infty), \\ \dot{m} + (um)' + p' = 0 & \text{in } \mathbb{R} \times (0,\infty), \end{cases} \tag{244.5}$$

where $m = \rho u$ is the momentum. To derive a linear equation, we consider small fluctuations $\bar{\rho}$, $\bar{u}$ and $\bar{p}$ around a constant state of density ρ_0, pressure p_0 and zero velocity, so that $\rho = \rho_0 + \bar{\rho}$, $p = p_0 + \bar{p}$ and $u = 0 + \bar{u}$. We assume that $\bar{p} = c^2\bar{\rho}$, where c is a constant representing the speed of sound, substitute the new variables into (244.5), and drop quadratic terms in the resulting equation, since these are very small if the fluctuations are small, to obtain

$$\begin{cases} \dot{\bar{\rho}} + \rho_0\bar{u}' = 0 & \text{in } \mathbb{R} \times (0,\infty), \\ \rho_0\dot{\bar{u}} + c^2\bar{\rho}' = 0 & \text{in } \mathbb{R} \times (0,\infty). \end{cases} \tag{244.6}$$

Eliminating either $\bar{\rho}$ or $\bar{p}$ leads to the wave equations $\ddot{\bar{\rho}} - c^2\bar{\rho}'' = 0$ and $\ddot{\bar{p}} - c^2\bar{p}'' = 0$.

244.7. (a) Verify the derivation of (244.6). (b) Show that (244.6) implies that $\bar{\rho}$ and $\bar{p}$ satisfy the wave equation under the assumptions of the derivation.

244.5 Structure of Solutions: d'Alembert's Formula

The general initial value problem for the wave equation,

$$\begin{cases} \ddot{u} - u'' = f & \text{in } \mathbb{R} \times (0, \infty), \\ u(x, 0) = u_0(x), \dot{u}(x, 0) = \dot{u}_0(x) & \text{for } x \in \mathbb{R}, \end{cases} \tag{244.7}$$

can be written as a system of transport equations by introducing the variable $w = \dot{u} - u'$ to get

$$\begin{cases} \dot{w} + w' = f & \text{in } \mathbb{R} \times (0, \infty), \\ \dot{u} - u' = w & \text{in } \mathbb{R} \times (0, \infty), \\ w(x, 0) = \dot{u}_0(x) - u_0'(x), u(x, 0) = u_0(x) & \text{for } x \in \mathbb{R}, \end{cases}$$

where the two transport equations in the new formulation correspond to transport of signals in opposite directions with speed one.

244.8. Verify that the two problems have the same solution u.

It is therefore natural, following d'Alembert and Euler, to look for a solution $u(x, t)$ of (244.7) with $f \equiv 0$ of the form $u(x, t) = \varphi(x-t)+\psi(x+t)$, where $\varphi(x - t)$ corresponds to a wave propagating in the positive direction with speed one and $\psi(x+t)$ corresponds to a wave propagating with speed one in the negative direction. It is easy to see that a function of this form satisfies the wave equation $\ddot{u} - u'' = 0$.

244.9. Verify this claim.

Determining the functions φ and ψ from the initial conditions, we find *d'Alembert's formula*:

$$u(x, t) = \frac{1}{2}(u_0(x - t) + u_0(x + t)) + \frac{1}{2}\int_{x-t}^{x+t} \dot{u}_0(y)\, dy. \tag{244.8}$$

244.10. Prove (244.8).

244.11. If the speed of the propagation of the waves is $c > 0$, then the corresponding wave equation takes the form $\ddot{u}-c^2u'' = 0$. Derive d'Alembert's formula for this case. Hint: seek a solution of the form $u(x, t) = \varphi(x - ct) + \psi(x + ct)$.

Using d'Alembert's formula, we can study the dependence of the solution on the initial data. For example, if $u_0(x)$ is an approximate "point" source supported in a small interval around $x = 0$ and $\dot{u}_0 \equiv 0$, then the solution $u(x,t)$ consists of two pulses propagating from $x = 0$ in the positive and negative directions with speed ± 1, see Fig. 244.5. This data corresponds to an elastic string being released at time zero with a displacement concentrated at 0 and with zero velocity. The d'Alembert formula shows that the solution $u(x,t)$ at a given time t is influenced only by the value of the initial data $u_0(x)$ at the points $x \pm t$, i.e. as for the transport equation, there is *sharp propagation* of the initial data u_0. The effect of an initial

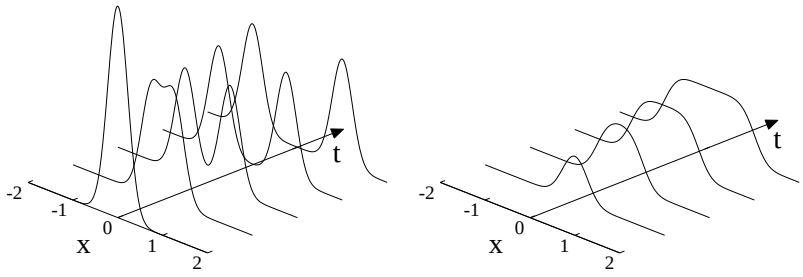


FIGURE 244.5. The evolution of solutions of the wave equation corresponding to an approximate "point" source in u_0 together with $\dot{u}_0 \equiv 0$ on the left and an approximate "point" source in $\dot{u}_0$ together with $u_0 \equiv 0$ on the right.

impulse in the derivative data $\dot{u}_0$ is different, as illustrated in Fig. 244.5. If $\dot{u}_0$ has support in a small interval centered at $x = 0$ and $u_0 \equiv 0$ then $u(x,t)$ is constant in most of the region $[x - t, x + t]$ and zero outside a slightly larger interval.

244.12. Define $g(x) = 10^8(x - .1)^4(x + .1)^4$ if $|x| < .1$ and 0 otherwise and show that g has continuous second derivatives. (a) Compute an explicit formula for the solution if $u_0(x) = g(x)$ and $\dot{u}_0 \equiv 0$ and plot the results for a few times. (b) Do the same if $u_0 \equiv 0$ and $\dot{u}_0(x) = g(x)$. (c) Referring to (b), given $t > 0$, determine the intervals on which u is constant.

The extension of the d'Alembert's formula to the nonhomogeneous problem (244.7) with $f \neq 0$ is

$$u(x,t) = \frac{1}{2}\big(u_0(x+t) + u_0(x-t)\big) + \frac{1}{2}\int_{x-t}^{x+t} \dot{u}_0(y)\,dy + \frac{1}{2}\iint\limits_{\Delta(x,t)} f(y,s)\,dy\,ds, \quad (244.9)$$

where $\Delta = \Delta(x,t) = \{(y,s) : |x - y| \leq t - s,\ s \geq 0\}$ denotes the *triangle of dependence* indicating the portion of space-time where data can influence

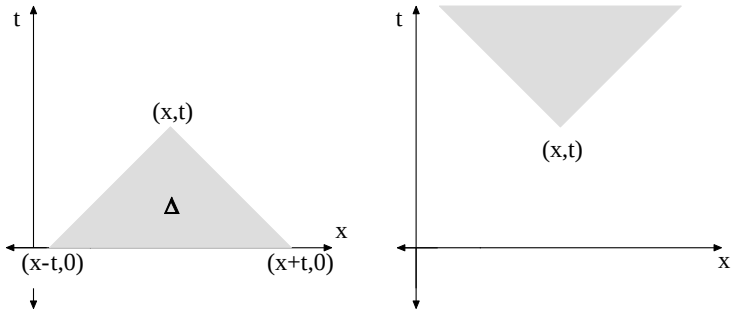


FIGURE 244.6. On the left, we show the triangle of dependence Δ of the point (x,t). On the right, we show the triangle of influence.

the value of the solution at the point (x,t), see Fig. 244.6. Turning the triangle of dependence upside-down gives the *triangle of influence* $\{(y,s) : |x-y| \le s-t\}$ indicating the points (y,s) which can be influenced by the values of the data at (x,t).

244.13. Prove (244.9).

We can handle problems with boundaries by modifying d'Alembert's formula. For example, to find a formula for the homogeneous wave equation $\ddot{u} - u'' = 0$ for $x > 0$, $t > 0$ together with the boundary condition $u(0,t) = 0$ for $t > 0$ and initial conditions $u_0(x)$ and $\dot{u}_0(x)$ as above, we use d'Alembert's formula for the solution of the wave equation $\ddot{w} - w'' = 0$ on $\mathbb{R} \times (0,\infty)$ together with odd initial data w_0 and $\dot{w}_0$, where w_0 is defined by

$$\bar{w}_0(x) = \begin{cases} -u_0(-x), & x < 0, \\ 0, & x = 0, \\ u_0(x), & x > 0, \end{cases}$$

and $\dot{w}_0$ is defined similarly. It is easy to verify that the solutions of the two problems agree in the region $x > 0$, $t > 0$. Using d'Alembert's formula and tracing the characteristic lines to their intersections with the x axis, see Fig. 244.7, we find that

$$u(x,t) = \begin{cases} \frac{1}{2}\big(u_0(x+t) + u_0(x-t)\big) + \frac{1}{2}\int_{x-t}^{x+t} \dot{u}_0(y)\,dy, & x > t \\ \frac{1}{2}\big(u_0(t+x) - u_0(t-x)\big) + \frac{1}{2}\int_{t-x}^{t+x} \dot{u}_0(y)\,dy, & x \le t. \end{cases} \tag{244.10}$$

244.14. (a) Verify (244.10). (b) Find a formula for the solution of the homogeneous wave equation posed with the Neumann boundary condition $u'(0,t) = 0$. Hint: extend the data to be even functions on $\mathbb{R}$.

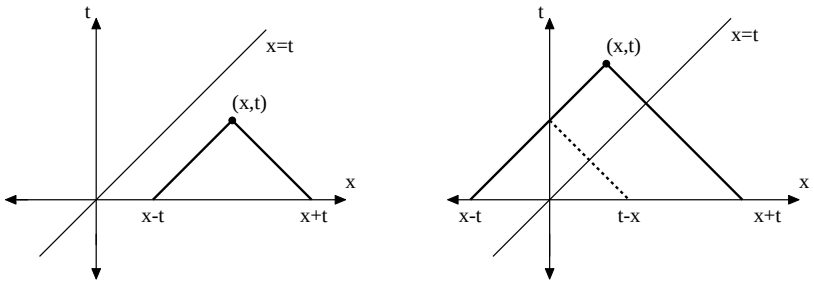


FIGURE 244.7. The two cases for applying d'Alembert's formula to the wave equation posed with a boundary at $x = 0$. We plot the characteristic lines for (x, t) with $x > t$ on the left and $x < t$ on the right. Note the reflection in the t axis of the point $x - t$ to $t - x$.

244.15. Use d'Alembert's formula to construct the solution of the homogeneous wave equation posed on $(0, 1)$ with periodic boundary conditions.

244.16. Give a d'Alembert solution formula for the vibrating string problem (244.4). Hint: extend u_0 and $\dot{u}_0$ to be functions on $\mathbb{R}$.

The existence of the triangles of dependence and influence and the sharp propagation of the data are the result of the *finite speed of propagation* of solutions of the wave equation. This contrasts to the behavior of solutions of the heat equation, where the value of the solution at one point depends on the data at every point (although the exponential decay of the fundamental solution implies that the dependence is very small from point far away) and the diffusion of the data as time passes. One consequence is that it is more difficult to send recognizable signals by heating a conducting wire than sending sound waves down a pipe.

244.6 Separation of Variables and Fourier's Method

The technique of separation of variables and Fourier's method can be used to write the solution of the wave equation as a Fourier series. To simplify the notation, we pose (244.4) on $(0, \pi)$ instead of $(0, 1)$. In this case, the solution is

$$u(x,t) = \sum_{n=1}^{\infty} (a_n \sin(nt) + b_n \cos(nt)) \sin(nx), \tag{244.11}$$

where the coefficients a_n and b_n are determined from the Fourier series of the initial conditions:

$$u_0(x) = \sum_{n=1}^{\infty} b_n \sin(nx), \quad \dot{u}_0(x) = \sum_{n=1}^{\infty} n a_n \sin(nx).$$

Note that the time factor, $a_n \sin(nt) + b_n \cos(nt)$, in the Fourier series of the solution of the wave equation does not decrease exponentially as time increases like the corresponding factor in the Fourier series of a solution of the heat equation. Therefore, the solution of the wave equation generally does not become smoother as time passes.

244.17. Verify the solution formula (244.11) formally.

244.18. Compute the solution for (a) $u_0(x) = x(\pi - x)$, $\dot{u}_0(x) \equiv 0$, (b) $\dot{u}_0(x) = x(\pi - x)$, $u_0(x) \equiv 0$.

244.7 Conservation of Energy

We saw that a solution of the heat equation tends to dissipate as time passes, with a corresponding decrease in the energy. In contrast, the total energy (the sum of kinetic and potential energies) of the solution u of the homogeneous wave equation (244.4) remains constant in time:

$$\|\dot{u}(\cdot,t)\|^2 + \|u'(\cdot,t)\|^2 = \|\dot{u}_0\|^2 + \|u_0'\|^2 \quad \text{for } t \geq 0,$$

where $\|\cdot\|$ denotes the $L_2(0,1)$ norm as usual. To prove this, we multiply (244.12) by $2\dot{u}$, integrate over $(0,1)$, and then integrate by parts to get

$$0 = \frac{\partial}{\partial t}\left(\int_0^1 \left(\dot{u}(x,t)^2 + u'(x,t)^2 \right) dx \right).$$

244.19. Provide the details of this derivation.

244.20. (a) Show that the only solution of (244.4) with $u_0 \equiv \dot{u}_0 \equiv 0$ is $u \equiv 0$. (b) Suppose that w solves (244.4) with initial data w_0 and $\dot{w}_0$. Estimate $u - w$, where u solves (244.4).

244.8 The wave equation in higher dimensions

Situations modelled by the wave equation in higher dimensions include the vibrations of a drum head and the propagation of sound waves in a volume

of gas. Letting Ω denote a domain in $\mathbb{R}^d$, $d = 2$ or 3, with boundary Γ, the initial-boundary value problem for the wave equation is

$$\begin{cases} \ddot{u} - \Delta u = f & \text{in } \Omega \times (0,\infty), \\ u = 0 & \text{on } \Gamma \times (0,\infty), \\ u(x,0) = u_0(x),\ \dot{u}(x,0) = \dot{u}_0(x) & \text{for } x \in \Omega, \end{cases} \tag{244.12}$$

where f, u_0, and $\dot{u}_0$ are given functions. The wave equation is also posed on all of $\mathbb{R}^d$ in some models.

Before turning to the approximation of (244.12), we recall some of the properties of the solutions. We emphasize the important fact that the behavior of solutions of the wave equation depends on the dimension, and in particular, the behavior in two dimensions is significantly different that in three dimensions.

244.9 Symmetric Waves

We begin by considering solutions of the homogeneous wave equation in $\mathbb{R}^d$ that are symmetric through the origin since this effectively reduces the problem to one dimension in space. In $\mathbb{R}^3$, these are called *spherically symmetric* waves. For simplicity, we assume that $\dot{u}_0 \equiv 0$. The wave equation (244.12) in spherical coordinates, assuming the solution depends only on r, i.e. the distance to the origin, reads

$$\ddot{u} - u_{rr} - \frac{2}{r}u_r = 0 \quad \text{for } r > 0,\ t > 0, \tag{244.13}$$

where $u_r = \partial u/\partial r$. Note the important factor two in the third term; by introducing the new unknown $v = ru$, this equation transforms into the one-dimensional wave equation,

$$\ddot{v} - v_{rr} = 0 \quad \text{for } r > 0,\ t > 0. \tag{244.14}$$

This equation is posed together with the boundary condition $v(0,t) = 0$ for $t > 0$ and initial conditions $v(r,0) = ru_0(r)$ and $\dot{v}(r,0) = 0$ for $r > 0$. Using (244.10) to write a formula for v and then changing back to u, we find that

$$u(r,t) = \frac{1}{2}\begin{cases} \big(u_0(r+t) + u_0(r-t)\big) + \dfrac{t}{r}\big(u_0(r+t) - u_0(r-t)\big), & r \geq t, \\ \big(u_0(t+r) + u_0(t-r)\big) + \dfrac{t}{r}\big(u_0(t+r) - u_0(t-r)\big), & r < t, \end{cases} \tag{244.15}$$

where we take $u(0,\cdot) = \lim_{r\to 0^+} u(r,\cdot)$. From this, we conclude that the initial data propagates sharply outwards in the positive r direction as time passes. In particular, if u_0 has support in the ball $\{x : |x| \le \rho\}$ for some $\rho > 0$, then at any point x with $x > \rho$, $u(x,t)$ is zero for $t < x - \rho$, then the solution is non-zero with values determined by u_0 for 2ρ time units, and finally after that the solution is once again zero.

244.21. Compute explicit formulas for the spherically symmetric solution corresponding to $u_0 \equiv 1$ for $|x| \le 1$ and 0 otherwise and $\dot{u}_0 \equiv 0$. Hint: there are six regions in the (r,t) plane that have to be considered. Plot the solution as a function of the radius at several times.

We can also look for symmetric solutions of the wave equation in $\mathbb{R}^2$. Unfortunately in this case, the wave equation reduces to

$$\ddot{u} - u_{rr} - \frac{1}{r}u_r = 0 \quad \text{for } r > 0,\ t > 0, \tag{244.16}$$

and there is no simple change of variables that reduces this problem to the wave equation in one dimension.

244.22. Verify (244.13), (244.16), and (244.14).

244.23. (a) Verify (244.15). (b) Treat the problem where $\dot{u}_0$ is not assumed to be zero.

244.10 Finite Speed of Propagation

As suggested by the spherically symmetric case, there is a finite speed of propagation of information in solutions of the wave equation in higher dimensions. By this, we mean that the value of $u(x,t)$ depends only on the values of the data given in the *cone of dependence*

$$\Delta(x,t) := \{(y,s) \in \mathbb{R}^d \times \mathbb{R} : |y - x| \le t - s, s \ge 0\}.$$

The cone of dependence is the multi-dimensional counterpart to the triangle of dependence. Specifically, for any proper subdomain ω of Ω, we may define the enlarged region $\omega(t) = \{x \in \mathbb{R}^d : \mathrm{dist}(x,\omega) < t\}$ assuming for simplicity that $t \ge 0$ is not too large so that $\omega(t)$ is also contained in Ω, see Fig. 244.8. Then we prove the following estimate on the value of u in ω at time t in terms of the values of the data in $\omega(t)$:

Theorem 244.1 *For any proper subdomain ω of Ω and $t > 0$ such that $\omega(t) \subset \Omega$, the solution u of the homogeneous wave equation satisfies*

$$\|\dot{u}(\cdot,t)\|^2_{L_2(\omega)} + \|\nabla u(\cdot,t)\|^2_{L_2(\omega)} \le \|\dot{u}_0\|^2_{L_2(\omega(t))} + \|\nabla u_0\|^2_{L_2(\omega(t))}.$$

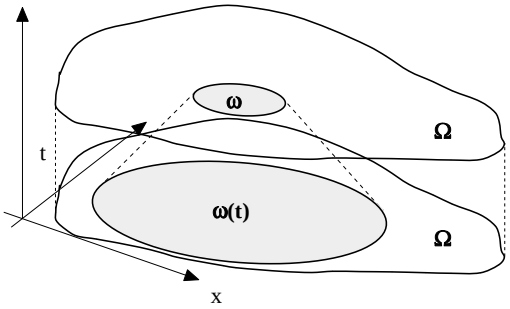


FIGURE 244.8. The generalized cone of dependence $\Delta(\omega, t)$ and enlarged region $\omega(t)$ associated to a subdomain ω of Ω.

Proof: We define the generalized cone of dependence $\Delta = \Delta(\omega, t) = \{\cup_{x\in\omega}\Delta(x, t)\}$, which is the union of all the cones of dependence $\Delta(x, t)$ with $x \in \omega$. Under the assumption on t, Δ is contained in the cylinder $\bar{\Omega} \times [0, t]$. We denote the exterior unit space-time normal to the boundary S of Δ by $n = (n_x, n_t)$, where n_x denotes the space components of n. To obtain the desired estimate, we multiply (244.12) by $2\dot{u}$, integrate over Δ, and then integrate by parts to obtain

$$\begin{aligned}
0 &= \int_\Delta (\ddot{u} - \Delta u) 2\dot{u}\, dx\, dt \\
&= \int_\Delta (2\ddot{u}\dot{u} + 2\nabla u \cdot \nabla \dot{u})\, dx\, dt - \int_S n_x \cdot \nabla u 2\dot{u}\, ds \\
&= \int_\Delta \frac{d}{dt}((\dot{u})^2 + |\nabla u|^2)\, dx\, dt - \int_S n_x \cdot \nabla u 2\dot{u}\, ds \\
&= \int_S n_t((\dot{u})^2 + |\nabla u|^2)\, ds - \int_S n_x \cdot \nabla u 2\dot{u}\, ds.
\end{aligned}$$

On the "sloping" sides of S, we have $n_t = |n_x| = 1/\sqrt{2}$ and thus by Cauchy's inequality, $n_t((\dot{u})^2 + |\nabla u|^2) - n_x \cdot \nabla u 2\dot{u} \geq 0$. We can therefore estimate the integral over the top part of S (with $n_t = 1$ and $n_x = 0$) corresponding to ω, in terms of the integral over the base of S corresponding to $\omega(t)$, and thus obtain the desired result. ■.

244.24. Write out the details of the last estimate.

244.25. Derive a version of Theorem 244.1 for the solution of (244.12) with $f \equiv 0$ without the restriction on t that keeps Δ inside the cylinder $\bar{\Omega} \times [0, t]$. Hint: define a generalized cone that includes part of the boundary of $\bar{\Omega} \times [0, t]$ when t is large.

244.26. Generalize the result of Lemma 244.1 to the case $f \neq 0$.

244.11 Conservation of Energy

Along with a finite speed of propagation, a solution u of (244.12) with $f = 0$ satisfies

$$\|\dot{u}(\cdot,t)\|^2 + \|\nabla u(\cdot,t)\|^2 = \|\dot{u}_0\|^2 + \|\nabla u_0\|^2, \quad t > 0,$$

where $\|\cdot\|$ denotes the $L_2(\Omega)$ norm.

244.27. Prove this by modifying the proof in one dimension.

244.11.1 Kirchhoff's formula and Huygens' principle

The generalization of d'Alembert's solution formula to the homogeneous wave equation (244.12) with $f = 0$ and $\Omega = \mathbb{R}^3$ is called *Kirchhoff's formula*, and was first derived by Poisson,

$$u(x,t) = \frac{1}{4\pi t} \int\limits_{S(x,t)} \dot{u}_0 \, ds + \frac{\partial}{\partial t}\left(\frac{1}{4\pi t} \int\limits_{S(x,t)} u_0 \, ds \right), \tag{244.17}$$

where $S(x,t) = \{y \in \mathbb{R}^3 : |y - x| = t\}$ is the sphere with radius t centered at x. This formula shows sharp propagation at speed one of both the initial data u_0 and $\dot{u}_0$, since the integrals involve only the surface $S(x,t)$ of the ball $B_3(x,t) = \{y \in \mathbb{R}^3 : |y - x| \le t\}$, which is the set of points in $\mathbb{R}^3$ from which a signal of speed one may reach x within the time t. In other words, only the values of the data on the surface of the cone of dependence actually have an influence on the value at a point. The sharp wave propagation in three dimensions is referred to as *Huygens' principle*.

244.28. Use (244.17) to write a formula for the solution of the wave equation with the data used in Problem 244.21.

A formula for the solution of the wave equation in two dimensions can be derived from (244.17) by considering the function to be a solution of the wave equation in three dimensions that happens to be independent of x_3. For $x \in \mathbb{R}^2$, we let $B_2(x,t) = \{y \in \mathbb{R}^2 : |y - x| \le t\}$, which may be thought of as the projection of $B_3(x,t)$ onto the plane $\{x : x_3 = 0\}$. The solution is given by

$$u(x,t) = \frac{1}{2\pi} \int\limits_{B_2(x,t)} \frac{\dot{u}_0(y)}{\left(t^2 - |y-x|^2\right)^{1/2}} \, dy + \frac{\partial}{\partial t}\left(\frac{1}{2\pi} \int\limits_{B_2(x,t)} \frac{u_0(y)}{\left(t^2 - |y-x|^2\right)^{1/2}} \, dy \right).$$

Note that this formula involves integration over the entire ball $B_2(x,t)$ and not just the surface as in three dimensions. As a result, wave propagation in two dimensions is not as sharp as in three dimensions. If we strike a circular drumhead at the center, the vibrations propagate outwards in a circular pattern. The vibrations first hit a point a distance d from the center at time $t = d$ and that point continues to vibrate for all time afterwards. The amplitude of the vibrations decays roughly like $1/t$. We illustrate this in Fig. 245.2 where we show a finite element approximation to a related problem. See Strauss ([**?**]) for more details on wave propagation.

244.29. Write down a formula for the solution of the homogeneous wave equation in two dimensions corresponding to $u_0 = 1$ for $|x| \leq 1$ and 0 otherwise, and $\dot{u}_0 \equiv 0$.

245
Wave Equation FEM

245.1 Reformulation as System

To discretize (244.12), we first rewrite this scalar second order equation as a system of first order equations in time using the notation of Chapter 212.1 setting $u_1 = \dot{u}$ and $u_2 = u$: find the vector (u_1, u_2) such that

$$\begin{cases} \dot{u}_1 - \Delta u_2 = f & \text{in } \Omega \times (0,\infty), \\ -\Delta \dot{u}_2 + \Delta u_1 = 0 & \text{in } \Omega \times (0,\infty), \\ u_1 = u_2 = 0 & \text{on } \Gamma \times (0,\infty), \\ u_1(\cdot,0) = \dot{u}_0,\ u_2(\cdot,0) = u_0 & \text{in } \Omega. \end{cases} \tag{245.1}$$

We choose this formulation, and in particular write $\Delta u_1 = \Delta \dot{u}_2$ instead of $u_1 = \dot{u}_2$, because this brings (245.1) into a form that is analogous to the hyperbolic model problem of Chapter 212.1 with the positive coefficient a corresponding to $-\Delta$. Thus, we can use the same trick of cancellation that we used for the analysis of the hyperbolic model problem. In particular when $f \equiv 0$, if we multiply the first equation by u_1 and the second by u_2 and add, the terms $-(\Delta u_2, u_1)$ and $(\Delta u_1, u_2)$ cancel, leading to the conclusion that $\|u_1\|^2 + \|\nabla u_2\|^2$ is constant in time. In other words, we get energy conservation very easily.

The finite element functions we use to approximate the solution of (245.1) are piecewise linear polynomials in space and time that are continuous in space and "nearly" continuous in time. By nearly, we mean that the approximation is continuous unless the mesh changes from one time level

to the next. We call this the cG(1) method. We discretize $\Omega \times (0, \infty)$ in the usual way, letting $0 = t_0 < \cdots < t_n < \cdots$ denote a partition of $(0, \infty)$ and to each time interval $I_n = (t_{n-1},\, t_n]$ of length $k_n = t_n - t_{n-1}$, associate a triangulation $\mathcal{T}_n$ of Ω with mesh function h_n and a corresponding finite element space V_n of continuous piecewise linear vector functions in Ω that vanish on Γ. For $q = 0$ and 1, we define the space

$$W_{kn}^{(q)} = \left\{ (w_1, w_2) :\ w_j(x,t) = \sum_{r=0}^{q} t^r v_j^{(r)}(x),\ v_j^{(r)} \in V_n,\ j = 1, 2 \right\}$$

on the space-time slab $S_n = \Omega \times (t_{n-1}, t_n)$ and then the space $W_k^{(q)}$ of piecewise polynomial functions (v_1, v_2) such that $(v_1, v_2)|_{S_n} \in W_{kn}^{(q)}$ for $n = 1, 2, ..., N$. The functions in $W_k^{(q)}$ are forced to be continuous in space, but may be discontinuous in time.

The cG(1) method for (244.12) is based on the variational formulation of (245.1) as usual and reads: Find $U = (U_1, U_2) \in W_k^{(1)}$ such that for $n = 1, 2, \ldots,$

$$\begin{cases} \displaystyle\int_{t_{n-1}}^{t_n} \left((\dot{U}_1, w_1) + (\nabla U_2, \nabla w_1) \right) dt = \int_{t_{n-1}}^{t_n} (f, w_1)\, dt, \\ \displaystyle\int_{t_{n-1}}^{t_n} \left((\nabla \dot{U}_2, \nabla w_2) - (\nabla U_1, \nabla w_2) \right) dt = 0, \\ U_{1,n-1}^{+} = P_n U_{1,n-1}^{-}, \quad U_{2,n-1}^{+} = \pi_n U_{2,n-1}^{-}, \end{cases} \tag{245.2}$$

for all $w = (w_1, w_2) \in W_{kn}^{(0)}$, where $U_{1,0}^{-} = \dot{u}_0$, $U_{2,0}^{-} = u_0$, and

$$U_j(x,t)|_{S_n} = U_{j,n}^{-}(x) \frac{t - t_{n-1}}{k_n} + U_{j,n-1}^{+}(x) \frac{t - t_n}{-k_n}, \quad j = 1, 2.$$

Further, π_n is the *elliptic projection* into V_n defined by $(\nabla \pi_n w, \nabla v) = (\nabla w, \nabla v)$ for all $v \in V_n$. Note that $\pi_n w \in V_h$ is the Galerkin approximation of the solution w of Poisson's equation on Ω with homogeneous Dirichlet boundary conditions.

Note that if the mesh is unchanged across t_{n-1}, i.e. $\mathcal{T}_n = \mathcal{T}_{n-1}$, then both P_n and π_n reduce to the identity and the approximation U is continuous across t_{n-1}. If $V_{n-1} \subset V_n$, which occurs for example when the mesh is refined using the customary strategies, then the coefficients of $U_{j,n-1}^{+}$, $j = 1, 2$, can be found by straightforward interpolation of $U_{j,n-1}^{-}$, i.e., for j=1,2,

$$U_{j,n-1}^{+}(N_{n,i}) = U_{j,n-1}^{-}(N_{n,i}),$$

where $\{N_{n,i}\}$ is the set of nodes in $\mathcal{T}_n$.

245.1. Prove this last claim.

In this case, the components U_j are continuous across t_{n-1} in the sense that

$$\lim_{t\to t_{n-1}^-} U_j(x,t) = \lim_{t\to t_{n-1}^+} U_j(x,t) \quad \text{for all } x \in \Omega.$$

However, when the mesh is changed so $V_{n-1} \not\subset V_n$, which typically happens when the mesh is coarsened, then U_j will in general be discontinuous across t_{n-1}. We illustrate this in Fig. 245.1.

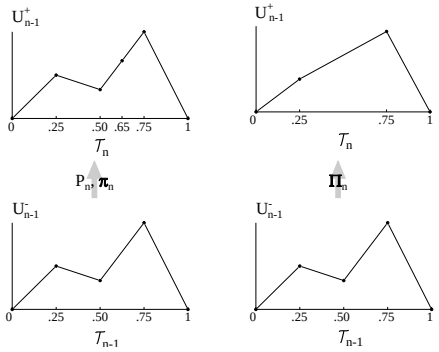


FIGURE 245.1. The effect of π_n in two cases of mesh changes. On the left, the mesh is refined so $V_{n-1} \subset V_n$ and π_n and P_n correspond to nodal interpolation. On the right, the mesh is coarsened and U_j is discontinuous across t_{n-1}.

245.2. Compute $U_{n-1}^+ = \pi_n U_{n-1}^-$ for the example on the right in Fig. 245.1 assuming that $U_{n-1}^-(.25) = 1/2$, $U_{n-1}^-(.5) = 1/3$, and $U_{n-1}^-(.75) = 1$.

We use B_n and A_n to denote the $M_n \times M_n$ mass and stiffness matrices associated to the nodal basis $\{\varphi_{i,n}\}$ for V_n with dimension M_n, and further $A_{n-1,n}$ to denote the $M_n \times M_{n-1}$ matrix with coefficients

$$\big(A_{n-1,n}\big)_{i,j} = (\nabla\varphi_{i,n}, \nabla\varphi_{j,n-1}), \quad 1 \le i \le M_n,\ 1 \le j \le M_{n-1},$$

and let $B_{n-1,n}$ be defined by (243.4). Finally $\xi_{j,n}^-$ and $\xi_{j,n-1}^+$ denote the vectors of coefficients with respect to $\{\varphi_{i,n}\}$ of $U_{j,n}^-$ and $U_{j,n-1}^+$ for $j = 1, 2$. With this notation, (245.2) is equivalent to the set of matrix equations

$$\begin{cases} B_n\big(\xi_{1,n}^- - \xi_{1,n-1}^+\big) + k_n A_n\big(\xi_{2,n}^- + \xi_{2,n-1}^+\big)/2 = F_n, \\ A_n\big(\xi_{2,n}^- - \xi_{2,n-1}^+\big) - k_n A_n\big(\xi_{1,n}^- + \xi_{1,n-1}^+\big)/2 = 0, \\ B_n\xi_{1,n-1}^+ = B_{n-1,n}\xi_{1,n-1}^-, \quad A_n\xi_{2,n-1}^+ = A_{n-1,n}\xi_{2,n-1}^-, \end{cases} \tag{245.3}$$

where F_n is the data vector with coefficients

$$\big(F_n\big)_i = \int_{t_{n-1}}^{t_n} (f, \varphi_{i,n})\, dt, \quad 1 \le i \le M_n.$$

245.3. Prove (245.3) is correct.

245.4. In the case the space mesh $\mathcal{T}_n$ does not change, show that (245.3) reduces to

$$\begin{cases} B(\xi_{1,n} - \xi_{1,n-1}) + k_n A(\xi_{2,n} + \xi_{2,n-1})/2 = F_n, \\ A(\xi_{2,n} - \xi_{2,n-1}) - k_n A(\xi_{1,n} + \xi_{1,n-1})/2 = 0, \end{cases}$$

where we have dropped the superscripts $+$ and $-$ on the coefficient vectors $\xi_{i,n}$ and the subscript n on A and B since U is continuous.

245.5. Formulate the cG(1) finite element method that uses the lumped mass quadrature rule in space and the midpoint rule in time to evaluate the integrals giving the approximation. Write out the discrete matrix equations for the approximation.

245.2 Energy Conservation

One reason that we use the cG(1) method (245.2) is that the approximation conserves the total energy when $f \equiv 0$ provided $V_{n-1} \subset V_n$ for all n. To prove this, we choose $w_j = (U^+_{j,n-1} + U^-_{j,n})/2$ in (245.2) and add the two equations to get

$$\int_{t_{n-1}}^{t_n} (\dot{U}_1, U_1)\,dt + \int_{t_{n-1}}^{t_n} (\nabla \dot{U}_2, \nabla U_2)\,dt = 0,$$

because of the terms that cancel. This gives

$$\|U^-_{1,n}\|^2 + \|\nabla U^-_{2,n}\|^2 = \|U^-_{1,n-1}\|^2 + \|\nabla U^-_{2,n-1}\|^2. \tag{245.4}$$

In other words, the total energy of the cG(1) approximation is conserved from one time step to the next, just as holds for the solution of the continuous problem. When the mesh is changed so $V_{n-1} \not\subset V_n$, then the energy is only approximately conserved because each projection onto the new mesh changes the total energy.

245.6. Provide the details of the proof of (245.4).

245.7. Compute the change in energy in U in Problem 245.2.

245.3 A Posteriori Error Estimates and Adaptivity

In this section, we present an a posteriori error analysis under some simplifying assumptions. The analysis of the cG(1) method for (245.1) is analogous to the analysis of the cG(1) method for the hyperbolic model problem in Chapter 212.1, but there are new technical difficulties in the case of the partial differential equation.

The adaptive error control is based on an a posteriori error estimate as usual. We prove the estimate under the assumptions that Ω is convex and the space mesh is kept constant, which simplifies the notation considerably. We use $\mathcal{T}_h$ to denote the fixed triangulation of mesh size $h(x)$ and we denote the corresponding finite element space by V_h. We use P_h to denote the L_2 projection into V_h and Δ_h to denote the discrete Laplacian on V_h. We use P_k to denote the L_2 projection into the set of piecewise constant functions on the partition $\{t_n\}$, and R_2, to denote the space residual associated to the discretization of the Laplacian as defined in Chapter **??**. Finally, since U is continuous, we set $U_{j,n-1} = U^+_{j,n-1} = U^-_{j,n-1}$. We shall prove the following a posteriori error estimate assuming that Ω is convex so that Theorem 240.6 applies.

Theorem 245.1 *There is a constant C_i such that for $N = 1, 2, ...,$*

$$\begin{aligned}\|u_2(t_N) - U_{2,N}\| &\leq C_i\Big(\|h^2 R_2(U_{2,N})\| + \|h^2 R_2(U_{2,0})\| \\ &\quad + \int_0^{t_N} \big(\|h(f - P_h f)\| + \|h^2 R_2(U_1)\|\big)\, dt \\ &\quad + \int_0^{t_N} \big(\|k(f - P_k f)\| + \|k\Delta_h(U_2 - P_k U_2)\| \\ &\quad + \|k\nabla(U_1 - P_k U_1)\|\big)\, dt\Big).\end{aligned}$$

Note that the first four quantities on the right arise from the space discretization and the last three quantities arise from the time discretization. The integrals in time implies that errors accumulate at most linearly with time, as expected from the analysis of the model hyperbolic problem.

Proof: The proof is based on using the continuous dual problem to get an error representation formula. The dual problem is: For $N \geq 1$, find $\varphi = (\varphi_1, \varphi_2)$ such that

$$\begin{cases} -\dot{\varphi}_1 + \Delta\varphi_2 = 0 & \text{in } \Omega \times (0, t_N), \\ \Delta\dot{\varphi}_2 - \Delta\varphi_1 = 0 & \text{in } \Omega \times (0, t_N), \\ \varphi_1 = \varphi_2 = 0 & \text{on } \Gamma \times (0, t_N), \\ -\Delta\varphi_2(\cdot, t_N) = e_{2,N} & \text{in } \Omega, \\ \varphi_1(\cdot, t_N) = 0 & \text{in } \Omega, \end{cases} \tag{245.5}$$

where $e_2 = u_2 - U_2$. We multiply the first equation in (245.5) by $e_1 = u_1 - U_1$, the second by e_2, add the two together, integrate over $\Omega \times (0, t_N)$, integrate by parts, and finally use the Galerkin orthogonality of the ap-

proximation, to obtain

$$\|e_{2,N}\|^2 = \sum_{n=1}^{N}\Big(\big(f-\dot U_1, \varphi_1 - P_kP_h\varphi_1\big)_n \\ - \big(\nabla U_2, \nabla(\varphi_1 - P_kP_h\varphi_1)\big)_n - \big(\nabla(\dot U_2 - U_1), \nabla(\varphi_2 - P_kP_h\varphi_2)\big)_n\Big),$$

where $(\cdot,\cdot)_n$ denotes the $L_2(S_n)$ inner product. The goal is to distinguish the effects of the space and time discretizations by using the splitting $v - P_kP_hv = (v - P_hv) + (P_hv - P_kP_hv)$ and the orthogonalities of the L_2 projections P_h and P_k to obtain

$$\begin{aligned}\|e_{2,N}\|^2 &= \sum_{n=1}^{N}\Big(\big(f-P_kf, P_h\varphi_1 - P_kP_h\varphi_1\big)_n + \big(f-P_hf, \varphi_1 - P_h\varphi_1\big)_n \\ &\quad - \big(\nabla U_2, \nabla(\varphi_1 - P_h\varphi_1)\big)_n \\ &\quad - \big(\nabla(U_2 - P_kU_2), \nabla(P_h\varphi_1 - P_kP_h\varphi_1)\big)_n \\ &\quad - \big(\nabla(\dot U_2 - U_1), \nabla(\varphi_2 - P_h\varphi_2)\big)_n \\ &\quad - \big(\nabla(\dot U_2 - U_1), \nabla(P_h\varphi_2 - P_kP_h\varphi_2)\big)_n\Big).\end{aligned}$$

Finally, using the fact that $\varphi_1 = \dot\varphi_2$ and integrating by parts in t, we obtain

$$\begin{aligned}\|e_{2,N}\|^2 &= \\ &\sum_{n=1}^{N}\Big(\big(f-P_kf, P_h\varphi_1 - P_kP_h\varphi_1\big)_n + \big(f-P_hf, \varphi_1 - P_h\varphi_1\big)_n \\ &\quad - \big(\nabla(U_2 - P_kU_2), \nabla(P_h\varphi_1 - P_kP_h\varphi_1)\big)_n \\ &\quad + \big(\nabla U_1, \nabla(\varphi_2 - P_h\varphi_2)\big)_n \\ &\quad + \big(\nabla(U_1 - P_kU_1), \nabla(P_h\varphi_2 - P_kP_h\varphi_2)\big)_n\Big) \\ &\quad - \big(\nabla U_{2,N}, \nabla(\varphi_{2,N} - P_h\varphi_{2,N})\big) + \big(\nabla U_{2,0}, \nabla(\varphi_{2,0} - P_h\varphi_{2,0})\big).\end{aligned}$$

To complete the proof, we use (243.3) and a standard estimate for $v - P_kv$ together with the following stability result for the dual problem (245.5).■

Lemma 245.2 *If Ω is convex, then the solution φ of (245.5) satisfies*

$$\|\ddot\varphi_2\|_{[0,t_N]} + \|D\dot\varphi_2\|_{[0,t_N]} + \|D^2\varphi_2\|_{[0,t_N]} \le C\|e_{2,N}\|. \tag{245.6}$$

Proof: Multiplying the first equation in (245.5) by $\Delta\varphi_1$ and the second by $\Delta\varphi_2$ and adding, after using Greens formula, we obtain

$$\frac{d}{dt}(\|\nabla\varphi_1\|^2 + \|\Delta\varphi_2\|^2) = 0.$$

It follows using the initial conditions that

$$\|\nabla\varphi_1\|^2 + \|\Delta\varphi_2\|^2 = \|e_{2,N}\|^2.$$

The desired conclusion results from using the elliptic regularity estimate (240.6) and the fact that $\ddot{\varphi}_2 = \Delta\varphi_2$. ■

245.8. (a) Fill in the details of the above proof. (b) (*Hard!*) Extend the a posteriori error estimate to the case $\mathcal{T}_n$ varies with n.

245.4 Adaptive Error Control

Following the ideas in the previous chapters, we can formulate an algorithm for adaptive error control by using the a posteriori error bound in Theorem 245.1 to give an estimate of the error on a given space-time mesh. We illustrate the use of the a posteriori error bound in two examples.[1]

In the first example, we compute the effects of a sharp strike at the center of a large square drumhead. We can model the problem for small amplitude vibrations by posing the wave equation on a finite domain $\Omega = [0,1]\times[0,1]$ with homogeneous Neumann boundary conditions. We assume the drumhead is initially at rest, i.e. $u(\cdot,0) = \dot{u}(\cdot,0) = 0$ and we model the strike by a source f located at the center of the square defined by

$$f(x,t) = \begin{cases} \sin^2(\pi t/T), & \text{for } t \le .1,\ |x - (.5,.5)| \le .1, \\ 0, & \text{otherwise.} \end{cases}$$

We plot the finite element approximation at time $t \approx .7$ in Fig. 245.2. We compute with a fixed time step $k_n \equiv .01$ for a relatively short time so that the error due to time discretization is remains small. The space mesh is adapted according to the error control algorithm based on using an a posteriori error bound to equidistribute the error across the elements. The a posteriori analysis presented above can be changed to cover Neumann boundary conditions in a straightforward way.

The second example is a computation on a model of wave propagation in an inhomogeneous, linear elastic, viscously damped solid. We assume that the displacements are small and perpendicular to the (x_1, x_2) plane and that the solid is relatively long in the x_3 direction, which reduces the model to a scalar, two-dimensional wave equation for the shear waves propagating in the (x_1, x_2) plane.

[1] These computations are provided courtesy of M. G. Larson and A. J. Niklasson. See *Adaptive finite element methods for elasto-dynamics*, preprint, Department of Mathematics, Chalmers University of Technology, S41296 Göteborg, Sweden, for further details.

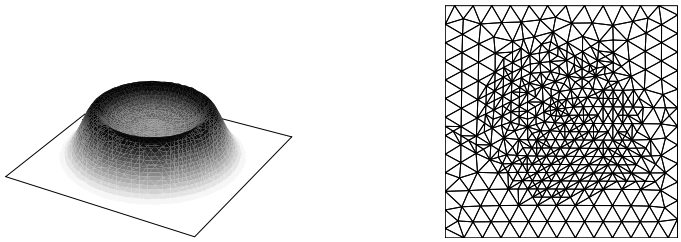


FIGURE 245.2. Plot of the finite element approximation and the corresponding space mesh for the model of the drumhead struck sharply at the center. In the plot of the displacement, a 15 level grey scale is used with black representing the largest and white the smallest displacement. The mesh on the right was adapted from an initial uniform mesh based on an a posteriori error bound.

In the specific example we present, the domain $\Omega = (0,1) \times (0,1)$ is composed of two isotropic materials joined along the line $x_2 = .59$. The material in the upper portion has a shear modulus that is five times larger than the material in the lower portion, so that the wave speed is five times greater in the upper portion of Ω. This gives the equation $\ddot{u} - a\Delta u = f$, where

$$a(x) = \begin{cases} 1, & x_2 \leq .59, \\ 5, & x_2 > .59. \end{cases}$$

We assume homogeneous Neumann (stress-free) boundary conditions and an approximate point source that is active for small time, so we define

$$f(x,t) = \begin{cases} \sin^2(\pi t/.07), & |x - (.4,.4)| \leq .1 \text{ and } t \leq .14, \\ 0, & \text{otherwise.} \end{cases}$$

This is the kind of forcing that might be found in nondestructive ultrasonic testing and seismology. The a posteriori error bound used to control the adaptivity is derived using techniques similar to those used to prove Theorem 245.1. We show contour plots of the approximation and the associated space meshes at times $t = .05, .15, .25$, and $.4$ in Fig. 245.3–Fig. 245.6. The material interface is marked with a horizontal line, and the difference in wave speed in the two materials is clear.

245.5 A Priori Error Estimate

We state an a priori error estimate for the cG(1) method in the case $\mathcal{T}_n$ is constant and assuming that Ω is convex.

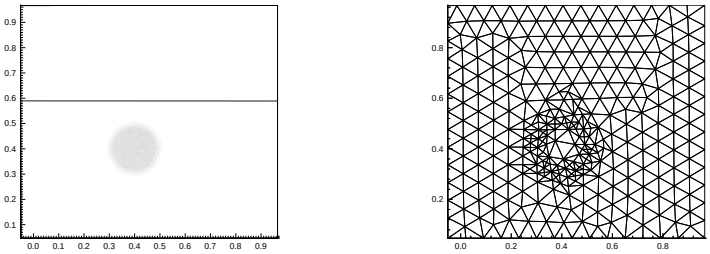


FIGURE 245.3. Density plot of the finite element approximation and the corresponding space mesh for the wave equation on an inhomogeneous material at time $t = .05$. The forcing has not reached maximum strength yet.

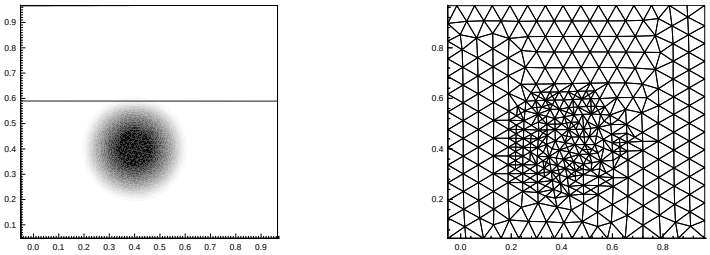


FIGURE 245.4. $t = .15$

Theorem 245.3 *There is a constant C_i such that if U satisfies (245.2), then for $N = 1, 2, ...,$*

$$\|u_2(\cdot, t_N) - U_{2,N}\| \leq C_i \int_0^{t_N} \left(\|k^2 \nabla \ddot{u}_2\| + \|k^2 \ddot{u}_1\| + \|h^2 D^2 \dot{u}_2\| \right) dt.$$

We note that the a priori error estimate is of order $\mathcal{O}(h^2 + k^2)$ and like the a posteriori estimate, the integral in time corresponds to a linear rate of accumulation of errors.

Proof: To simplify the analysis, we only analyze the time discretization. This corresponds to setting V_n equal to the space of functions with square-integrable gradients on Ω and that vanish on Γ. We use $\hat{u}$ to denote the piecewise linear time interpolant of u. Since we already know how to estimate $\rho = u - \hat{u}$, we only have to estimate $e = \hat{u} - U$.

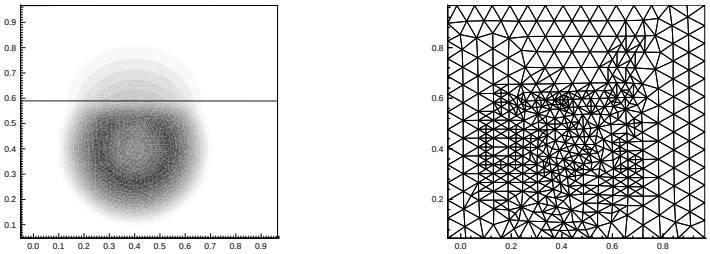


FIGURE 245.5. $t = .25$

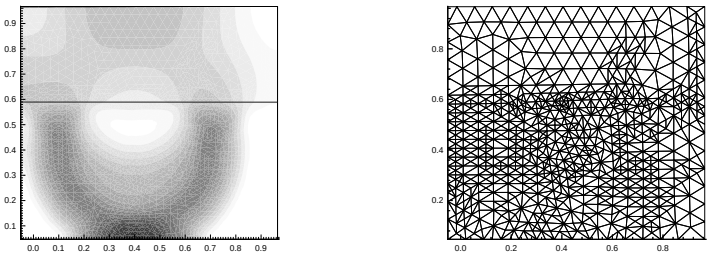


FIGURE 245.6. $t = .40$

To this end, we use the discrete dual problem: Find $(\Phi_1, \Phi_2) \in W_k^{(1)}$ that satisfies for $n = N, n-1, ..., 1$,

$$\begin{cases} -(v_1, \dot{\Phi}_1)_n - (\nabla v_1, \nabla \Phi_2)_n = 0, \\ -(\nabla v_2, \nabla \dot{\Phi}_2)_n + (\nabla v_2, \nabla \Phi_1)_n = 0, \end{cases} \tag{245.7}$$

for all $(v_1, v_2) \in W_{kn}^{(0)}$, where $\Phi_{1,N} \equiv 0$ and $-\Delta \Phi_{2,N} = e_{2,N}$ with $e_2 = \hat{u}_2 - U_2$. Because the test functions v_1 and v_2 are piecewise constant, we may replace Φ_1 and Φ_2 in the second terms in each equation by their mean values $\bar{\Phi}_{i,n} := (\Phi_{i,n-1} + \Phi_{i,n})/2$ on I_n. After this, we can replace the test functions v_1 and v_2 by arbitrary piecewise linear functions, because both $\dot{\Phi}_i$ and $\bar{\Phi}_i$ are piecewise constant. In particular, replacing v_1 by $e_1 = \hat{u}_1 - U_1$ and v_2 by e_2, adding the resulting equations and summing over n, and then

integrating by parts in time, we obtain the error representation

$$\begin{aligned}
&\|e_{2,N}\|^2 \\
&\quad = \sum_{n=1}^{N}\Big(\big(\nabla \dot{e}_2, \nabla \Phi_2\big)_n + \big(\nabla e_2, \nabla \bar{\dot{\Phi}}_1\big)_n + \big(\dot{e}_1, \Phi_1\big)_n - \big(\nabla e_1, \nabla \bar{\dot{\Phi}}_2\big)_n \Big) \\
&\qquad = \sum_{n=1}^{N}\Big(\big(\dot{e}_1, \bar{\dot{\Phi}}_1\big)_n + \big(\nabla e_2, \nabla \bar{\dot{\Phi}}_1\big)_n + \big(\nabla \dot{e}_2, \nabla \bar{\dot{\Phi}}_2\big)_n - \big(\nabla e_1, \nabla \bar{\dot{\Phi}}_2\big)_n \Big) \\
&\quad = -\sum_{n=1}^{N}\Big(\big(\dot{\rho}_1, \bar{\dot{\Phi}}_1\big)_n + \big(\nabla \rho_2, \nabla \bar{\dot{\Phi}}_1\big)_n \\
&\qquad\qquad + \big(\nabla \dot{\rho}_2, \nabla \bar{\dot{\Phi}}_2\big)_n - \big(\nabla \rho_1, \nabla \bar{\dot{\Phi}}_0\big)_n \Big).
\end{aligned}$$

We also replaced Φ_j by their mean values $\bar{\dot{\Phi}}_j$ and then used Galerkin orthogonality to replace U by u. Since the terms involving $\dot{\rho}_i$, $i = 1, 2$ vanish, we arrive at the following error representation

$$\|e_{2,N}\|^2 = -\sum_{n=1}^{N}\Big(\big(\nabla \rho_2, \nabla \bar{\dot{\Phi}}_1\big)_n - \big(\rho_1, \Delta \bar{\dot{\Phi}}_2\big)_n \Big).$$

Choosing $v_1 = -\Delta \bar{\dot{\Phi}}_1$ and $v_2 = -\Delta \bar{\dot{\Phi}}_2$ in (245.7), we obtain the stability estimate:

$$\|\nabla \Phi_{1,n}\|^2 + \|\Delta \Phi_{2,n}\|^2 = \|e_{2,N}\|^2 \quad \text{for all } n \le N.$$

Combining this with the error representation and then using standard estimates for the interpolation error ρ, we obtain the a priori error estimate.■

245.9. Supply the details of the proof.

245.10. Show that assuming the solution $u(x,t)$ of the wave equation $\ddot{u} - \Delta u = f$ has the form $u(x,t) = \exp(i\omega t)w(x)$, where $f(x,t) = \exp(i\omega t)g(x)$ and $\omega > 0$ is a given frequency, leads to the stationary Helmholtz's equation $-\Delta w - \omega^2 w = g$ for the amplitude $w(x)$. Show that a fundamental solution of Helmholtz's equation in $\mathbb{R}^3$ is given by $\frac{\exp(i\omega|x|)}{4\pi|x|}$. Solve Helmholtz's equation using Femlab on a bounded two-dimensional domain with suitable boundary conditions in a configuration of physical interest.

245.11. Derive the wave equation from Maxwell's equations under suitable assumptions.

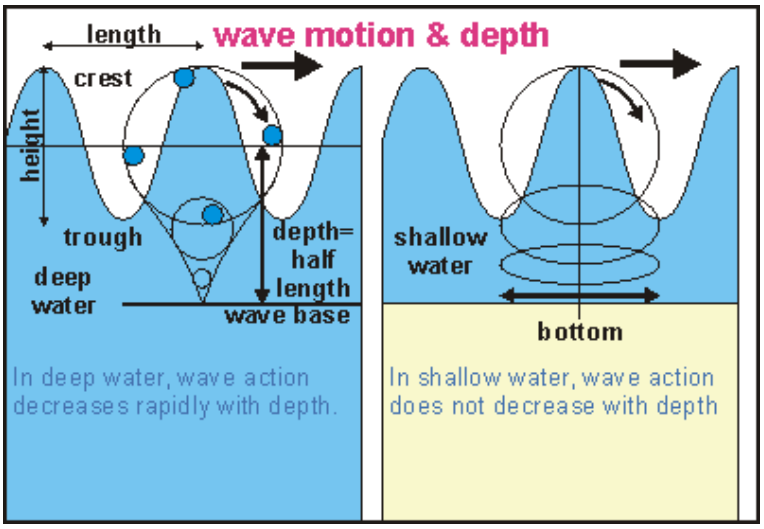


FIGURE 245.7. Particle motion in deep and shallow water waves.

246

Stationary Convection-Diffusion Analysis

> I have always found it difficult to read books that cannot be understood without too much meditation. For, when following one's own meditation, one follows a certain natural inclination and gains profit along with pleasure; but one is enormously cramped when having to follow the meditaton of others. (Leibniz)

246.1 Introduction

In this chapter and the next, we consider a linear model for a problem that includes the effects of convection, diffusion, and absorption, which is an example of a *multi-physics* problem coupling several physical phenomena. We begin by deriving the model and discussing the basic properties of solutions. In this chapter, we continue by considering the discretization of the stationary case, starting with a discussion that explains why a straightforward application of Galerkin's method yields disappointing results for a convection dominated problem. We then present a modified Galerkin method that resolves the difficulties, that we call the *streamline diffusion finite element method* or *Sd method*. We continue with the time dependent case in the next chapter. The material of these two chapters lay the foundation for the application of the finite element method to incompressible and compressible fluid flow including reactive flow, multi-phase flow and free-boundary flow, developed in the advanced companion volume.

246.2 A basic model

We consider the transport of heat in a current flowing between two regions of a relatively large body of water, for example from a warm region to a cold region, taking into account the dissipation of the heat, the advection of the heat by the current, and the absorption of heat into the air. An example of such a physical situation is the North American Drift flowing from Newfoundland, where it continues the Gulf Stream, to the British Isles, where it splits into two branches. The North American Drift is responsible for the warm climate of Western Europe. Our interest is focused on the water temperature in the Drift at different locations at different times. The full problem takes place in three space dimensions, but we simplify the model to two dimensions assuming all functions are independent of the depth.

The model is a time-dependent scalar convection-diffusion-absorption problem posed on a *space-time* domain $Q = \Omega \times I$, where Ω is a polygonal domain in $\mathbb{R}^2$ with boundary Γ and $I = (0, T)$, of the form

$$\begin{cases} \dot{u} + \nabla \cdot (\beta u) + \alpha u - \nabla \cdot (\epsilon \nabla u) = f & \text{in } Q, \\ u = g_- & \text{on } (\Gamma \times I)_-, \\ u = g_+ \text{ or } \epsilon \partial_n u = g_+ & \text{on } (\Gamma \times I)_+, \\ u(\cdot, 0) = u_0 & \text{in } \Omega, \end{cases} \tag{246.1}$$

where u represents the temperature, $\beta = (\beta_1, \beta_2)$, α and $\epsilon > 0$ are functions of (x, t) representing the *convection velocity*, *absorption coefficient*, and *diffusion coefficient*, respectively. Further, $f(x, t)$, u_0, g, and u_0 are given data, and

$$(\Gamma \times I)_- = \{(x, t) \in \Gamma \times I : \beta(x, t) \cdot n(x) < 0\},$$
$$(\Gamma \times I)_+ = \{(x, t) \in \Gamma \times I : \beta(x, t) \cdot n(x) \geq 0\},$$

where $n(x)$ is the outward normal to Γ at point x, are the *inflow* and *outflow* parts of the space-time boundary $\Gamma \times I$, respectively. We illustrate this in Fig. 246.1.

246.1. Let $\Omega = (0, 1) \times (0, 1)$, $I = (0, 1)$, and $\beta = \big(\cos\big(\frac{\pi}{2}t + \frac{\pi}{4}\big), \sin\big(\frac{\pi}{2}t + \frac{\pi}{4}\big)\big)$. Identify the inflow and outflow boundaries of Q.

The model is the result of expressing conservation of heat as

$$\frac{\partial}{\partial t}(\lambda u) + \nabla \cdot q + \alpha u = f,$$

where q is the heat flow and λ the heat capacity, and assuming that the constitutive law is the following generalization of Fourier's law (232.3)

$$q = \beta \lambda u - \epsilon \nabla u.$$

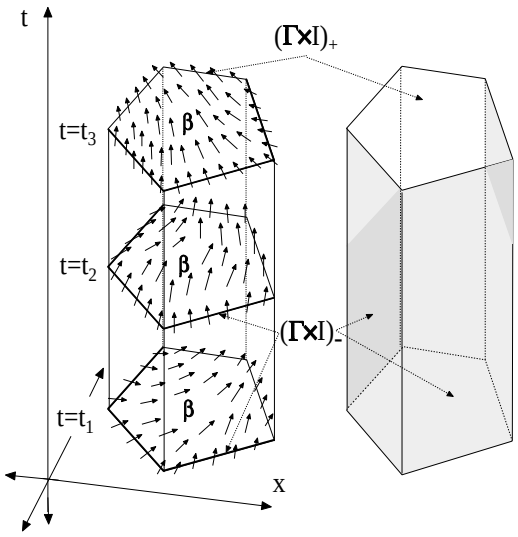


FIGURE 246.1. The space time domain Q indicating the inflow and outflow boundaries. The inflow boundary is shaded in the figure on the right.

Setting the heat capacity $\lambda = 1$, gives (246.1). This model is a natural extension of the model for heat flow considered in Chapter **??** with the addition of terms corresponding to convection of heat with the current β and absorption of heat at the rate α.

Using the identity

$$\nabla \cdot (\beta u) = \beta \cdot \nabla u + (\nabla \cdot \beta)u,$$

we may replace the convection term $\nabla \cdot (\beta u)$ by $\beta \cdot \nabla u$ by modifying the term αu to $(\alpha + \nabla \cdot \beta)u$.

The model (246.1) models a variety of phenomena with the variable u representing a quantity subject to convection, diffusion and absorption. Another example is the evolution of a contaminant dropped into fluid running in a pipe, see Fig. 246.2, where u represents the concentration of the contaminant in the fluid. A system of the form (246.1) may also serve as a simple model for fluid flow described by the Navier-Stokes equations, in which case u represents mass, momentum and energy. Thus, (246.1) is a fundamental model.

246.2. The motion of the rotor of an electrical motor gives rise to an additional contribution to the electric field E of the form $\beta \times B$ where β is the velocity and B the magnetic flux. Show that introducing this term into the derivation of (242.2) leads to the convection-diffusion equation

$$\sigma \frac{\partial u}{\partial t} + \sigma\beta \cdot \nabla u - \nabla \cdot \left(\frac{1}{\mu}\nabla u\right) = f.$$

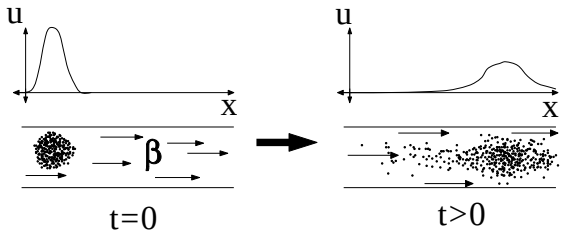


FIGURE 246.2. The convection and diffusion of a dye inside a water pipe. $u(x,t)$ represents the concentration of the dye at (x,t).

246.3 The stationary convection-diffusion problem

We begin by considering the stationary convection-diffusion-absorption problem associated to (246.1),

$$\begin{cases} \beta \cdot \nabla u + \alpha u - \nabla \cdot (\epsilon \nabla u) = f & \text{in } \Omega, \\ u = g_- & \text{on } \Gamma_-, \\ u = g_+ \text{ or } \epsilon \partial_n u = g_+ & \text{on } \Gamma_+, \end{cases} \tag{246.2}$$

with all functions independent of time, and α modified to include $\nabla \cdot \beta$ as indicated above. In this case, the definitions of the inflow and outflow boundaries Γ_- and Γ_+ are given by

$$\Gamma_- = \{x \in \Gamma : \beta(x) \cdot n(x) < 0\} \quad \text{and} \quad \Gamma_+ = \{x \in \Gamma : \beta(x) \cdot n(x) \geq 0\},$$

see Fig. 246.3. We first discuss basic features of solutions of the problem

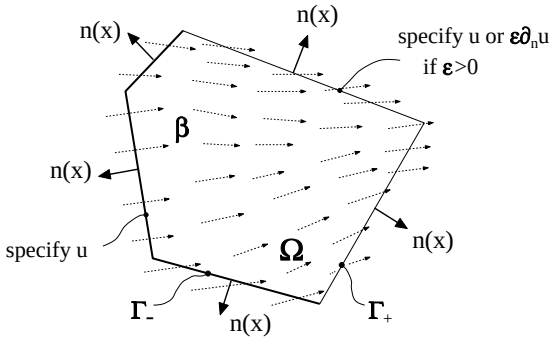


FIGURE 246.3. The notation for a stationary convection-diffusion problem.

(246.2) and then consider the computation of approximate solutions using finite element methods. Special care has to be taken in the design of the finite element method, because direct application of Galerkin's method to

(246.2) when the convection is the dominant feature leads to numerical solutions with spurious oscillations, which is illustrated in Problem 246.6 below.

246.3.1 Convection versus diffusion

Generally, the relative size of ϵ and β govern the qualitative nature of (246.2). If $\epsilon/|\beta|$ is small, then (246.2) is *convection dominated* and has hyperbolic character. If $\epsilon/|\beta|$ is not small, then (246.2) is *diffusion dominated* and has elliptic character. Thus, the problem (246.2) changes character from hyperbolic to elliptic as $\epsilon/|\beta|$ increases. In the diffusion dominated case the material on elliptic problems in Chapter **??** is applicable since the convection terms are dominated by the diffusion terms.

We now focus on the convection-dominated hyperbolic case and then first consider the extreme case with $\epsilon = 0$.

246.3.2 The reduced problem

The *reduced problem* with $\epsilon = 0$ takes the form

$$\begin{cases} \beta \cdot \nabla u + \alpha u = f & \text{in } \Omega, \\ u = g_- & \text{on } \Gamma_-, \end{cases} \tag{246.3}$$

where u is specified only on the inflow boundary Γ_-. The reduced problem couples convection and absorption.

The *streamlines* associated to the stationary convection velocity field $\beta(x)$ are curves $x(s)$, parametrized by $s \geq 0$, satisfying

$$\begin{cases} \dfrac{dx}{ds} = \beta(x(s)) & \text{for } s > 0, \\ x(0) = \bar{x}, \end{cases} \tag{246.4}$$

for the streamline starting at $\bar{x}$, see Fig. 246.4. This is the path followed by a particle starting at $\bar{x}$ that is convected with velocity $\beta(x)$. In this interpretation, s is time and dx/ds represents the particle velocity. A streamline

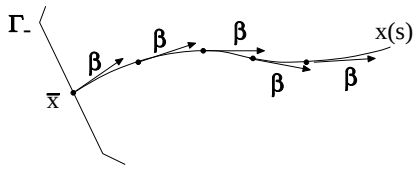


FIGURE 246.4. A streamline has tangent vector $\beta(x(s))$ at every point $x(s)$.

is *closed* if the particle returns to the point of departure, i.e. $x(s) = \bar{x}$ for

some $s > 0$. A problem with closed streamlines requires special care, so we assume for now that there aren't any. The reduced equation becomes an ordinary differential equation along a streamline since by the chain rule,

$$\frac{d}{ds}u(x(s)) + \alpha(x(s))u(x(s)) = (\beta \cdot \nabla u + \alpha u)(x(s)) = f(x(s)), \quad s > 0,$$

where the inflow data $g_-(\bar{x})$ at $\bar{x} \in \Gamma_-$ gives the "initial data" $u(x(0))$. The solution of the reduced problem (246.3) therefore can be found by solving for each streamline $x(s)$ an ordinary differential equation of the form (**??**):

$$\dot{v}(s) + a(s)v(s) = \bar{f}(s), \quad s > 0, \ v(0) = g_-(x(0)),$$

where $v(s) = u(x(s))$, $a(s) = \alpha(x(s))$ and $\bar{f}(s) = f(x(s))$, corresponding to "solving along streamline starting at inflow". We note that the case of non-negative absorption with $\alpha(x) \geq 0$ corresponds to the parabolic case with $a(s) \geq 0$.

We conclude that in the reduced problem without diffusion information is propagated sharply along streamlines from the inflow boundary to the outflow boundary. We see in particular that if there is a discontinuity in the inflow data at some point $\bar{x}$ on the inflow boundary Γ_-, then the solution of (246.3) will in general be discontinuous across the entire streamline staring at $\bar{x}$. As an example, the solution of the problem

$$\begin{cases} \dfrac{\partial u}{\partial x_1} = 0 \quad \text{in } x \in \Omega, \\ u(0, x_2) = \begin{cases} 0, & 0 < x_2 < 1/2, \\ 1, & 1/2 \leq x_2 < 1, \end{cases} \end{cases}$$

corresponding to (246.2) with $\beta = (1, 0)$, $\alpha = 0$ and $\Omega = [0, 1] \times [0, 1]$, is given by

$$u(x_1, x_2) = \begin{cases} 0, & 0 < x_2 < 1/2, \ 0 < x_1 < 1, \\ 1, & 1/2 \leq x_2 < 1, \ 0 < x_1 < 1, \end{cases}$$

with a discontinuity across the streamline $x(s) = (s, 1/2)$.

246.3. Suppose $\beta = (1, 1 - x_1)$ and $\Omega = [0, 1] \times [0, 1]$. (a) Plot Ω and identify the inflow and outflow boundaries. (b) Compute the streamlines corresponding to each point on the inflow boundary (Hint: there are two cases). Plot enough of the streamlines so that you can describe the "flow" over Ω.

246.4. Solve the problem $x_1 \frac{\partial u}{\partial x_1} + x_2 \frac{\partial u}{\partial x_2} = 0$ on $\Omega = \{x : 1 < x_1, x_2 < 2\}$, with some choice of inflow data.

246.3.3 Layers of difficulty

The features of the reduced problem with $\epsilon = 0$ are present also in the hyperbolic case with $\epsilon/|\beta|$ small positive but now the presence of positive

diffusion makes the solution continuous and "spreads out" a discontinuity over a *layer* in the solution, which is a narrow region where the solution changes significantly. For example, a discontinuity across a streamline becomes a *characteristic layer* of width $O(\sqrt{\epsilon})$, see Fig. 246.5. Further, if Dirichlet boundary conditions are specified on the outflow boundary Γ_+ in the case $\epsilon > 0$, then in general the solution u of (246.2) has an outflow *boundary layer* of width $O(\epsilon)$ close to Γ_+ where u changes rapidly to meet the boundary condition; see Fig. 246.5.

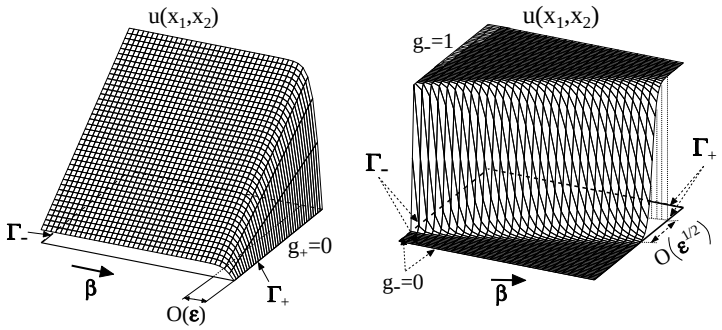


FIGURE 246.5. Illustrations of an outflow (on the left) and a characteristic layer caused by a discontinuity in g_- (on the right).

To give a concrete example of an outflow layer we consider the one-dimensional analog of (246.2), which takes the form

$$\begin{cases} -(\epsilon u')' + \beta u' + \alpha u = f \quad \text{for } 0 < x < 1, \\ u(0) = 0, \ u(1) = 0, \end{cases} \tag{246.5}$$

in the case of homogeneous Dirichlet boundary conditions. We present computational results in Fig. 246.6 for the case $\epsilon = 0.02$, $\beta = 1$, $\alpha = 0$ and $f = 1$ using L_2 norm error control on the tolerance level .02. The flow is from left to right with inflow at $x = 0$ and outflow at $x = 1$. Note the outflow layer in u in the boundary layer near $x = 1$ resulting from the convection in the positive x direction and how the mesh is refined in that region.

246.5. Show that the width of an outflow layer is approximately of order ϵ by explicitly solving the one-dimensional convection-diffusion problem $-\epsilon u'' + u' = 0$ for $0 < x < 1$ with $u(0) = 1$, $u(1) = 0$.

We now present a problem showing that Galerkin's method may go berserk under certain conditions. We urge the reader to do this problem before continuing.

246.6. Consider the continuous Galerkin cG(1) method for the one-dimensional problem $-\epsilon u'' + u' = 0$ in $(0, 1)$ with $u(0) = 1$, $u(1) = 0$. (a) Write down the

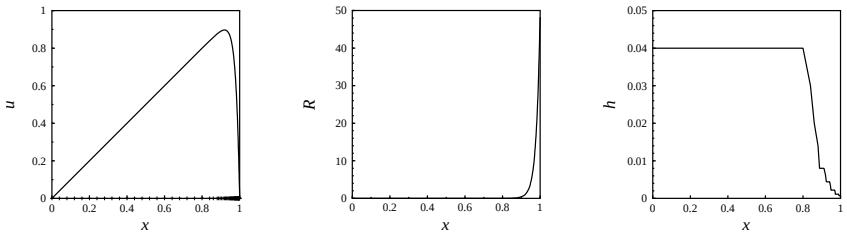


FIGURE 246.6. Solution, error, and meshsize for (246.5) with $\epsilon = .02$, $\beta = 1$, $\alpha = 0$, $f = 1$, and TOL=.02.

discrete equations for the cG(1) approximation computed on a uniform mesh with M interior nodes. (b) With $\epsilon = 0.01$, compute the approximation for $M = 10$ and $M = 11$ and compare to the true solution. (c) Compute the approximation with $M \approx 100$ and compare with the exact solution. (d) Write out the discrete equations when $\epsilon = h/2$. Explain why this scheme is called the *upwind method* for the reduced problem. How is the convection term approximated by Galerkin's method? Compare with the upwind method. Compare the nature of propagation of effects (in particular the outflow boundary condition) in Galerkin's method with $\epsilon > 0$ much smaller that h and the upwind method.

247
Stationary Convection-Diffusion FEM

247.1 The Streamline Diffusion Method

Convection dominated problems present difficulties for computation that are not present in diffusion dominated problems, mainly because the stability properties of convection dominated problems cause the standard Galerkin finite element method to be non-optimal compared to interpolation. Recall that Galerkin methods are typically optimal for elliptic and parabolic problems, and in general for diffusion dominated problems. However, the standard Galerkin method for convection dominated problems may be far from optimal if the exact solution is nonsmooth, in which case the Galerkin approximations contain "spurious" oscillations not present in the true solution. This is illustrated in Problem 246.6. The oscillations occur whenever the finite element mesh is too coarse to resolve layers, which typically is the case in the early stages of an adaptive refinement process. The oscillations result from a lack of stability of the standard Galerkin finite element method for convection dominated problems, and may have disastrous influence on the performance of an adaptive method leading to refinements in large regions where no refinement is needed.

We conclude that it is important to improve the stability properties of the Galerkin finite element method. However, this has to be done cleverly, because additional stability is often obtained at the price of decreased accuracy. For example, increasing artificially the diffusion term (e.g. by simply setting $\epsilon = h$) will increase the stability of the Galerkin method, but may

also decrease accuracy and prevent sharp resolution of layers. Thus, the objective is to improve stability without sacrificing accuracy.

We consider two ways of enhancing the stability of the standard Galerkin finite element method:

(a) introduction of weighted least squares terms

(b) introduction of artificial viscosity based on the residual.

We refer to the Galerkin finite element method with these modifications as the streamline diffusion, or Sd-method, and motivate this terminology below. The modification (a) adds stability through least squares control of the residual and the modification (b) adds stability by the introduction of an elliptic term with the size of the diffusion coefficient, or *viscosity*, depending on the residual with the effect that viscosity is added where the residual is large, i.e., typically where the solution is nonsmooth. Both modifications enhance stability without a strong effect on the accuracy because both modifications use the residual.

247.1.1 Abstract formulation

We begin by describing the Sd-method for an abstract linear problem of the form

$$Au = f, \tag{247.1}$$

for which the standard Galerkin finite element method reads: compute $U \in V_h$ such that

$$(AU, v) = (f, v) \quad \text{for all } v \in V_h, \tag{247.2}$$

where A is a linear operator on a vector space V with inner product $(\cdot,\cdot)$ and corresponding norm $\|\cdot\|$, and $V_h \subset V$ is a finite element space. Typically, A is a convection-diffusion differential operator, $(\cdot,\cdot)$ is the L_2 inner product over some domain Ω. We assume that A is positive semi-definite, i.e. $(Av, v) \geq 0$ for all $v \in V$.

The *least squares method* for (247.1) is to find $U \in V_h$ that minimizes the residual over V_h, that is

$$\|AU - f\|^2 = \min_{v \in V_h} \|Av - f\|^2.$$

This is a convex minimization problem and the solution U is characterized by

$$(AU, Av) = (f, Av) \quad \text{for all } v \in V_h. \tag{247.3}$$

We now formulate a Galerkin/least squares finite element method for (247.1) by taking a weighted combination of (247.2) and (247.3): compute $U \in V_h$ such that

$$(AU, v) + (\delta AU, Av) = (f, v) + (\delta f, Av) \quad \text{for all } v \in V_h, \tag{247.4}$$

where $\delta > 0$ is a parameter to be chosen. Rewriting the relation (247.4) as

$$(AU, v + \delta Av) = (f, v + \delta Av) \quad \text{for all } v \in V_h, \tag{247.5}$$

we can alternatively formulate the Galerkin/least squares method as a *Petrov-Galerkin method*, which is a Galerkin method with the space of test functions being different from the space of trial functions V_h. In our case, the test functions have the form $v + \delta Av$ with $v \in V_h$.

Adding the artificial viscosity modification (b) yields (with a typical choice of diffusion operator) the Sd-method in abstract form: find $U \in V_h$ such that

$$(AU, v + \delta Av) + (\hat{\epsilon}\nabla U, \nabla v) = (f, v + \delta Av) \quad \text{for all } v \in V_h, \tag{247.6}$$

where $\hat{\epsilon}$ is the artificial viscosity defined in terms of the residual $R(U) = AU - f$ through

$$\hat{\epsilon} = \gamma_1 h^2 |R(U)|, \tag{247.7}$$

with γ_1 a positive constant to be chosen, and $h(x)$ the local mesh size of V_h.

Choosing $v = U$ in (247.6) we see that the modifications improve the stability of the approximation as compared to (247.2).

247.1. Assume $(Av, v) \geq c\|v\|^2$ for some positive constant c. (a) Choose $v = U$ in (247.6) and derive a stability result for U. (b) Compare the result from (a) to the stability result obtained by choosing $v = U$ in (247.2). How does the stability result from (a) depend on δ and γ_1?

247.1.2 The streamline diffusion method for a convection-diffusion problem

We now formulate the streamline diffusion method for (246.2) with constant ϵ and homogeneous Dirichlet boundary conditions using the standard space of piecewise linear functions $V_h \subset V = H_0^1(\Omega)$ based on a triangulation $\mathcal{T}_h$ of Ω: compute $U \in V_h$ such that

$$(AU, v + \delta Av) + (\hat{\epsilon}\nabla U, \nabla v) = (f, v + \delta Av) \quad \text{for all } v \in V_h, \tag{247.8}$$

where $(\cdot, \cdot)$ is the $L_2(\Omega)$ inner product,

$$Aw = \beta \cdot \nabla w + \alpha w, \ \delta = \frac{1}{2}\frac{h}{|\beta|},$$
$$\hat{\epsilon}(U, h) = \max\{\epsilon, \gamma_1 h^2 |f - (\beta \cdot \nabla U + \alpha U)|, \gamma_2 h^{3/2}\},$$

where the γ_j are positive constants to be specified. We obtain (247.8) by multiplying the terms in (246.2) that appear in the reduced equation by the modified test function $v + \delta(\beta \cdot \nabla v + \alpha v)$, which corresponds to a least

squares modification of the convection/absorption terms, while multiplying the diffusion term in (246.2) by v after replacing ϵ by $\hat{\epsilon}$. If ϵ is variable or higher order polynomials are used, then the diffusion term should be included in the least squares modification.

In general, $\hat{\epsilon}$ depends on U and the discrete problem (247.8) is nonlinear, even though the continuous problems (246.2) and (246.3) are linear. When iterative methods are used to solve the discrete equations, the additional complication in solving the discrete equations due to the nonlinearity introduced by $\hat{\epsilon}$ has little effect on the computational overhead. The artificial viscosity $\hat{\epsilon}$ is proportional to $|f - (\beta \cdot \nabla U + \alpha U)|$, which plays the role of the residual. For simplicity, the jump terms related to the diffusion term has been left out; see the statement of Theorem 247.5.

The size of the artificial viscosity $\hat{\epsilon}$ relative to the mesh size h (assuming $\epsilon \leq h$) gives a measure of the smoothness of the exact solution u. In regions where u is smooth, $\hat{\epsilon} \approx h^{3/2}$, while in outflow layers in general $\hat{\epsilon} = \gamma_1 h^2 |f - (\beta \cdot \nabla U + \alpha U)| \propto h$, because there $|f - (\beta \cdot \nabla U + \alpha U)| \propto h^{-1}$ on a general mesh. In characteristic layers, typically $|f - (\beta \cdot \nabla U + \alpha U)| \approx h^{-1/2}$ so that again $\hat{\epsilon} \propto h^{3/2}$. Thus, we distinguish two basic cases:

(a) u is "smooth" with $\hat{\epsilon} \propto h^{3/2}$, including characteristic layers,

(b) u is non-smooth with $\hat{\epsilon} \propto h$, typically resulting in outflow layers.

We assume for the sake of simplicity that $\hat{\epsilon} = \epsilon$, which can be guaranteed during a computation by adding this requirement to the stopping criterion in the adaptive algorithm. The case $\hat{\epsilon} > \epsilon$ typically occurs in initial stages of adaptive refinements when the mesh is coarse. We focus on the case with $h^2 \leq \epsilon \leq h$. If ϵ is larger than h then all layers are resolved by the mesh, and if ϵ is smaller than h^2 then the mesh is much too coarse.

In Fig. 247.1, we present the results of a computation using the adaptive streamline diffusion method on the convection-diffusion problem with $\Omega = (0,1) \times (0,1)$, $\beta = (2,1)$, $\alpha = 0$, $\epsilon = 0.01$, and discontinuous inflow data $u(0,y) \equiv 1$, $0 \leq y \leq 1$ and $u(x,0) \equiv 0$, $0 < x < 1$. Note the form and thickness of the layers and the corresponding shape of the adapted mesh.

247.2. Plot Ω for this computation and identify the streamlines and the inflow and outflow boundaries.

247.2 A framework for an error analysis

We describe the basic ingredients of the analysis of the streamline diffusion method. The goal is an a posteriori error estimate that can be used to guide the mesh adaptivity in order to control the error. After presenting the general points, we analyze a specific case in the following section.

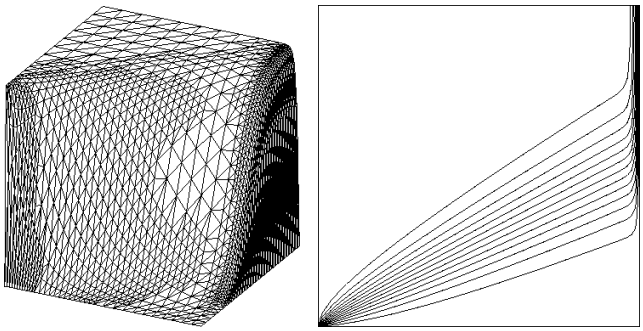


FIGURE 247.1. A surface plot with the mesh indicated and associated level curves of the approximate solution obtained using the streamline diffusion method for a convection-diffusion problem with both outflow and characteristic layers.

247.2.1 Basic stability estimates

We assume that

$$\alpha - \frac{1}{2}\nabla \cdot \beta \geq c > 0, \tag{247.9}$$

for some constant c. In the case of non-closed streamlines this condition may be satisfied by a change of variable; cf. Problem 247.4. The weak stability estimate for the solution u of (246.2) has the form:

$$\|\sqrt{\epsilon}\nabla u\| + \|u\| \leq C\|f\|, \tag{247.10}$$

with $C = (\sqrt{c}+2)/(2c)$, and $\|\cdot\|$ denotes the $L_2(\Omega)$ norm. In what follows, the exact value of C changes, but it is always a constant that depends on the constant c in (247.9). The estimate (247.10) follows after multiplying the differential equation in (246.2) by u, integrating over Ω, and using the fact that $(\beta \cdot \nabla u, u) = -\frac{1}{2}(\nabla \cdot \beta\, u, u)$.

A corresponding stability estimate for the discrete problem (247.8) is obtained by choosing $v = U$, which gives

$$\|\sqrt{\hat{\epsilon}}\nabla U\| + \|\sqrt{\delta}(\beta \cdot \nabla U + \alpha U)\| + \|U\| \leq C\|f\|. \tag{247.11}$$

We note that the control of the $\|\sqrt{\delta}(\beta \cdot \nabla U + \alpha U)\|$ term results from the least squares modification of the streamline diffusion method, and that the artificial viscosity $\hat{\epsilon}$ is present in the gradient term. The $\|U\|$ term allows the $\|\sqrt{\delta}(\beta \cdot \nabla U + \alpha U)\|$ term to be replaced by $\|\sqrt{\delta}\beta \cdot \nabla U\|$, yielding a weighted control of the streamline derivative $\beta \cdot \nabla U$. This control corresponds to adding diffusion in the streamline direction with coefficient δ, and this is the motivation for the name "streamline diffusion method".

Below, we also use an analog of the stability estimate (247.11) with $\hat{\epsilon} = \epsilon = 0$ that has the following form: for piecewise continuous functions w with $w = 0$ on Γ_-,

$$\|\sqrt{\delta}Aw\|^2 + c\|w\|^2 \leq (Aw, w + \delta Aw), \tag{247.12}$$

where as above $A = \beta \cdot \nabla w + \alpha w$. This estimate follows from the following version of Green's formula after noting that the boundary term is guaranteed to be non-negative if $w = 0$ on Γ_-, because $\beta \cdot n \geq 0$ on Γ_+.

Lemma 247.1

$$(\beta \cdot \nabla w, w) = -\frac{1}{2}(\nabla \cdot \beta\, w, w) + \frac{1}{2}\int_\Gamma w^2 \beta \cdot n\, ds. \tag{247.13}$$

We note that the stability estimate (247.12) requires w to be specified (to be zero) on the inflow boundary Γ_-. The estimate gives a motivation why it is natural to specify data on Γ_-, rather than on Γ_+, in the case $\epsilon = 0$.

247.3. Provide the details in the derivations of (247.10), (247.11), (247.12) and (247.13).

247.4. Show that the equation $u'(s) = f(s)$, where $s \in \mathbb{R}$, takes the form $v'(s) + v(s) = \exp(-s)f(s)$ using the change of dependent variable $v(s) = \exp(-s)u(s)$.

247.2.2 A strong stability estimate

In addition to the weak stability estimate (247.10), we also use the following estimate for a dual continuous problem which can be written in the form (246.2) with Neumann outflow boundary conditions:

$$\|\beta \cdot \nabla u + \alpha u\| + \|\epsilon D^2 u\| \leq C\|f\|, \tag{247.14}$$

where C is a moderately sized constant that does not depend in a significant way on ϵ if Ω is convex. We refer to this estimate as a *strong stability estimate* because second derivatives are bounded, in addition to the control of the term $\beta \cdot \nabla u + \alpha u$. The "price" of the second derivative control is a factor ϵ^{-1}, which is natural from the form of the equation. Since ϵ is small, the "price" is high, but nevertheless there is a net gain from using this estimate, because the presence of the second derivatives brings two powers of h to compensate the ϵ^{-1} factor.

We are not able to prove the analog of the strong stability estimate (247.14) for the discrete problem, which would be useful in deriving a priori error estimates. Instead, we use (247.11) as a substitute, yielding a weighted control of $\beta \cdot \nabla U + \alpha U$ with the weight $\sqrt{\delta}$.

We summarize the effects of the two modifications used to create the streamline diffusion method: the least squares modification gives added control of the derivative in the streamline direction with a weight $\sqrt{\delta}$, while the artificial viscosity modification gives control of the gradient ∇U with the weight $\sqrt{\hat{\epsilon}}$.

247.2.3 Basic forms of the error estimates

Assuming that $\hat{\epsilon} = \epsilon$, the a posteriori error estimate for the streamline diffusion method (247.8) has the form:

$$\|u - U\| \leq C_i S_c \left\| \frac{h^2}{\epsilon} R(U) \right\|, \tag{247.15}$$

where $S_c \approx 1$ and $R(U)$ is the residual of the finite element solution U defined in terms of the differential equation in a natural way. We note the presence of the factor h^2/ϵ that results from combining strong stability with Galerkin orthogonality. In many cases, we have $h^2/\epsilon << 1$. For example if $\epsilon \approx h^{3/2}$, which corresponds to a "smooth" exact solution such as a solution with a characteristic layer, then (247.15) reduces to

$$\|u - U\| \leq C\|h^{1/2}\, R(U)\|.$$

If $\epsilon \approx h$, which corresponds to a "non-smooth" exact solution such as a solution with an outflow layer, then (247.15) reduces to

$$\|u - U\| \leq C\|hR(U)\|.$$

To understand the gain in (247.15), compare it to the "trivial" a posteriori error estimate

$$\|u - U\| \leq C\|R(U)\| \tag{247.16}$$

that follows directly from the weak stability estimate and even holds for non-Galerkin methods. This estimate is almost useless for error control in the case the exact solution is non-smooth, because the right-hand side in general increases with decreasing mesh size until all layers are resolved.

The a priori error estimate takes the form

$$\|h^{1/2}R(U)\| + \|u - U\| \leq C_i S_{c,h} \|h^{3/2} D^2 u\|,$$

where $S_{c,h} \approx 1$. In the case of a smooth solution, the a posteriori and a priori error estimates match and both are non-optimal with a loss of $h^{1/2}$, while in the non-smooth case with $\hat{\epsilon} = h$, the a posteriori estimate appears in optimal form.

247.5. Prove (247.16). Assuming that $R(U) \approx h^{-1}$ in an outflow layer of width of order h, estimate $\|R(U)\|$ and discuss what would happen if an adaptive method tried to control $\|u - U\|$ by using (247.16). Do the same in the case of a characteristic layer assuming $R(U) \approx h^{-1/2}$ in a layer of width $h^{1/2}$.

247.3 A posteriori error analysis in one dimension

We consider the one-dimensional convection-diffusion-absorption problem (246.5) with $\beta = 1$, $\alpha = 1$ and ϵ a small positive constant:

$$\begin{cases} -\epsilon u'' + u' + u = f & \text{in } (0,1), \\ u(0) = u(1) = 0. \end{cases} \tag{247.17}$$

This problem in general has an outflow layer of width $O(\epsilon)$ at $x = 1$ where the solution rapidly changes to adjust to the imposed outflow boundary value $u(1) = 0$.

For simplicity, we consider the streamline diffusion method for (247.17) without the artificial viscosity modification, which takes the form : Compute $U \in V_h$ such that

$$(U' + U, v + \delta(v' + v)) + (\epsilon U', v') = (f, v + \delta(v' + v)) \qquad \text{for all } v \in V_h, \tag{247.18}$$

where $\delta = h/2$ when $\epsilon < h$ and $\delta = 0$ otherwise, V_h is the usual space of continuous piecewise linear functions that vanish at $x = 0, 1$, and $(\cdot, \cdot)$ the $L_2(\Omega)$ inner product. We note that the streamline diffusion method is essentially obtained by multiplication by the modified test function $v + \delta(v' + v)$. The modification has a stabilizing effect, which is manifested by the presence of the positive term $(U' + U, \delta(U' + U))$, obtained by choosing $v = U$ in (247.18). If δ is increased, the stability is improved but at the cost of accuracy. If δ is decreased, then the reverse is true. Choosing $\delta \approx h/2$ gives the best compromise and results in a satisfactory combination of stability and accuracy.

We now prove an L_2 a posteriori error estimate for the streamline diffusion method (247.18). For simplicity, we consider a case with $h \leq \epsilon$ and $\delta = 0$.

Theorem 247.2 *There is a constant C independent of ϵ and h such that the solution U of (247.18) satisfies the following estimate for all $\epsilon \geq 0$*

$$\|u - U\| \leq C_i \Big\| \frac{h^2}{\epsilon^*}(f - U_x) \Big\| + |\epsilon u'(0)| + |\epsilon U'(0)|,$$

where $\epsilon^*(x) = h^{1/2}\epsilon$ on the interval of the subdivison underlying V_h with left-hand endpoint $x = 0$, $\epsilon^* = \epsilon$ elsewhere, and $\|\cdot\|$ denotes the $L_2(\Omega)$ norm

Proof: Let φ be the solution of the dual problem

$$\begin{cases} -\epsilon\varphi'' - \varphi' + \varphi = g & \text{for } 0 < x < 1, \\ \varphi'(0) = 0, \varphi(1) = 0, \end{cases} \tag{247.19}$$

with the direction of the convection from right to left, which is opposite to that of (247.17). We pose the dual problem with Dirichlet inflow boundary condition at the inflow at $x = 1$ and it is convenient to choose a Neumann outflow condition at the outflow at $x = 0$. Choosing $g = u - U$ in (247.19), multiplying by $u-U$ and integrating by parts, we get and using the Galerkin orthogonality,

$$\|u - U\|^2 = \int_0^1 (f - U' - U)(\varphi - \pi_h\varphi)\,dx \\ - \int_0^1 \epsilon U'(\varphi - \pi_h\varphi)'\,dx + \epsilon u'(0)\varphi(0),$$

where $\pi_h\varphi \in V_h$ interpolates φ at the interior mesh points. Using the following stability result this proves the desired result, up to the small modification of ϵ required because in general $\varphi(0) \neq 0$, while $\pi_h\varphi(0) = 0$. ■

Lemma 247.3 *There is a constant C such that if φ solves (247.19), then*

$$|\varphi(0)| \leq \|g\| \text{ and } \|\epsilon\varphi''\| \leq \|g\|. \tag{247.20}$$

Proof: Multiplication with φ and integration gives

$$\int_0^1 (\epsilon\varphi')^2\,dx + \int_0^1 \varphi^2\,dx + \frac{1}{2}\varphi(0)^2 \leq \frac{1}{2}\int_0^1 g^2\,dx + \frac{1}{2}\int_0^1 \varphi^2\,dx,$$

which proves the estimate for $|\varphi(0)|$. Next, multiplication with $-\epsilon\varphi''$ gives

$$\int_0^1 (\epsilon\varphi'')^2\,dx + \int_0^1 \varphi'\epsilon\varphi''\,dx + \int_0^1 \epsilon(\varphi')^2\,dx = -\int_0^1 g\epsilon\varphi''.$$

Since

$$2\int_0^1 \varphi'\varphi''\,dx = \varphi'(1)^2 \geq 0,$$

this proves the desired estimate for $\epsilon\varphi''$ by Cauchy's inequality. ■

247.6. Determine the *Green's function* $g_z(x)$ for the boundary value problem

$$\begin{cases} -\epsilon u'' + bu' = f, & 0 < x < 1, \\ u(0) = u(1) = 0, \end{cases} \tag{247.21}$$

where b is constant. This is the function $g_z(x)$ defined for $0 < z < 1$ that satisfies

$$\begin{cases} -\epsilon g_z'' - bg_z' = \delta_z, & 0 < x < 1, \\ g_z(0) = g_z(1) = 0, \end{cases}$$

where δ_z denotes the delta function at z. Prove the representation formula

$$u(z) = \int_0^1 g_z(x)f(x)\,dx, \quad 0 < z < 1, \tag{247.22}$$

where $u(x)$ is the solution of (247.21). Consider first the case $\epsilon = 1$ and $b = 0$, and then the case $\epsilon > 0$ and $b = 1$, paying particular attention to the limit $\epsilon \to 0$.

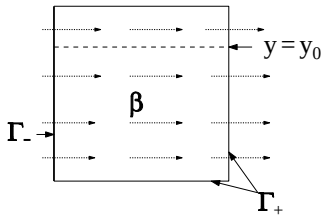


FIGURE 247.2. The model problem.

247.4 Error analysis in two dimensions

We prove error estimates for a model problem of the form (246.2) with $\beta = (1,0)$, $\alpha = 1$, ϵ constant and $\Omega = (0,1) \times (0,1)$. For convenience, we denote the coordinates in $\mathbb{R}^2$ by (x,y), and we write

$$u_x = \partial u/\partial x = \beta \cdot \nabla u = (1,0) \cdot \nabla u,$$

and formulate the model problem (see Fig. 247.2) as follows:

$$\begin{cases} u_x + u - \epsilon \Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma. \end{cases} \tag{247.23}$$

This problem is a direct extension of the one-dimensional model problem (246.5) to two dimensions. Solutions of (247.23) can have an outflow layer of width $O(\epsilon)$ along Γ_+ and also characteristic layers of width $O(\sqrt{\epsilon})$ along characteristics $\{(x,y) : y = y_0\}$ that do not occur in the corresponding one-dimensional problem.

247.4.1 Strong stability analysis

We use the following strong stability estimate for the associated dual problem with homogeneous Neumann outflow boundary data.

Lemma 247.4 *The solution φ of the dual problem*

$$\begin{cases} -\varphi_x + \varphi - \epsilon \Delta \varphi = g, & \text{in } \Omega, \\ \varphi = 0, & \text{on } \Gamma_+, \\ \varphi_x = 0, & \text{on } \Gamma_-, \end{cases} \tag{247.24}$$

satisfies the stability estimates

$$\left(\|\varphi\|^2 + 2\|\epsilon^{1/2} \nabla \varphi\|^2 + \int_{\Gamma_-} \varphi^2 \, ds \right)^{1/2} \leq \|g\|, \tag{247.25}$$

$$\|\epsilon D^2 \varphi\| \leq \|g\|. \tag{247.26}$$

Proof: Multiplying (247.24) by 2φ and integrating, we obtain (247.25) after using the fact that $-2(\varphi_x, \varphi) = \int_{\Gamma_-} \varphi^2 ds$. Next, multiplying (247.24) by $-\epsilon\Delta\varphi$ and integrating, we obtain

$$\|\epsilon\Delta\varphi\|^2 + (\epsilon\nabla\varphi, \nabla\varphi) + (\epsilon\varphi_{xx}, \varphi_x) + (\epsilon\varphi_{yy}, \varphi_x) = (f, -\epsilon\Delta\varphi).$$

Since $\varphi_x = 0$ on Γ_-, we have

$$2(\varphi_{xx}, \varphi_x) = \int_{\Gamma_+} \varphi_x^2 \, ds.$$

On the two sides of Ω with $y = 0$ and 1, $\varphi_x = 0$, while $\varphi_y = 0$ on Γ_+. This gives

$$2(\varphi_{yy}, \varphi_x) = -2(\varphi_y, \varphi_{xy}) = \int_{\Gamma_-} \varphi_y^2 \, ds.$$

We conclude that

$$\|\epsilon\Delta\varphi\|^2 \le (f, -\epsilon\Delta\varphi) \le \|f\|\|\epsilon\Delta u\|.$$

The desired estimate follows using the elliptic regularity result $\|D^2\varphi\| \le \|\Delta\varphi\|$, see (240.37). ■

247.4.2 The a posteriori error estimate

We prove an a posteriori error estimate in the case $\delta = 0$ and $\hat{\epsilon} = \epsilon$ constant. The proof when $\delta \approx h$ is obtained by a simple modification.

Theorem 247.5 *There is a constant C such that*

$$\|u - U\| \le C\left(\left\|\frac{h^2}{\epsilon^*}R(U)\right\| + \|\epsilon\partial_n u\|_{\Gamma_-} + \|\epsilon\partial_n U\|_{\Gamma_-}\right), \tag{247.27}$$

where $R(U) = R_1(U) + R_2(U)$ with

$$R_1(U) = |f - U_x - U|$$

and

$$R_2(U) = \frac{\epsilon}{2} \max_{S\subset\partial K} h_K^{-1}|[\partial_S U]| \quad \text{on } K \in \mathcal{T}_h, \tag{247.28}$$

where $[\partial_S v]$ denotes the jump across the side $S \subset \partial K$ in the normal derivative of the function v in V_h, and $\epsilon^* = \epsilon h^{1/2}$ on K if $K \cap \Gamma_- \neq 0$ and $\epsilon^* = \epsilon$ otherwise.

Proof: Letting φ denote the solution of the dual problem (247.24) with $g = e = u - U$, we obtain the following error representation by using

Galerkin orthogonality and the equations defining u and U,

$$\begin{aligned}\|e\|^2 &= (e, -\varphi_x + \varphi - \epsilon\Delta\varphi)\\ &= (e_x + e, \varphi) + (\epsilon\nabla e, \nabla\varphi)\\ &= (u_x + u, \varphi) + (\epsilon\nabla u, \nabla\varphi) - (U_x + U, \varphi) - (\epsilon\nabla U, \nabla\varphi)\\ &= (f, \varphi) + \int_{\Gamma_-} \epsilon\,\partial_n u\,\varphi\,ds - (U_x + U, \varphi) - (\epsilon\nabla U, \nabla\varphi)\\ &= (f - U_x - U, \varphi - \pi_h\varphi) - (\epsilon\nabla U, \nabla(\varphi - \pi_h\varphi)) + \int_{\Gamma_-} \epsilon\,\partial_n u\,\varphi\,ds,\end{aligned}$$

from which the desired estimate follows by standard interpolation error estimates and Lemma 247.4.

247.7. Supply the details to finish the proof.

247.8. Prove a similar result when $\delta \approx h$.

We note that the $*$ modification of ϵ is required to deal with the incompatibility of boundary conditions for φ and functions in V_h on Γ_-. ■

247.4.3 The a priori error estimate

We prove the a priori error estimate (247.6) in the simplified case that $\epsilon = \hat{\epsilon} = 0$. The Dirichlet boundary condition is specified only on the inflow boundary. Using Galerkin orthogonality in the analog of (247.12) for the error $e = u - U$ with $A = \frac{\partial}{\partial x} + I$, we get

$$\begin{aligned}\|\sqrt{\delta}Ae\|^2 + c\|e\|^2 \le (Ae, e + \delta Ae) &= (Ae, u - \pi_h u + \delta A(u - \pi_h u))\\ &\le \frac{1}{2}\|\sqrt{\delta}Ae\|^2 + \|\delta^{-1/2}(u - \pi_h u)\|^2 + \|\sqrt{\delta}A(u - \pi_h u)\|^2,\end{aligned}$$

where as usual $\pi_h u$ denotes the nodal interpolant of u. Choosing $\delta = h$ and using standard interpolation results, yields

$$\frac{1}{2}\|\sqrt{h}Ae\|^2 + c\|e\|^2 \le C_i^2\|h^{3/2}D^2u\|^2.$$

We state the final result, which extends directly to the case with ϵ small, as a theorem.

Theorem 247.6 *If $\alpha - \frac{1}{2}\nabla\cdot\beta \ge c > 0$ and $\epsilon \le h$, then*

$$\|u - U\| \le CC_i\|h^{3/2}\,D^2u\|.$$

247.4.4 The propagation of information

It is possible to prove a "local" form of an a priori error estimate for the streamline diffusion method in which the L_2 error over a domain $\Omega_1 \subset \Omega$ that excludes layers is estimated in terms of the L_2 norm of $h^{3/2}D^2u$ over a slightly larger domain Ω_2 that also excludes layers. The upshot is that the presence of e.g. an outflow layer where the error may be locally large if ϵ is small, does not degrade the accuracy away from the layer. This is because in the streamline diffusion method, effects are propagated more or less along streamlines from inflow to outflow in the direction of the "wind" β just as effects are propagated in the continuous problem. In particular, the streamline diffusion method does not have the spurious propagation in the opposite direction to the wind that occurs in the standard Galerkin method.

247.9. (a) Consider the problem $-\epsilon u'' + u' + u = f$ for $0 < x < 1$, together with $u(0) = 1$, $u(1) = 0$. Let $\psi(x)$ be a positive weight function on I such that $0 \leq -\psi' \leq C\psi/\epsilon$, with C a suitable constant. Prove a stability estimate of the form $\|\sqrt{\psi}u\| \leq C\|\sqrt{\psi}f\|$. Use this estimate to draw a conclusion on the decay of information in the "upwind" direction. Hint: multiply by ψu. (b) (*Hard*) Extend to the streamline diffusion method. Hint: multiply by $\pi_h(\psi U)$ and estimate the effect of the perturbation $\psi U - \pi_h(\psi U)$.

247.5 79 A.D.

The figure below, adapted with the permission of the National Geographic Society, shows the ash fall resulting from the eruption of Mount Vesuvius in 79 A.D. This is an example of a full scale convection-diffusion problem with the convection velocity corresponding to a wind from north-west and an approximate delta function source. The level curves of the ash downfall (levels at 0.1, 1 and 2 meters are faintly shaded) give a measure of the concentration of ash in the atmosphere in various directions from the crater. Note the pronounced propagation of ash in the direction of the wind due to convection. The propagation against the wind due to diffusion is much smaller.

248

Time Dependent Convection-Diffusion Analysis

> The fact that in nature "all is woven into one whole", that space, matter, gravitation, the forces arising from the electromagnetic field, the animate and inanimate are all indissolubly connected, strongly supports the belief in the unity of nature and hence in the unity of scientific method. (Weyl)

248.1 Introduction

We return to the time-dependent problem (246.1), considering mainly the *convection dominated* case with $\epsilon/|\beta|$ small since the case when $\epsilon/|\beta|$ is not small can be analyzed by extending the results for the heat equation presented in Chapter **??**.

The cG(1)dG(r) method for the heat equation, using cG(1) in space and dG(r) in time on space-time slabs, can be applied to (246.1) with modifications like those used to create the streamline diffusion method for stationary convection diffusion problems (246.2). We discuss this approach briefly in Section 249.3. However, it turns out that using space-time meshes that discretize space and time independently is not optimal for convection dominated problems. It is better to use a mesh that is *oriented* along characteristics or space-time particle paths. We illustrate this in Fig. 248.1. We refer to a finite element method using oriented meshes as a *characteristic Galerkin* method, or a chG method.

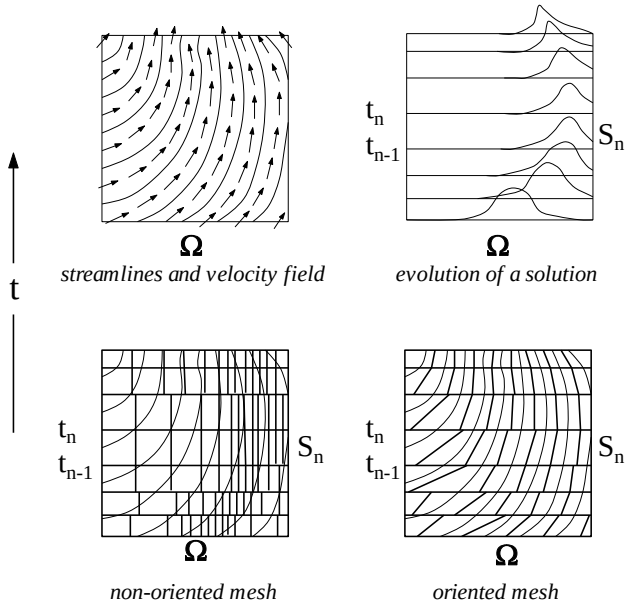


FIGURE 248.1. The upper figures show the space-time domain with the flow field in space-time, the space-time streamlines, and an illustration of the evolution of a solution. The two lower figures show non-oriented and oriented space-time meshes.

In particular, we study the chG(0) method obtained applying the cG(1)dG(0) method on a mesh oriented along particle paths in space and time inside each slab. In its most elementary form, the chG(0) method reduces on each space-time slab to an L_2 projection from the previous mesh onto the current mesh followed by an exact transport in the directions of the characteristics. The main computational work is spent on the L_2 projections. However for the purpose of analysis, it is more useful to view the chG method as a Galerkin method on a space-time mesh oriented in space-time along space-time particle paths. In addition, this opens the way to generalizations in which the space-time meshes are oriented in different ways.

We begin by describing the two fundamental ways to represent solutions of a convection problem, called respectively *Euler coordinates* and *Lagrange coordinates*.

248.2 Euler and Lagrange coordinates

We describe the coordinates systems in the context of measuring the temperature in the North Atlantic stream. Dr. Euler and Dr. Lagrange each lead a team of assistants provided with boats and thermometers. Dr. Eu-

ler's assistants anchor their boats at specific locations and measure the temperature of the water continuously as it flows past their positions. Dr. Lagrange's assistants, on the other hand, drift with the current while measuring the temperature. An assistant to Dr. Euler measures the temperature of the water as it is affected by the current in contrast to an assistant to Dr. Lagrange who measures the temperature of the same "piece" of water, albeit at different positions. In order to correlate the measurements of the two groups, it is necessary to record the stationary positions of Dr. Euler's assistants and to keep track of the changing positions of Dr. Lagrange's assistants.

To simplify the mathematical description of the two sets of coordinates, we consider the model problem,

$$\begin{cases} \dot{u} + \beta \cdot \nabla u - \epsilon \Delta u = f & \text{in } Q = \mathbb{R}^2 \times (0, \infty), \\ u(x,t) \to 0 & \text{for } t > 0 \text{ as } |x| \to \infty, \\ u(\cdot, 0) = u_0 & \text{in } \mathbb{R}^2, \end{cases} \tag{248.1}$$

where we assume that β is smooth, f and u_0 have *bounded support*, which means that they are zero outside some bounded set, and $\epsilon \geq 0$ is constant. This means in particular that we avoid here discussing complications rising from boundaries in space.

248.2.1 Space-time particle paths

The *space-time particle paths*, or *characteristics*, corresponding to the convection part $\dot{u} + \beta \cdot \nabla u$ of (248.1) are curves $(x,t) = (x(\bar{x},\bar{t}), t(\bar{t}))$ in space and time parameterized by $\bar{t}$, where $x(\bar{x},\bar{t})$ and $t(\bar{t})$ satisfy

$$\begin{cases} \dfrac{dx}{d\bar{t}} = \beta(x,\bar{t}) & \text{for } \bar{t} > 0, \\ \dfrac{dt}{d\bar{t}} = 1 & \text{for } \bar{t} > 0, \\ x(\bar{x},0) = \bar{x}, \quad t(0) = 0. \end{cases} \tag{248.2}$$

This is analogous to the stationary case with the operator $\beta \cdot \nabla$ replaced by $1 \cdot \partial/\partial t + \beta \cdot \nabla$, where the coefficient of the time-derivative is one and t acts as an extra coordinate. Here, the time coordinate has a special role and in fact $t = \bar{t}$. The projection of the space-time particle path into space is given by the curve $x(\bar{x},\bar{t})$ satisfying

$$\begin{cases} \dfrac{dx}{d\bar{t}} = \beta(x,\bar{t}) & \text{for } \bar{t} > 0, \\ x(\bar{x},0) = \bar{x} \end{cases} \tag{248.3}$$

which is the time-dependent analog of the particle path in the stationary case and gives the path in space of a particle moving with speed $\beta(x,t)$.

Note that for time-dependent velocity fields, it is important to distinguish between particle paths and streamlines, unlike the case of stationary velocities when the two concepts are the same. Streamlines are related to a time-independent velocity, for instance we may "freeze" the velocity $\beta(x,t)$ for $t=\hat{t}$ and consider the *streamlines* of $\beta(x,\hat{t}\,)$ that solve $dx/d\bar{t}=\beta(x,\hat{t}\,)$. The streamlines are therefore different from the particle paths, which satisfy $dx/d\bar{t}=\beta(x,\bar{t})$, if $\beta(x,\bar{t})$ depends on $\bar{t}$.

It is also important to distinguish between a space-time particle path $(x(\bar{x},\bar{t}),\bar{t})$ and its projection into space $x(\bar{x},\bar{t})$. Space-time particle paths are essential for the construction of the oriented space-time mesh we describe below.

248.1. Compute and plot the space-time particle paths if (a) $\beta=(x_1,1)$. (b) $\beta=(-x_2,x_1)$. (c) $\beta=(\sin(t),\cos(t))$.

248.2.2 Changing from Lagrange to Euler coordinates

The solution of (248.2) defines a map $(\bar{x},\bar{t})\to(x,t)$ by setting $(x,t)=(x(\bar{x},\bar{t}),\bar{t})$ where $x(\bar{x},\bar{t})$ is the position at time $\bar{t}$ of a particle starting at $\bar{x}$ at time zero. Because particle paths fill up space-time and cannot cross, the map is invertible. We illustrate this in Fig. 248.2. We refer to (x,t) as the

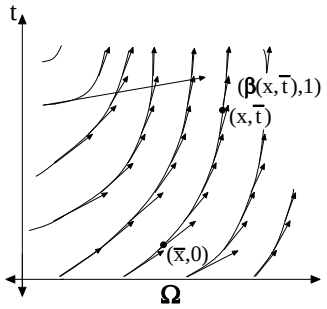


FIGURE 248.2. The vector field $(\beta,1)$ and the corresponding streamlines define a map between the Euler and Lagrange coordinate systems.

Euler coordinates and $(\bar{x},\bar{t})$ as the Lagrange coordinates. An observer using Euler coordinates is anchored at a fixed location x in space and observes the change of some quantity, such as the temperature at x, as time passes, where the change may be caused by the convection bringing new "particles" to the point x. On the other hand, the Lagrange coordinate system moves with the velocity field so that there is no convection relative to the moving coordinate system. The coordinate $\bar{x}$ then acts as a "label" or "marker" attached to "particles" moving along streamlines, where the $\bar{x}$ denotes the position of a particle at time zero, and $x=x(\bar{x},\bar{t})$ is its position at time $\bar{t}$.

In the context of the Dr. Euler and Dr. Lagrange's assistants, the mapping describes the positions of the Dr. Lagrange's crew as they are transported by the current.

248.2. Compute the coordinate map between Euler and Lagrange coordinates corresponding to β in Problem 248.1.

Given a function $u(x,t)$ in Euler coordinates, we define a corresponding function $\bar{u}(\bar{x},\bar{t})$ in Lagrange coordinates by $\bar{u}(\bar{x},\bar{t}) = u(x(\bar{x},\bar{t}),\bar{t})$. By the chain rule,

$$\frac{\partial \bar{u}}{\partial \bar{t}} = \frac{\partial u}{\partial t} + \beta \cdot \nabla u,$$

since $\dfrac{dx}{d\bar{t}} = \beta(x,\bar{t})$ and $t = \bar{t}$. Thus, the convection equation

$$\frac{\partial u}{\partial t} + \beta \cdot \nabla u = f \tag{248.4}$$

in the Euler coordinates (x,t), which is (248.1) with $\epsilon = 0$, takes the simple form

$$\frac{\partial \bar{u}}{\partial \bar{t}} = \bar{f} \tag{248.5}$$

in the *global* Lagrange coordinates $(\bar{x},\bar{t})$, where $\bar{f}(\bar{x},\bar{t}) = f(x(\bar{x},\bar{t}),\bar{t})$. We conclude that in global Lagrange coordinates, the convection term disappears and the original partial differential equation (248.4) reduces to a set of first order ordinary differential equations with respect to $\bar{t}$ indexed by the "marker" $\bar{x}$. In particular, if $f = 0$ then $\bar{u}(\bar{x},\bar{t})$ is independent of time. This makes the job easy for a Lagrange assistant in the sense that if $f = 0$ then it is sufficient to measure the temperature at time equal to zero since the temperature following particles is constant. The Euler assistants on the other hand have to measure the temperature continuously at their fixed location since it may vary even though $f = 0$. Of course, the Lagrange assistants have to keep track of their positions as time passes.

248.3. Compute the solution of $\dot{u} + xu' = f$ for $x \in \mathbb{R}$ and $t > 0$ with

$$f(t) = \begin{cases} t(1-t), & 0 \le t \le 1, \\ 0, & 1 < t \end{cases} \quad \text{and} \quad u_0(x) = \begin{cases} 0, & |x| > 1, \\ 1, & |x| \le 1. \end{cases}$$

by computing the characteristics and changing to Lagrange coordinates.

248.4. Compute the solution of $\dot{u} + (x_1,t) \cdot \nabla u = 0$ for $(x_1,x_2) \in \mathbb{R}^2$ and $t > 0$ with

$$u_0(x) = \begin{cases} 1, & (x_1,x_2) \in [0,1] \times [0,1], \\ 0, & \text{otherwise.} \end{cases}$$

by computing the characteristics and changing to Lagrange coordinates.

Because of the simplicity of (248.5), it is tempting to use Lagrange coordinates. But there is a hook: the Lagrange coordinates have to be computed by solving (248.2) and this is as difficult to solve as the original convection-diffusion problem formulated in Euler coordinates. However, using a kind of "local" Lagrange coordinates, we can avoid solving the equations (248.2) for the global characteristics, while keeping the advantage of the simple form (248.5) in the Lagrange description. The Lagrange coordinates associated to (248.1) underlie the construction of the space-time mesh on each slab S_n used in the chG(0) method in the sense that the space-time mesh in the chG(0) method is oriented approximately along the characteristics of the flow locally on S_n, as shown in Fig. 248.1.

249

Time-Dependent Convection-Diffusion FEM

249.1 The characteristic Galerkin method

The characteristic chG(0) method is based on piecewise constant approximation along space-time characteristics and piecewise linear approximation in space. As usual we let $\{t_n\}$ be an increasing sequence of discrete time levels and associate to each time interval $I_n = (t_{n-1}, t_n)$ a finite element space V_n of piecewise linear continuous functions on a triangulation $\mathcal{T}_n$ of $\mathbb{R}^2$ with mesh function h_n. We use S_n to denote the space-time slab $\mathbb{R}^2 \times I_n$.

249.1.1 Approximate particle paths

We let $\beta_n^h \in V_n$ denote the nodal interpolant of $\beta_n = \beta(\cdot, t_{n-1})$ and introduce the *approximate space-time particle path* $(x_n(\bar{x}, \bar{t}), \bar{t})$ in S_n, where

$$\begin{cases} \dfrac{dx_n}{d\bar{t}} = \beta_n^h(\bar{x}) & \text{in } I_n, \\ x_n(\bar{x}, t_{n-1}) = \bar{x}, \end{cases}$$

or

$$x_n(\bar{x}, \bar{t}) = \bar{x} + (\bar{t} - t_{n-1})\beta_n^h(\bar{x}) \quad \text{for } \bar{t} \in I_n. \tag{249.1}$$

The approximate particle path $(x_n(\bar{x}, \bar{t}), \bar{t})$ is a straight line segment with slope $\beta_n^h(\bar{x})$ starting at $\bar{x}$. We illustrate this in Fig. 249.1.

249.1. Suppose that $\beta = (x_1, 1)$. Plot some of the particle paths and corresponding approximate particle paths for $0 \le t \le .1$ associated to mesh points on the standard uniform triangulation of the square.

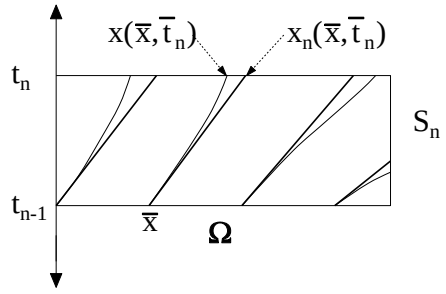


FIGURE 249.1. Exact and approximate particle paths in S_n.

249.1.2 The coordinate mapping

We introduce the coordinate map $F_n : S_n \to S_n$ defined by

$$(x,t) = F_n(\bar{x},\bar{t}) = (x_n(\bar{x},\bar{t}),\bar{t}) \quad \text{for } (\bar{x},\bar{t}) \in S_n,$$

where the coordinates $(\bar{x},\bar{t})$ acts like local Lagrange coordinates on S_n. We illustrate this in Fig. 249.2. We call β_n^h the *tilting velocity* for F_n. Note

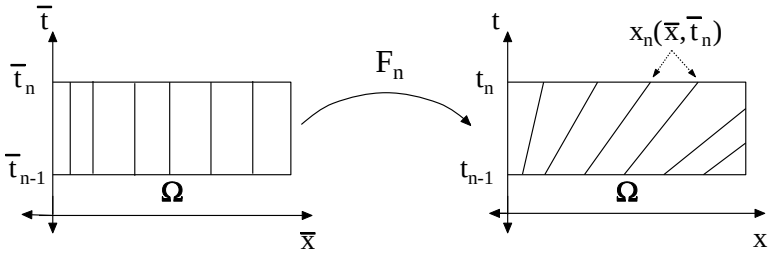


FIGURE 249.2. The map F_n between local Lagrange and Euler coordinates takes a non-oriented grid in $(\bar{x},\bar{t})$ to an oriented grid in (x,t).

that these coordinates are similar but not the same as the global Lagrange coordinates unless β is constant.

Denoting the Jacobian with respect to $\bar{x}$ by $\overline{\nabla}$, we have from (249.1)

$$\overline{\nabla} x_n(\bar{x},\bar{t}) = I + (\bar{t} - t_{n-1})\overline{\nabla}\beta_n^h(\bar{x}),$$

where I denotes the identity, It follows from the inverse function theorem that the mapping $F_n : S_n \to S_n$ is invertible if

$$k_n \|\overline{\nabla}\beta_n^h\|_{L_\infty(R^2)} \leq c, \tag{249.2}$$

with c a sufficiently small positive constant. This condition guarantees that approximate particle paths don't cross in S_n.

249.2. Give an argument showing that F_n is invertible under the condition (249.2)

249.1.3 The finite element spaces for chG(0)

We introduce the space-time finite element space

$$\overline{W}_{kn} = \{\bar{v} : \bar{v}(\bar{x},\bar{t}) = \bar{w}(\bar{x}),\ (\bar{x},\bar{t}) \in S_n \text{ for some } \bar{w} \in V_n\}.$$

To each function $\bar{v}(\bar{x},\bar{t})$ defined on S_n, we associate a function $v(x,t)$ on S_n by

$$v(x,t) = \bar{v}(\bar{x},\bar{t}) \quad \text{for } (x,t) = F_n(\bar{x},\bar{t}).$$

The analog of $\overline{W}_{kn}$ in (x,t) coordinates is

$$W_{kn} = \left\{v : v(x,t) = \bar{v}(\bar{x},\bar{t}),\ (x,t) = F_n(\bar{x},\bar{t}) \in S_n \text{ for some } \bar{v} \in \overline{W}_{kn}\right\}. \tag{249.3}$$

A function v belongs to W_{kn} if the limit v_{n-1}^+ is a continuous piecewise linear function on $\mathcal{T}_n$ and $v(x,t)$ is constant on the straight lines $x = \bar{x} + (t - t_{n-1})\beta_n^h(\bar{x})$ for t in I_n. The corresponding space-time mesh on S_n consists of the elements

$$\mathcal{T}_n^\beta = \{K : K = F_n(\bar{K} \times I_n) \text{ for some } \bar{K} \in \mathcal{T}_n\},$$

which are prisms in space-time "tilted" in the direction of β_n^h illustrated in Fig. 249.3. We use W_k to denote the space of functions v on Q such that

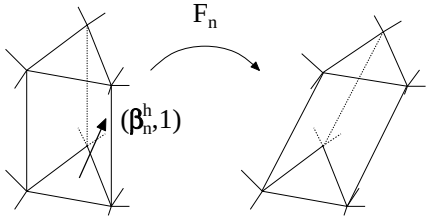


FIGURE 249.3. The normal and the tilted prism elements.

$v|_{S_n} \in W_{kn}$ for $n = 1, 2, ...$

249.3. Assume that $\beta = (x_1, 1)$ and that the standard triangulation is used to discretize the square. Draw a few of the "tilted prisms" for $S_1 = \Omega \times [0, k_1]$.

There are two space meshes associated to each time level t_{n-1}: the mesh $\mathcal{T}_n$ associated to S_n, that is the "bottom mesh" on the slab S_n, and $\mathcal{T}_n^- = \{F_{n-1}(\bar{K} \times \{t_{n-1}\}); \bar{K} \in \mathcal{T}_{n-1}\}$, that is the"top mesh" on the previous slab S_{n-1}, which results from letting the previous "bottom mesh" $\mathcal{T}_{n-1}$ be transported by the flow. The two meshes $\mathcal{T}_n$ and $\mathcal{T}_n^-$ may or may not

coincide. In case they do not match, the L_2 projection is used to interpolate a function on $\mathcal{T}_n^-$ into V_n. Depending on the regularity of the velocity field β, it is possible to maintain matching meshes over a certain length of time simply by choosing $\mathcal{T}_n = \mathcal{T}_n^-$, until the mesh $\mathcal{T}_n^-$ is so distorted that this becomes infeasible. At the other extreme, we may use the same mesh $\mathcal{T}_n$ for all slabs S_n and perform the projection from $\mathcal{T}_n^-$ to $\mathcal{T}_n$ at every time step. We illustrate this in Fig. 249.4.

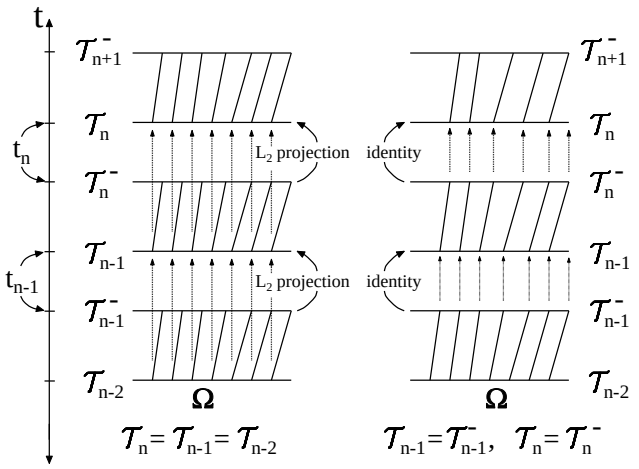


FIGURE 249.4. Two possibilities for constructing grids on succeeding slabs S_n.

249.1.4 The characteristic Galerkin method

The characteristic Galerkin method chG(0) reads: Compute $U \in W_k$ such that for $n = 1, 2, ...,$

$$\int_{I_n} (L(U), v)\, dt \;+\; ([U_{n-1}], v_{n-1}^+) + \int_{I_n} (\hat{\epsilon}\nabla U, \nabla v)\, dt = \int_{I_n} (f, v)\, dt \quad \text{for all } v \in W_{kn}, \quad (249.4)$$

with

$$\begin{aligned} L(U) &= \dot{U} + \beta \cdot \nabla U \quad \text{on } S_n, \\ \hat{\epsilon} &= \max\{\epsilon, \gamma_1 h^2 R(U), \gamma_2 h^{3/2}\} \quad \text{on } S_n, \\ R(U) &= |L(U) - f| + |[U_{n-1}]|/k_n \quad \text{on } S_n, \end{aligned}$$

where γ_1 and γ_2 are non-negative constants to be specified and $[U_{n-1}]$ is extended to S_n as a constant along the characteristics $x_n(\bar{x}, \cdot)$. We have

chosen the streamline diffusion parameter $\delta = 0$ because, as we shall see, the use of tilted elements effectively reduces the convection term, so that no streamline diffusion is needed unless β is non-smooth.

Rewriting (249.4) in local Lagrange coordinates on S_n displays the effect of the orientation. Extending β_n^h to S_n by setting $\beta_n^h(x,t) = \beta_n^h(\bar{x})$ if $(x,t) = F_n(\bar{x},\bar{t})$, the chain rule implies

$$\begin{aligned}\frac{\partial v}{\partial t} + \beta\cdot\nabla v &= \frac{\partial v}{\partial t} + \beta_n^h\cdot\nabla v + (\beta-\beta_n^h)\cdot\nabla v\\ &= \frac{\partial \bar{v}}{\partial \bar{t}} + (\bar{\beta}-\bar{\beta}_n^h)\cdot J_n^{-1}\overline{\nabla}\bar{v}\\ &= \frac{\partial \bar{v}}{\partial \bar{t}} + \bar{\alpha}\cdot\overline{\nabla}\bar{v},\end{aligned}$$

where $J_n(\bar{x},\bar{t}) = \dfrac{\partial x}{\partial \bar{x}}(\bar{x},\bar{t})$ and $\bar{\alpha} = J_n^{-T}(\bar{\beta}-\bar{\beta}_n^h)$. Now, (249.2) implies that there is a constant C such that

$$|\bar{\alpha}| \le C|\bar{\beta}-\bar{\beta}_n^h| \quad \text{on } S_n,$$

so that $|\bar{\alpha}| \le C(k_n + h_n^2)$ if β is smooth. Reformulated in $(\bar{x},\bar{t})$-coordinates, the characteristic Galerkin method takes the form: for $n = 1, 2, ...$, compute $\bar{U} = \bar{U}|_{S_n} \in \bar{W}_{kn}$ such that,

$$\int_{I_n}\Big(\frac{\partial \bar{U}}{\partial \bar{t}} + \bar{\alpha}\cdot\overline{\nabla}\bar{U},\bar{v}|J_n|\Big)\,dt + ([\bar{U}_{n-1}],\bar{v}_{n-1}^+|J_n|) + \int_{I_n}(\bar{\epsilon}\widehat{\nabla}\bar{U},\widehat{\nabla}\bar{v}|J_n|)\,dt = \int_{I_n}(\bar{f},\bar{v}|J_n|)\,dt \quad \text{for all } \bar{v}\in\bar{W}_{kn}, \quad (249.5)$$

where $\widehat{\nabla} = J_n^{-1}\overline{\nabla}$.

Comparing (249.4) and (249.5), we see that using the oriented space-time elements transforms the original problem with velocity β on each slab S_n to a problem with small velocity $\bar{\alpha}$ to which the cG(1)dG(0) method is applied on a tensor-product mesh in $(\bar{x},\bar{t})$ coordinates with no tilting. Thus, the tilting essentially eliminates the convection term, which both improves the precision and facilitates the solution of the corresponding discrete system. The price that is payed is the L_2 projection at mesh changes.

249.2 Extension

The chG(0) method directly extends to the higher order chG(r) method with $r \ge 1$ by using an approximate velocity β_n^h on S_n defined by

$$\bar{\beta}_n^h(\bar{x},\bar{t}) = \sum_{j=0}^{r}\bar{t}^j\beta_{nj}^h(\bar{x})$$

where $\beta^h_{nj}(\bar{x}) \in V_n$. The corresponding approximate characteristics are given by $x(\bar{x},\bar{t}) = \bar{x} + \sum_{j=1}^{r+1} \frac{(\bar{t}-t_{n-1})^j}{j}\beta^h_{nj}(\bar{x})$.

249.4. Prove the last statement.

249.3 The streamline diffusion method on an Euler mesh

The cG(1)dG(r) method for the heat equation extends to (246.1) using the streamline diffusion and artificial viscosity modifications of Section 246.3. This corresponds to using a non-oriented space-time mesh. The coresponding cG(1)dG(r) streamline diffusion method is based on the space W^r_k of functions on Q which on each slab S_n belong to the space W^r_{kn} defined by

$$W^r_{kn} = \{v : v(x,t) = \sum_{j=0}^{r} t^j v_j(x), \quad v_j \in V_n, \, (x,t) \in S_n\}.$$

The method takes the form: compute $U \in W^r_k$ such that for $n = 1, 2, \ldots$, and for $v \in W^r_{kn}$,

$$\int_{I_n} \big(L(U), v \, + \, \delta L(v)\big)\, dt + \int_{I_n} (\hat{\epsilon}\nabla U, \nabla v)\, dt \, + \, ([U_{n-1}], v^+_{n-1}) = \int_{I_n} (f, v + \delta L(v))\, dt \quad (249.6)$$

where

$$\begin{aligned} L(w) &= w_t + \beta \cdot \nabla w, \\ \delta &= \frac{1}{2}(k_n^{-2} + h_n^{-2}|\beta|^2)^{-1/2}, \\ \hat{\epsilon} &= \max\{\epsilon, \gamma_1 h^2 R(U), \gamma_2 h^{3/2}\}, \\ R(U) &= |L(U) - f| + |[U_{n-1}]|/k_n \quad \text{on } S_n, \end{aligned}$$

for positive constants γ_i. Note that the streamline diffusion modification $\delta L(v)$ only enters in the integrals over the slab S_n.

249.3.1 Two examples

We present two examples to illustrate the advantages gained in using the chG method, that is the Sd method on an oriented space-time mesh, as compared to the Sd method on a non-oriented space-time mesh. These examples bring up the point that comparing numerical results purely by

comparing the errors in the L_2 norm may not give a complete picture. This is obvious after a moment of thought since a norm does not contain as much information about a function as a picture of the function. In the following examples, we compare results using the Sd method on non-oriented and oriented space-time meshes in computations with roughly the same accuracy in the L_2 norm. We will see, that the quality in the "picture norm" differs considerably.

The first example is a common quite difficult test problem with pure convection. The initial data consisting of a cylinder with a slit shown in Fig. 249.5, which is rotated counterclockwise by $\beta = (\sin(t), \cos(t))$ until time $t = \pi$, or a rotation of 180 degrees. We first plot in Fig. 249.6 the

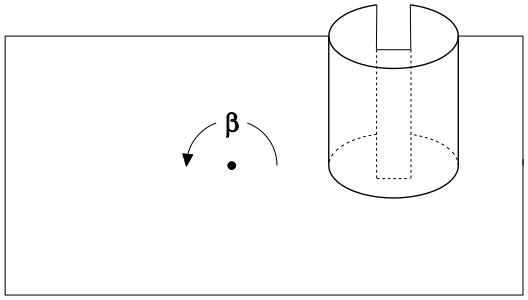


FIGURE 249.5. The initial data for the first example.

results obtained by using the cG(1)dG(1) method on a non-oriented space-time grid reaching one half rotation after 251 constant time steps. Next,

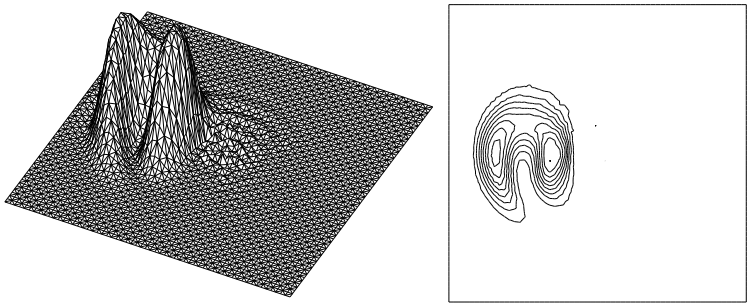


FIGURE 249.6. The approximation and associated level curves from the cG(1)dG(1) streamline diffusion method on a fixed space-time grid applied to the data shown in Fig. 249.5.

in Fig. 249.7 we plot the solution obtained using the chG(0) method. The mesh was tilted in space-time according to the rotating velocity field locally on each space-time slab and an L_2 projection back to a fixed uniform space

mesh was performed at each time step, following the principle illustrated to the left in Fig. 249.4. The solution after a half revolution, is visibly much better than the previous computation, and also computationally much less expensive, because only 21 constant time steps were used and piecewise constants in time where used instead of piecewise linears.

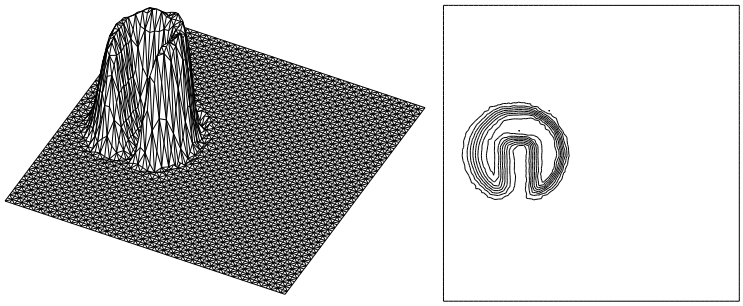


FIGURE 249.7. The approximation and associated level curves from the characteristic Galerkin chG(0) method applied to the data shown in Fig. 249.5.

The next problem, called the Smolarkiewicz example, is a very demanding test problem. The initial data is a cone of height 1 and base radius 15 centered in the rectangular region $\Omega = (25, 75) \times (12.5, 87.5)$. The cone is convectively "folded" in the velocity field

$$\beta = \frac{8\pi}{25}\big(\sin(\frac{\pi x}{25})\sin(\frac{\pi x}{25}), \big(\cos(\frac{\pi x}{25})\cos(\frac{\pi x}{25})\big),$$

which is periodic in x and y with six "vortex cells" inside Ω. We illustrate this in Fig. 249.8. We compute the approximations using 1000 fixed time

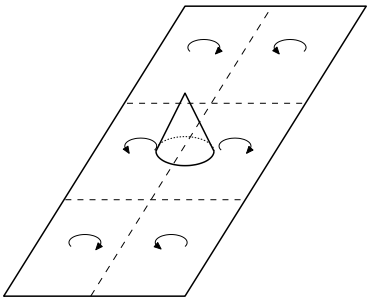


FIGURE 249.8. The initial data for the Smolarkiewicz problem. We also plot the convective vortex cells.

steps to reach the final time $t = 30$. In the first case, we use the chG(0) method without mesh change according to the principle on the right of

Fig. 249.4 with $\mathcal{T}_n^- = \mathcal{T}_n$, so that no L_2 projections from changing the space mesh at a discrete time level were required. We plot the result in Fig. 249.9. Note the extreme mesh distortion that develops as a result of avoiding projections into new, less distorted space meshes. In the second computation, shown in Fig. 249.10, the mesh was changed into a new uniform mesh every one hundred time steps. This limits the mesh distortion but introduces L_2 projection errors at the mesh changes that gradually destroy sharp features of the solution.

249.4 Error analysis

We analyze the chG(0) method applied to the model problem

$$\begin{cases} u_t + \beta \cdot \nabla u - \epsilon \Delta u = f & \text{in } \mathbb{R}^2 \times (0, \infty), \\ u(x,t) \to 0 & \text{for } t > 0 \text{ as } |x| \to \infty, \\ u(\cdot, 0) = u_0 & \text{on } \mathbb{R}^2, \end{cases} \tag{249.7}$$

where $\beta = (\beta_1, \beta_2)$ and $\epsilon \geq 0$ are constant, and f and u_0 are given data with bounded support.

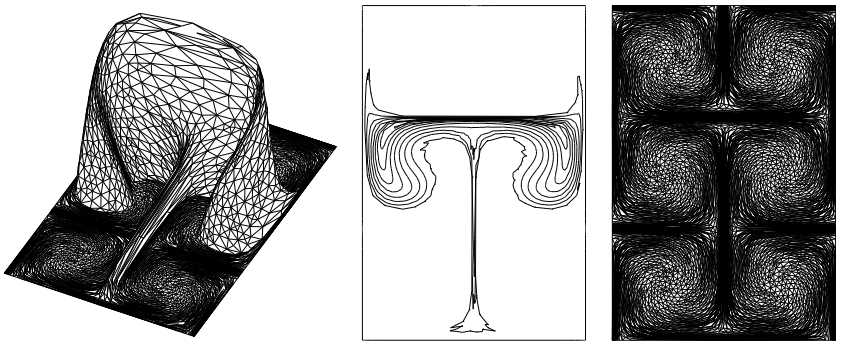


FIGURE 249.9. The approximation, level curves, and mesh resulting from the characteristic Galerkin chG(0) method applied to the Smolarkiewicz problem at $t = 30$. In this computation, the mesh passively follows the flow for all times and no L_2 projections are used.

The transformation between the Euler (x,t) and Lagrange coordinates $(\bar{x}, \bar{t})$ is simply $(x,t) = (\bar{x} + \bar{t}\beta, \bar{t})$ in this case. Reformulating (249.7) in Lagrange coordinates for $\bar{u}(\bar{x}, \bar{t}) = u(x,t)$, after noting that

$$\frac{\partial \bar{u}}{\partial \bar{t}} = \frac{\partial}{\partial \bar{t}} u(\bar{x} + \bar{t}\beta, \bar{t}) = \frac{\partial u}{\partial t} + \beta \cdot \nabla u,$$

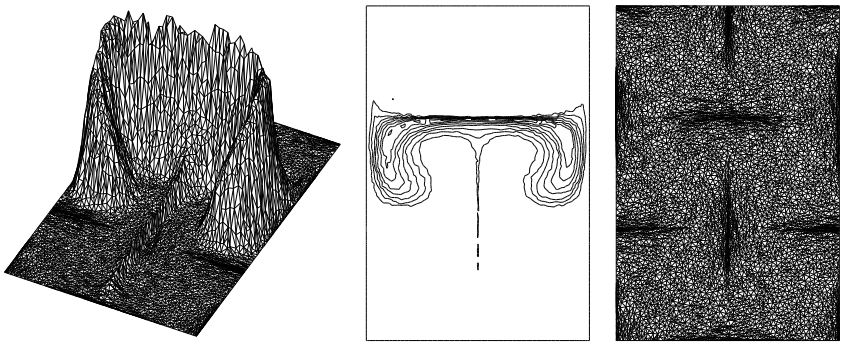


FIGURE 249.10. The approximation, level curves, and mesh resulting from the characteristic Galerkin chG(0) method applied to the Smolarkiewicz problem at $t = 30$. In this computation, an L_2 projection into a uniform mesh is used every one hundred time steps to limit the mesh distortion.

we obtain

$$\begin{cases} \dfrac{\partial \bar{u}}{\partial \bar{t}} - \epsilon \Delta \bar{u} = \bar{f} & \text{in } \mathbb{R}^2 \times (0,\infty), \\ \bar{u}(\bar{x},\bar{t}) \to 0 & \text{for } \bar{t} > 0 \text{ as } |\bar{x}| \to \infty, \\ \bar{u}(\bar{x},0) = u_0(\bar{x}) & \bar{x} \in \mathbb{R}^2. \end{cases} \tag{249.8}$$

We see that the Lagrange formulation (249.8) is the familiar heat equation with constant diffusion coefficient ϵ and the characteristic Galerkin chG(0) method for (249.7) is simply the cG(1)dG(0) method for (249.8).

Before presenting the analysis, we write out the chG(0) method for (249.7) explicitly. By construction, the functions v in W_{kn} are constant in the direction β so that $v_t + \beta \cdot \nabla v = 0$ for $v \in W_{kn}$. Thus, the chG(0) method for (249.7) reduces to: compute $U \in W_k$ such that

$$([U_{n-1}], v_{n-1}^+) + \int_{I_n} (\hat{\epsilon} \nabla U, \nabla v)\, dt = \int_{I_n} (f, v)\, dt \quad \text{for all } v \in W_{kn}, \tag{249.9}$$

where

$$\hat{\epsilon} = \max\left\{\epsilon, \gamma_1 h^2 (|[U_{n-1}]|/k_n + |f|), \gamma_2 h^{3/2}\right\} \quad \text{on } S_n,$$

with $h(x,t) = h_n(x - (t - t_{n-1})\beta)$, where h_n is the mesh function for V_n, and $[U_{n-1}]$ is similarly extended. If $f = 0$ (and ϵ is small), then (249.9) can be written: compute $U_{n-1}^+ \in V_n$ such that

$$\int_{\mathbb{R}^2} U_{n-1}^+ v\, dx + \int_{\mathbb{R}^2} \tilde{\epsilon} \nabla U_{n-1}^+ \cdot \nabla v\, dx = \int_{\mathbb{R}^2} U_{n-1}^- v\, dx \quad \text{for all } v \in V_n, \tag{249.10}$$

with $\tilde{\epsilon} = \gamma_1 h_n^2 |[U_{n-1}]|$ and $U_0^- = u_0$.

249.4.1 Projection and transport

This leads to an alternate formulation of the chG(0) method. Introducing the translation operator τ_n : $\tau_n v(x) = v(x - k_n\beta)$ and the nonlinear projection $\tilde{P}_n$ into V_n defined by

$$(\tilde{P}_n w, v) + (\tilde{\epsilon}\nabla\tilde{P}_n w, \nabla v) = (w, v) \quad \text{for all } v \in V_n,$$

where $\tilde{\epsilon} = \gamma_1 h_n^2 |w - \tilde{P}_n w|$, we can write (249.10) using the notation $U_n = U_n^-$ as

$$U_n = \tau_n \tilde{P}_n U_{n-1}, \tag{249.11}$$

and $U_0 = u_0$.

249.5. Assuming that $\gamma_1 = 0$, show that (249.11) reduces to

$$U_n = \tau_n P_n U_{n-1}, \tag{249.12}$$

where P_n is the L_2-projection into V_n.

Thus, the chG(0) method in the simplest case may be viewed as an algorithm of the form "projection then exact transport". This view is useful for understanding some properties of the chG(0) method, but the chG(0) method is not derived from this concept because this complicates the extension to more complex situations with β variable and ϵ positive.

249.4.2 A direct a priori error analysis in a simple case

We first derive an a priori error estimate for the chG(0) method in the simple case with $f = \epsilon = 0$, where the solution of (249.8) is simply given by $\bar{u}(\bar{x}, t) = u_0(\bar{x})$ and that of (249.7) by

$$u(x, t) = u_0(x - t\beta).$$

Using the formulation (249.12), we write the error as

$$u_n - U_n = \tau_n(u_{n-1} - P_n U_{n-1}) = \tau_n(u_{n-1} - P_n u_{n-1} + P_n(u_{n-1} - U_{n-1})).$$

Using the facts $\|\tau_n v\| = \|v\|$ and $\|P_n v\| \le \|v\|$, we obtain by Pythagoras' theorem

$$\begin{aligned}\|u_n - U_n\|^2 &= \|u_{n-1} - P_n u_{n-1}\|^2 + \|P_n(u_{n-1} - U_{n-1})\|^2 \\ &\le \|u_{n-1} - P_n u_{n-1}\|^2 + \|u_{n-1} - U_{n-1}\|^2.\end{aligned}$$

Iterating this inequality and using a standard error estimate for the L_2 projection, we obtain

$$\begin{aligned}\|u_N - U_N\| &\le \left(\sum_{n=1}^{N} \|u_{n-1} - P_n u_{n-1}\|^2\right)^{1/2} \\ &\le C_i \left(\sum_{n=1}^{N} \|h_n^2 D^2 u_{n-1}\|^2\right)^{1/2} \le C_i \sqrt{N} h^2,\end{aligned}$$

provided u is smooth and we set $h = \max_n h_n$. This estimate is slightly sub-optimal because of the factor $\sqrt{N}$. In the generic case with $k_n \approx h$, we conclude that

$$\|u_N - U_N\| \leq C_i h^{3/2}. \tag{249.13}$$

An optimal result can be derived if the viscosity is positive, as we show in the next section.

249.6. Assuming that the time steps are constant $k_n = k = T/N$, prove that

$$\|u_N - U_N\| \leq C_i \sqrt{N/T}\, \|(I - P)u\|_{L_2(Q)}, \tag{249.14}$$

where $P = P_n$ on S_n. This result is also sub-optimal in comparison with the accuracy of the L_2 projection.

249.4.3 Orientation of Space-Time Mesh

The orientation of the space-time mesh in the characteristic Galerkin method is chosen according to the flow velocity. In general, we could choose an arbitrary mesh translation velocity. We refer to this variant as the *oriented streamline diffusion method*. For example, if the solution is constant in time, this suggests an orientation with zero velocity, which in general is different from orientation along the flow velocity.

249.4.4 An error analysis based on error estimates for parabolic problems

The a priori and a posteriori results for the cG(1)dG(0) method for the heat equation apply to the chG(0) method for (249.7) written in the Lagrange form (249.5). We write out the a priori and a posteriori error estimates, which translate to corresponding optimal estimates for the chG(0) method immediately.

Theorem 249.1 *If $\mu k_n \epsilon \geq h_n^2$, μ sufficiently small, $\bar{\alpha} = 0$ and $\hat{\epsilon} = \epsilon$, then*

$$\|\bar{u}(\cdot, t_N) - \bar{U}_N\| \leq L_N C_i \max_{1 \leq n \leq N} \left(k_n \left\| \frac{\partial \bar{u}}{\partial t} \right\|_{I_n} + \|h_n^2 D^2 \bar{u}\|_{I_n} \right), \tag{249.15}$$

and

$$\|\bar{u}(\cdot, t_N) - \bar{U}_N\| \leq L_N C_i \max_{1 \leq n \leq N} \big(\|k R_{0k}(\bar{U})\|_{I_n} + \|\frac{h_n^2}{\epsilon k_n} [\bar{U}_{n-1}]\|^\star + \|h_n^2 R(\bar{U})\|_{I_n} \big), \tag{249.16}$$

where $L_N = (\max((log(t_N/k_N))^{1/2}, log(t_N/k_N)) + 2$, $R_{0k}(\bar{U}) = |f| + |[\bar{U}]|/k$, $R(\bar{U}) = \frac{1}{\epsilon}|f| + R_2(\bar{U})$ with R_2 defined by (247.28), and a star indicates that the corresponding term is present only if $V_{n-1} \not\subseteq V_n$.

The assumption that

$$k_n\epsilon \geq h_n^2$$

means that $\epsilon > 0$ is needed to get optimal estimates. In the case $k_n = h$ and $\epsilon \approx h^{\frac{3}{2}}$, the estimates reduce to (249.13) if $\partial\bar{u}/\partial\bar{t}$ is small. In the case of pure convection with $f = 0$, when $\frac{\partial\bar{u}}{\partial\bar{t}} = 0$, (249.15) reduces to

$$\|u_N - U_N\| \leq C_i\|(I-P)u\|_{[0,t_N]},$$

where $(I - P)u = (I - P_n)u$ on I_n. This shows that the chG(0) method in the convection dominated case is optimal compared to projection if the viscosity is not too small, cf. (249.14).

249.7. Prove Theorem 249.1.

> Leibniz's spirit of inquiry is apparent even in his report to the Académie des Sciences in Paris about a talking dog. Leibniz describes the dog as a common middle-sized dog owned by a peasant. According to Leibniz, a young girl who heard the dog make noises resembling German words decided to teach it to speak. After much time and effort, it learned to pronounce approximately thirty words, including "thé", "caffé", "chocolat", and "assemblée"- French words which had passed into German unchanged. Leibniz also adds the crucial observation that the dog speaks only "as an echo", that is. after the master has pronounced the word; "it seems that the dog speaks only by force, though without ill-treatment". (The Cambridge Companion to Leibniz)

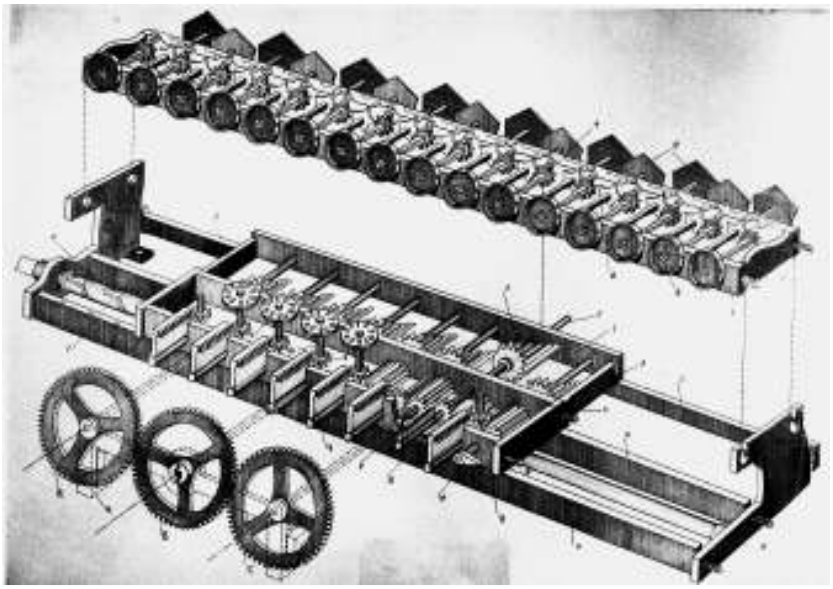


FIGURE 249.11. Leibniz' calculator.

250

The Eigenvalue Problem for an Elliptic Operator

> For since the fabric of the universe is most perfect and the work of a most wise Creator, nothing at all takes place in the universe in which some rule of maximum or minimum does not appear. (Euler)

In this chapter, we briefly consider the *eigenvalue problem* of finding non-zero functions φ and real numbers $\lambda \in \mathbb{R}$ such that

$$\begin{cases} -\nabla \cdot (a\nabla\varphi) + c\varphi = \lambda\varphi & \text{in } \Omega, \\ \varphi = 0 & \text{on } \Gamma, \end{cases} \tag{250.1}$$

where $\Omega \subset \mathbb{R}^d$, Γ is the boundary of Ω, and $a = a(x) > 0$ and $c = c(x)$ are given coefficients. We refer to φ as an *eigenfunction* corresponding to the *eigenvalue* λ. Recall that we discussed the eigenvalue problem in reference to Fourier's method in Chapter **??**. It turns out that the eigenvalues of (250.1) may be arranged as a sequence $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \to \infty$ with one eigenfunction φ_n corresponding to each eigenvalue λ_n. The eigenfunctions corresponding to different eigenvalues are orthogonal with respect to the $L_2(\Omega)$ scalar product and the eigenfunctions corresponding to the same eigenvalue form (together with the zero function) a finite dimensional vector space called the *eigenspace*. The eigenfunctions $\{\varphi_n\}_{n=1}^{\infty}$ may be chosen as an orthonormal basis in $L_2(\Omega)$. In particular, any function $v \in L_2(\Omega)$ can be represented as a series $v = \sum_n v_n\varphi_n$, where $v_n = \int_\Omega v\,\varphi_n\,dx$ are called the *generalized Fourier coefficients*. See Strauss ([**?**]) for more information on these results.

With $a = 1$ and $c = 0$, we obtain the eigenvalue problem for the Laplace operator with Dirichlet boundary conditions

$$\begin{cases} -\Delta\varphi = \lambda\varphi & \text{in } \Omega, \\ \varphi = 0 & \text{on } \Gamma. \end{cases} \tag{250.2}$$

We recall that in the corresponding problem in one dimension with $\Omega = (0,\pi)$, the eigenfunctions are (modulo normalization) $\varphi_n(x) = \sin(nx)$ corresponding to eigenvalues $\lambda_n = n^2$, $n = 1, 2, ...$ In the case $d = 2$ and $\Omega = (0,\pi)\times(0,\pi)$, the eigenfunctions are $\varphi_{nm}(x_1, x_2) = \sin(nx_1)\sin(mx_2)$, $n, m = 1, 2, ...$, with eigenvalues $\lambda_{nm} = n^2 + m^2$. In the first case, all of the eigenspaces have dimension one, but in higher dimensions, all of the eigenspaces except for the eigenspace corresponding to the smallest eigenvalue have dimension larger than one.

250.1. Prove that eigenvalues of (250.2) are positive and that eigenfunctions corresponding to different eigenvalues are orthogonal in $L_2(\Omega)$.

The drum and the guitar

The motion of an elastic membrane supported at the edge along a curve Γ in the plane bounding the domain Ω, is described by the homogeneous wave equation

$$\begin{cases} \ddot{u} - \Delta u = 0 & \text{in } \Omega\times(0,T), \\ u = 0 & \text{on } \Gamma\times(0,T), \\ u(0) = u_0,\ \dot{u}(0) = \dot{u}_0 & \text{in } \Omega, \end{cases} \tag{250.3}$$

where $u(x,t)$ represents the transversal deflection of the membrane. If φ_n is an eigenfunction with corresponding eigenvalue λ_n of the eigenvalue problem (250.2), then the functions $\sin(\sqrt{\lambda_n}t)\varphi_n(x)$ and $\cos(\sqrt{\lambda_n}t)\varphi_n(x)$ satisfy the homogeneous wave equation (250.3) with specific initial data. These functions are called the *normal modes* of vibration of the membrane. A general solution u of the homogeneous wave equation (250.3) with initial values $u(0)$ and $\dot{u}(0)$ can be expressed as a linear combination of the normal modes $\sin(\sqrt{\lambda_n}t)\varphi_n(x)$ and $\cos(\sqrt{\lambda_n}t)\varphi_n(x)$. This is the Fourier's solution of the wave equation on $\Omega\times(0,T)$, which is analogous to the solution in one dimension given by (244.11).

If Ω is a circular disc, then (250.3) describes the vibrations of a drum head. The smallest eigenvalue corresponds to the basic tone of the drum. This can be changed by changing the tension of the drum head, which corresponds to changing the coefficient a in the generalization (250.1).

In Fig. 250.1, we show contour plots for the first four eigenfunctions, corresponding to $\lambda_1 \approx 38.6$, $\lambda_2 \approx 83.2$, $\lambda_3 \approx 111.$, and $\lambda_4 \approx 122.$, computed using Femlab in a case where (250.2) describes the vibrations of the lid of a guitar with Dirichlet boundary conditions on the outer boundary, described

as an ellipse, and Neumann boundary conditions at the hole in the lid, described as a circle.[1] The distribution of the eigenvalues determine the sound produced by the guitar lid and the computational results could be used to find good shapes of a guitar lid.

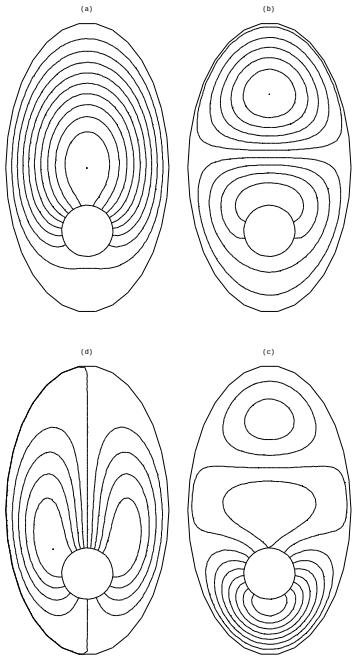


FIGURE 250.1. Contour plots of the first four eigenfunctions of the guitar lid corresponding to (a) $\lambda_1 \approx 38.6$, (b) $\lambda_2 \approx 83.2$, (c) $\lambda_3 \approx 111.$, and (d) $\lambda_4 \approx 122.$. These were computed with Femlab with a fixed mesh size of diameter .02.

Often the smaller eigenvalues are the most important in considerations of design. This is the case for example in designing suspension bridges, which must be built so that the lower eigenvalues of vibrations in the bridge are not close to possible wind-induced frequencies. This was not well understood in the early days of suspension bridges which caused the famous collapse of the Tacoma bridge in 1940.

[1] Computations provided courtesy of Marten Levenstam. The eigenvalues were computed in Femlab using a filtered k-step Arnoldi method as described in D.C. Sorensen, SIAM J. Matrix Anal. Appl. 13 (1992), pp. 357–385.

250.0.5 The Rayleigh quotient

The variational form of the eigenvalue problem (250.1) is to find $\lambda \in \mathbb{R}$ and a non-zero $\varphi \in V$ such that

$$(a\nabla\varphi, \nabla\psi) + (c\varphi, \psi) = \lambda(\varphi, \psi) \quad \text{for all } \psi \in V, \tag{250.4}$$

where

$$V = \left\{ v : \int_\Omega (a|\nabla v|^2 + v^2)\, dx < \infty,\ v = 0 \text{ on } \Gamma \right\},$$

and $(\cdot, \cdot)$ as usual denotes $L_2(\Omega)$ inner product. Setting $\psi = \varphi$ gives a formula for the eigenvalue corresponding to φ,

$$\lambda = \frac{(a\nabla\varphi, \nabla\varphi) + (c\varphi, \varphi)}{(\varphi, \varphi)}.$$

Introducing the *Rayleigh quotient*

$$RQ(\psi) = \frac{(a\nabla\psi, \nabla\psi) + (c\psi, \psi)}{(\psi, \psi)} \quad \text{for } \psi \in V,$$

the previous equality can be rewritten as $\lambda = RQ(\varphi)$, or in words: the Rayleigh quotient of an eigenfunction is equal to the corresponding eigenvalue.

We can turn this argument around and consider how $RQ(\psi)$ varies as ψ varies in V. In particular, there is a function $\varphi_1 \in V$ that minimizes the Rayleigh quotient over all functions in V and this function is the eigenfunction corresponding to the smallest eigenvalue λ_1:

$$\lambda_1 = \min_{\psi \in V} RQ(\psi) = RQ(\varphi_1). \tag{250.5}$$

More generally, the eigenfunction φ_j minimizes the Rayleigh quotient over all functions in V orthogonal to the eigenfunctions φ_i, $i = 1, 2, \ldots j-1$, and $\lambda_j = RQ(\varphi_j)$.

250.2. State and prove the analog of the Rayleigh quotient minimum principle for a diagonal matrix.

250.3. Suppose $\varphi_1 \in V$ minimizes the Rayleigh quotient. Prove that φ_1 is the eigenfunction corresponding to a smallest eigenvalue λ_1 satisfying (250.4) with $\lambda = \lambda_1$. Hint: Define the function $f(\epsilon) = RQ(\varphi + \epsilon\psi)$, where $\psi \in V$ and $\epsilon \in \mathbb{R}$, and use that $f'(0) = 0$.

250.4. Consider the problem of finding the smallest interpolation constant C_i in an error estimate of the form $\|v - \pi v\|_{L_2(0,1)} \le C_i \|v'\|_{L_2(0,1)}$, where $\pi v \in \mathcal{P}^1(0,1)$ interpolates $v(x)$ at $x = 0, 1$. Hint: show first that it suffices to consider the case $v(0) = v(1) = 0$ with $\pi v = 0$. Then rewrite this problem as a problem of determining the smallest eigenvalue and show that $C_i = 1/\pi$. Show similarly that the best constant C_i in the estimate $\|v - \pi v\|_{L_2(0,1)} \le C_i \|v''\|_{L_2(0,1)}$ is equal to $1/\pi^2$.

250.1 Computation of the smallest eigenvalue

We consider the computation of the smallest eigenvalue in the eigenvalue problem (250.2) for the Laplacian by minimizing the Rayleigh quotient over the usual finite element subspace $V_h \subset V$ consisting of continuous piecewise linear functions vanishing on Γ,

$$\lambda_1^h = \min_{\psi \in V_h} RQ(\psi). \tag{250.6}$$

The difference between λ_1 given by (250.5) and λ_1^h given by (250.6) is that the minimization in (250.6) is over the finite dimensional vector space V_h instead of V. Since $V_h \subset V$, we must have $\lambda_1^h \geq \lambda_1$. The question is thus how much larger λ_1^h is than λ_1. To answer this question, we prove an a priori error estimate showing the error in the smallest eigenvalue is bounded by the square of the energy norm interpolation error of the eigenfunction φ_1. This result extends to approximation of larger eigenvalues λ_j with $j > 1$, but the proof is more subtle in this case. We comment on computation of larger eigenvalues in the next section and in the companion volume.

Theorem 250.1 *There is a constant C_i such that for h sufficiently small,*

$$0 \leq \lambda_1^h - \lambda_1 \leq C_i \|hD^2\varphi_1\|^2. \tag{250.7}$$

Assume φ satisfies (250.2) with $\|\varphi\| = 1$ with corresponding eigenvalue $\lambda = RQ(\varphi) = \|\nabla\varphi\|^2$. We shall use the following identity for all $v \in V$ with $\|v\| = 1$, which follows from the definition of φ

$$\|\nabla v\|^2 - \lambda = \|\nabla(\varphi - v)\|^2 - \lambda\|\varphi - v\|^2.$$

250.5. Prove this identity.

Using this identity with $v \in V_h$, $\lambda = \lambda_1$ and $\varphi = \varphi_1$, and recalling the characterization (250.6), we obtain

$$\lambda_1^h - \lambda_1 \leq \|\nabla v\|^2 - \lambda_1 \leq \|\nabla(\varphi_1 - v)\|^2. \tag{250.8}$$

We now take $v \in V_h$ to be a suitable approximation of φ_1 such that $\|\nabla(\varphi_1 - v)\| \leq C\|hD^2\varphi_1\|$, which may put a condition on the size of h because of the restriction $\|v\| = 1$, and the desired result follows.

250.6. Verify that it is possible to find the approximation v to φ_1 required in the proof of (250.7).

250.7. Derive an a posteriori error error estimate for $\lambda_1^h - \lambda_1$. Hint: multiply the equation $-\Delta\varphi_1 - \lambda_1\varphi_1 = 0$ satisfied by the continuous eigenfunction φ_1 corresponding to λ_1, by the discrete eigenfunction $\Phi_1 \in V_h$ satisfying $(\nabla\Phi_1, \nabla v) = \lambda_1^h(\Phi_1, v)$ for all $v \in V_h$, to get

$$(\lambda_1 - \lambda_1^h)(\varphi_1, \Phi_1) = (\nabla\Phi_1, \nabla(\varphi_1 - \pi_h\varphi_1) - \lambda_1^h(\Phi_1, \varphi_1 - \pi_h\varphi_1),$$

where φ_1 and Φ_1 are normalized to have L_2 norm equal to one. Assuming that $(\varphi_1, \Phi_1) \geq c > 0$, where c is a positive constant, derive an a posteriori error estimate in the usual way. (see M. Larsson, A posteriori error estimates for eigenvalue problems, to appear).

250.2 On computing larger eigenvalues

We give an example illustrating the approximation of the larger eigenvalues. In principle, larger eigenvalues and their associated eigenfunctions could be computed in the same fashion as the first eigenpair by finding the stationary points of the Rayleigh quotient over the appropriate finite element space. However, since the eigenfunctions corresponding to larger eigenvalues generally oscillate at larger frequencies, we expect the accuracy of the approximations on a fixed mesh to deteriorate with increasing eigenvalues. In fact, the eigenvalues of the continuous problem tend to infinity, while those of the finite element approximation are finite, so some of the eigenvalues of the continuous problem cannot be captured in the approximation no matter how small we choose the mesh size.

As an example, we consider a finite element discretization of the weak form of the eigenvalue problem (250.2) with $\Omega = (0, \pi)$, which reads: compute $\Phi \in V_h$ and $\lambda_h \in \mathbb{R}$ such that

$$(\Phi', \psi') = \lambda_h(\Phi, \psi) \quad \text{for all } \psi \in V_h, \tag{250.9}$$

where V_h is the space of continuous piecewise linear functions, vanishing at $x = 0, \pi$, on a uniform discretization of $(0, \pi)$ into $M + 1$ elements with meshsize $h = \pi/(M + 1)$ and nodes $x_j = jh$. We also use lumped mass quadrature to evaluate the integral on the right-hand side of (250.9). This gives the matrix eigenvalue problem

$$A\xi = \lambda\xi, \tag{250.10}$$

where ξ denotes the vector of nodal values of Φ and the coefficient matrix A is the product of the inverse of the diagonal lumped mass matrix and the stiffness matrix; cf. Section 240.5. Let ξ_n, $n = 1, 2, ..., M$ be the eigenvectors of (250.10) and Φ_n the corresponding finite element approximations.

250.8. Compute the finite element approximation of (250.9) using lumped mass quadrature and derive (250.10).

We know that the eigenvalues of the continuous problem are n^2, $n = 1, 2, ...$, with corresponding eigenfunctions $\varphi_n(x) = \sin(nx)$. It turns out in this very special case that the nodal values of the discrete eigenfunctions Φ_n agree with the nodal values of the exact eigenfunctions $\sin(nx)$ for $n = 1, ..., M$, that is $\Phi_n(jh) = \sin(njh)$, $n, j = 1, 2, ..., M$.

250.9. Prove by substitution that Φ_n is an eigenvector satisfying (250.10) with eigenvalue $\lambda_n^h = 2(1-\cos(nh))/h^2$ for $n = 1, 2, ..., N$.

When n is small the discrete eigenvalue λ_n^h is a good approximation of the continuous eigenvalue λ_n since by Taylor's theorem

$$\frac{2(1-\cos(nh))}{h^2} \approx n^2 + O(n^4h^2), \tag{250.11}$$

However, despite the interpolation property of the discrete eigenfunctions the L_2 norm of the error $\|\Phi_n - \varphi_n\|$, or even worse the energy norm of the error $\|\Phi_n' - \varphi_n'\|$, becomes large when n gets close to M, see Fig. 250.2. In this case,

$$\frac{2(1-\cos(nh))}{h^2} \approx \frac{4}{h^2} - \frac{(M-n)^2}{2}, \tag{250.12}$$

which is not close to n^2. In Fig. 250.3 we show the first 100 continuous and discrete eigenvalues for $M = 100$. We conclude that eigenvalues corresponding to eigenfunctions that oscillate with a wavelength on the order of the meshsize and smaller are not well approximated.

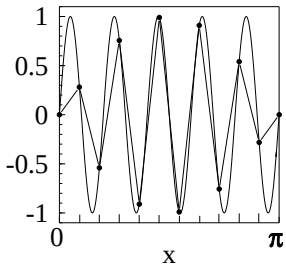


FIGURE 250.2. $\sin(10x)$ and $\Phi_{10}(x)$ for $M = 10$.

250.10. Verify estimates (250.11) and (250.12).

250.11. Define $f(x) = 100^2 \sin(100x)$ (i.e. $f(x) = \lambda_{100}\varphi_{100}$). Use Femlab to solve $-u'' = f$ on $(0, \pi)$ together with $u(0) = u(\pi) = 1$. Plot the approximation together with the true solution. How many elements did Femlab use to get an accurate approximation? Explain the significance of this for the discussion above.

This can have a strong consequences for the time behavior of a discrete approximation to a time dependent problem such as the heat equation or the wave equation. The following problem is an interesting illustration.

250.12. Consider the initial-boundary value problem for the wave equation:

$$\begin{cases} \ddot{u} - u'' = 0, & x \in [0, \pi],\ t > 0, \\ u(0,t) = u(\pi,t) = 0, & t > 0, \\ u(x,0) = u_0(x), \dot{u}(x,0) = 0, & x \in [0, \pi]. \end{cases} \tag{250.13}$$

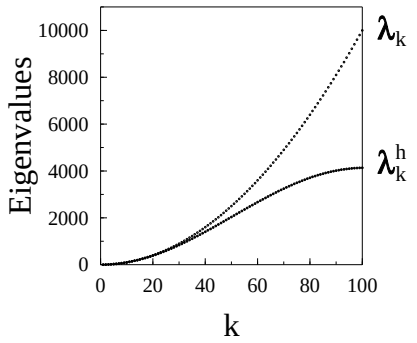


FIGURE 250.3. The continuous and discrete eigenvalues with $M = 100$.

Let U denote the continuous piecewise linear semi-discrete finite element approximation computed on a uniform mesh on $[0, \pi]$. Compare the time behavior of U to that of u when the initial data u_0 is nonsmooth. Can you say something about the time behavior of a finite element approximation that is discrete in time and space? Hint: discretize (250.13) in space using the finite element method on a uniform mesh as indicated. Now use separation of variables to get a scalar ordinary differential equation in time and a matrix eigenvalue problem in space, then solve both problems. Nonsmooth functions are characterized by large Fourier coefficients in the higher modes, so choose the data to be the discrete eigenfunction Φ_M. Compare the solution of (250.13) to the solution of the system of ordinary differential equations as time passes. Plot the two solutions.

250.13. Consider the finite element approximation of (250.2) with $\Omega = (0, \pi) \times (0, \pi)$ computed using the standard triangulation and continuous piecewise linear functions. (a) Compute the discrete eigenvalues and eigenfunctions. Hint: use separation of variables and Problem 250.9. (b) Estimate the convergence rate of the Jacobi iterative method for solving $A\xi = b$.

250.3 The Schrödinger equation for the hydrogen atom

> It does not require much imagination to see an analogy between the mirroring activity of the Leibniz monad, which appears to our confused vision like a casual activity, emanating from one monad and impinging on the other, and the modern view in which the chief activity of the electrons consists in radiating to one another. (Wiener)

The quantum mechanical model of a hydrogen atom consisting of one electron orbiting around one proton at the origin, takes the form of the follow-

ing eigenvalue problem in $\Omega = \mathbb{R}^3$:

$$\begin{cases} -\Delta\varphi - \frac{2}{r}\varphi = \lambda\varphi \quad \text{in } \Omega, \\ \int_\Omega \varphi^2\, dx = 1. \end{cases} \tag{250.14}$$

The eigenfunction φ is a *wave function* for the electron describing the position of the electron in the sense that the integral $\int_\omega \varphi^2 dx$ represents the probability that the electron is in the domain $\omega \subset \mathbb{R}^3$. In fact, (250.14) is the eigenvalue problem associated with the *Schrödinger equation*

$$i\dot{\varphi} - \Delta\varphi - \frac{2}{r}\varphi = 0$$

describing the motion of the electron.

The Rayleigh quotient for the eigenvalue problem (250.14) is given by

$$RQ(\psi) = \frac{\int_\Omega |\nabla\psi|^2\, dx - 2\int_\Omega \psi^2/r\, dx}{\int_\Omega \psi^2\, dx}, \tag{250.15}$$

and is defined for $\psi \in V = \{\psi : \int_{\mathbb{R}^3}(|\nabla\psi|^2 + \psi^2/r)dx < \infty\}$. The quantity $\int_\Omega |\nabla\psi|^2\, dx$ represents the kinetic energy of an electron with wave function ψ and $-2\int_\Omega \psi^2/r\, dx$ represents the potential energy corresponding to the attractive Coulomb force between the proton and electron. The equation (250.14) is one of the few equations of quantum mechanics that can be solved analytically and this is due to the spherical symmetry. The eigenvalues are $\lambda_n = -1/n^2$, for integers $n \geq 1$, called the *principal quantum number* and represent energy levels. There are n^2 eigenfunctions corresponding to each energy level λ_n, of which one depends only on the radius, see Strauss ([**?**]). The eigenfunctions are called the *bound states* and the unique eigenfunction corresponding to the smallest eigenvalue is called the *ground state* since it is the bound state "closest" to the proton with the smallest energy. As soon as more than one electron or proton are involved, that is for all atoms except the hydrogen atom, analytical solution of Schrödinger's equation is practically impossible and a variety of approximate solution methods have been developed.

Among other things, the model (250.14) predicts that the electron may jump from one state with eigenvalue λ_i to another with eigenvalue λ_j by emitting or absorbing a corresponding "quantum" of energy $\lambda_i - \lambda_j$ as was observed in the famous experiments of Bohr. Note that the fact that $\lambda_i \geq -1$ implies that the hydrogen atom is stable in the sense that the electron does not fall into the proton.

We note that the domain in (250.14) is the whole of $\mathbb{R}^3$. Looking for solutions in a space V of functions that are square integrable functions means that we exclude certain oscillating solutions of the Schrödinger equation corresponding to free states of the electron. This is related to the existence of solutions $u(x,t)$ of the problem $i\dot{u} - u'' = 0$ in $\mathbb{R}\times\mathbb{R}$ of the form

$u(x,t) = \exp(i\lambda^2 t)\exp(i\lambda x)$ for any $\lambda \geq 0$. The value λ belongs to the "continuous spectrum" for which the corresponding "eigen-functions" are not square integrable. The eigenvalues with eigenfunctions in V belong to the "discrete spectrum".

To discretize the Schrödinger eigenvalue problem (250.14) in $\mathbb{R}^3$, we generally truncate the domain to be finite, say $\{x : |x| < R\}$ for some $R > 0$, and impose suitable boundary conditions, such as Dirichlet boundary conditions, on the boundary $\{x : |x| = R\}$. The relevant choice of R is related to the eigenvalue/eigenfunction being computed and the tolerance level.

250.14. (For amateur quantum physicists) Prove that the hydrogen atom is stable in the sense that the Rayleigh quotient (250.15) satisfies

$$\min_{\psi \in V} RQ(\psi) \geq -4,$$

showing that the electron does not fall into the proton. Hint: estimate $\int_\Omega \psi \frac{\psi}{r}$ using Cauchy's inequality and the following Poincaré inequality for functions $\psi \in V$:

$$\int_\Omega \frac{\psi^2}{r^2}\,dx \leq 4\int_\Omega |\nabla\psi|^2\,dx. \tag{250.16}$$

This shows that the potential energy cannot outpower the kinetic energy in the Rayleigh quotient. To prove the last inequality, use the representation

$$\int_\Omega \frac{\psi^2}{r^2}\,dx = -\int_\Omega 2\psi\nabla\psi\cdot\nabla\ln(|x|)\,dx.$$

resulting from Green's formula, together with Cauchy's inequality.

250.15. (a) Show that the eigenvalue problem (250.14) for the hydrogen atom for eigenfunctions with radial dependence only, may be formulated as the following one-dimensional problem

$$-\varphi_{rr} - \frac{2}{r}\varphi_r - \frac{2}{r}\varphi = \lambda\varphi, \quad r > 0, \quad \varphi(0) \text{ finite}, \quad \int_{\mathbb{R}} \varphi^2 r^2\,dr < \infty, \tag{250.17}$$

where $\varphi_r = \dfrac{d\varphi}{dr}$. (b) Show that $\psi(r) = \exp(-r)$ is an eigenfunction corresponding to the eigenvalue $\lambda = -1$. (b) Is this the smallest eigenvalue? (c) Determine λ_2 and the corresponding eigenfunction by using a change of variables of the form $\varphi(r) = v(r)\exp(-\frac{r}{2})$. (d) Solve (250.17) using Femlab.

250.16. Formulate a two-dimensional analog of (250.14) of physical significance and compute approximate solutions using Femlab.

250.4 The special functions of mathematical physics

The one-dimensional analog of (250.1) is called a *Sturm-Liouville problem*. Such eigenvalue problems occur for example when separation of variables is

used in various coordinate systems, and the corresponding eigenfunctions are the classical *special functions* of mathematical physics. We list some of these functions below together with the corresponding Sturm-Liouville problem.

Bessel's equation

The eigenfunctions $u_n(x)$ and eigenvalues λ_n of Bessel's equation

$$\begin{cases} -\big(xu'\big)' + x^{-1}m^2u = \lambda xu \quad \text{for } 0 < x < 1, \\ |u(0)| < \infty, \quad u(0) = 1, \end{cases} \tag{250.18}$$

are given by $u_n(x) = J_m(\lambda_n^{1/2}x)$ and $\lambda_n = \mu^2$, where μ is a zero of the Bessel function J_m satisfying (250.18) with $\lambda = 1$ for $x \in \mathbb{R}$ and $|u(0)| < \infty$.

Legendre's equation

The eigenfunctions $u_n(x)$ and eigenvalues λ_n of Legendre's equation

$$\begin{cases} -((1-x^2)u')' + (1-x^2)^{-1}m^2u = \lambda u \quad \text{for } 0 < x < 1, \\ |u(-1)| < \infty, \quad |u(1)| < \infty, \end{cases} \tag{250.19}$$

are given by $\lambda_n = n(n+1)$ and

$$u_n(x) = \frac{1}{2^n n!}(1-x^2)^{m/2}\,\frac{d^{m+n}((x^2-1)^n)}{dx^{m+n}}.$$

Tchebycheff's equation

The eigenfunctions $u_n(x)$ and eigenvalues λ_n of Tchebycheff's equation

$$\begin{cases} -\big((1-x^2)^{1/2}u'\big)' = (1-x^2)^{-1/2}\lambda u \quad \text{for } 0 < x < 1, \\ |u(-1)| < \infty, \quad |u(1)| < \infty, \end{cases} \tag{250.20}$$

are given by $\lambda_n = n^2$ and $u_n(x) = 2^{-(n-1)}\cos\big(n\cos^{-1}x\big)$.

250.17. Use the method of separation of variables to solve the Poisson equation on the disc $\{x \in \mathbb{R}^2 : |x| < 1\}$ with homogeneous Dirichlet boundary conditions. Hint: use polar coordinates and the eigenfunctions of Bessel's equation with $m = 0$.

> Plowhand has been my name
> seems like a thousand years or more
> I ain't gonna pick no more cotton,
> I declare I ain't gonna plant no more corn.
> If a mule wants to run away with the world
> oooh Lord, I'll let it go it on.
> I wouldn't tell a mule to get up,
> Naah, if he sit down in my lap.
> I'm through with plowin'
> cause it killed my old grandpap. (R. Howard)

Part XV

Complex Calculus

1872

(From Applied Mathematics Body and Soul, Vol 3, Springer 2003, coauthored with Kenneth Eriksson and Don Estep).

251
Analytic Functions

> A mathematician of the first rank, Laplace quickly revealed himself as only a mediocre administrator, from his first work we saw we had been deceived. Laplace saw no question from its true point of view, he sought subtleties everywhere, had only doubtful ideas, and finally carried the spirit of the infinitely small into administration. (Napoleon)
>
> We arrive at truth, not by reason only, but also by the heart. (Pascal)

In this chapter we give a short account of *analytic functions*, that is, differentiable functions $f : \mathbb{C} \to \mathbb{C}$, taking complex arguments and having complex values. We use heavily the material developed above on Calculus in $\mathbb{R}^d$, $d = 1, 2$, including, the definition of derivative of a function $f : \mathbb{R}^d \to \mathbb{R}^d$, and Green's formulas in $\mathbb{R}^2$.

251.1 The Definition of an Analytic Function

We recall that we can write each complex number $z \in \mathbb{C}$ in the form $z = x + iy$, with $x, y \in \mathbb{R}$ and i the imaginary unit, and we can identify $\mathbb{C}$ with $\mathbb{R}^2$ by identifying $x + iy \in \mathbb{C}$ with $(x, y) \in \mathbb{R}^2$. In particular, $i = (0, 1)$, and $|z| = (x^2 + y^2)^{1/2}$.

Let $f : \Omega \to \mathbb{C}$ be a complex-valued function of a complex variable $z = x + iy \in \Omega$, where $x, y \in \mathbb{R}$ and Ω is an open domain of the complex plane. Decomposing into real and imaginary parts, we can write

$$f(z) = f(x + iy) = u(x, y) + iv(x, y),$$

where $u : \mathbb{R}^2 \to \mathbb{R}$ and $v : \mathbb{R}^2 \to \mathbb{R}$ are the real and imaginary parts of $f(z)$, that is, $u(x, y) = \text{Re } f(z)$ and $v(x, y) = \text{Im } f(z)$, where we thus view $u(x, y)$ and $v(x, y)$ as functions of $(x, y) \in \mathbb{R}^2$ with values in $\mathbb{R}$.

We say that $f : \Omega \to \mathbb{C}$ is *differentiable* at $z_0 \in \Omega$ with derivative $f'(z_0) \in \mathbb{C}$, if for z close to z_0, we have

$$|f(z) - f(z_0) - f'(z_0)(z - z_0)| \leq K_f(z_0)|z - z_0|^2, \tag{251.1}$$

where $K_f(z_0)$ is a non-negative real constant depending on f and z_0. This is a direct extension of the corresponding definition of the derivative of a function $f : \mathbb{R} \to \mathbb{R}$ to a function $f : \mathbb{C} \to \mathbb{C}$, and the usual rules for differentiation of sums, products and quotients directly extend.

We recall that differentiability of a function $f : \mathbb{R} \to \mathbb{R}$ at a point x_0 means that $f(x)$ is well approximated (up to a quadratic term) by the linear function $c_0 + c_1(x - x_0) = f(x_0) + f'(x_0)(x - x_0)$ for x close to x_0, where $c_0 = f(x_0)$ and $c_1 = f'(x_0)$ are real constants. Similarly, differentiability of a function $f : \mathbb{C} \to \mathbb{C}$ at a point z_0 means that $f(z)$ for z close to z_0 is well approximated by the linear function $c_0 + c_1(z - z_0) = f(z_0) + f'(z_0)(z - z_0)$, involving a translation and multiplication by a complex constant. We conclude that differentiability of a function $f : \mathbb{C} \to \mathbb{C}$ at a point z_0 means that $f(z)$ in a neighborhood of z_0 acts like a combination of a translation, rotation and change of modulus, see Fig. 251.1

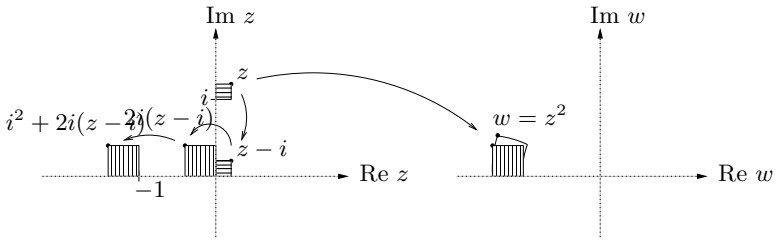


FIGURE 251.1. Linear approximation of a function $f(z)$ near $z = z_0$, with $z_0 = i$ and $w = f(z) = z^2$ approximated by $f(z_0) + f'(z_0)(z - z_0) = i^2 + 2i(z - i)$.

We say that $f : \Omega \to \mathbb{C}$ is *analytic* in the open domain Ω of the complex plane if $f(z)$ is differentiable at all $z_0 \in \Omega$ with derivative $f'(z_0)$. The derivative f' of an analytic function $f : \Omega \to \mathbb{C}$ is again a function $f' : \Omega \to \mathbb{C}$. We shall shortly prove the surprising fact that if $f : \Omega \to \mathbb{C}$ is analytic, then also $f' : \Omega \to \mathbb{C}$ is analytic with derivative $f'' : \Omega \to \mathbb{C}$, which is also analytic, and so on. An analytic function $f : \Omega \to \mathbb{C}$ thus has derivatives of all orders $f^{(n)} : \Omega \to \mathbb{C}$, $n = 1, 2, ...$, which are all analytic. We recall that a differentiable function $f : \mathbb{R} \to \mathbb{R}$ need not have a differentiable derivative, and therefore does not have this very special property in general.

We can view a function $f : \mathbb{C} \to \mathbb{C}$ alternatively as a function $f : \mathbb{R}^2 \to \mathbb{R}^2$ if we identify $\mathbb{C}$ and $\mathbb{R}^2$ as indicated. The Jacobian $f'(x, y)$ of a function $f :$

$\mathbb{R}^2 \to \mathbb{R}^2$ is a 2×2-matrix consisting of 4 real numbers, while the derivative $f'(z) \in \mathbb{C}$ of a function $f : \mathbb{C} \to \mathbb{C}$ is supposed to be a complex number being represented by 2 real numbers. We conclude that differentiability of a complex-valued function $f : \mathbb{C} \to \mathbb{C}$, is a more stringent requirement than differentiability of the corresponding function $f : \mathbb{R}^2 \to \mathbb{R}^2$, which only requires the partial derivatives of the real and imaginary parts $u(x, y)$ and $v(x, y)$ of $f(z)$ to exist. In fact, we shall see that the partial derivatives of the real and imaginary parts of an analytic function must be coupled in a specific way, which is expressed through the *Cauchy-Riemann equations* stated below.

251.2 The Derivative as a Limit of Difference Quotients

Note that (251.1) implies that if $z \neq z_0$, then

$$|\frac{f(z) - f(z_0)}{z - z_0} - f'(z_0)| \leq K_f(z_0)|z - z_0|,$$

which we can write in the form

$$\lim_{z\to z_0} \frac{f(z) - f(z_0)}{z - z_0} = f'(z_0), \tag{251.2}$$

by which we mean, of course, that

$$|\frac{f(z) - f(z_0)}{z - z_0} - f'(z_0)|$$

is as small as we please if we only choose $|z - z_0|$ small enough (respecting that $z \neq z_0$). In view of (251.2) we write as usual $\frac{df}{dz} = f'$.

251.3 Linear Functions Are Analytic

We consider the function $f : \mathbb{C} \to \mathbb{C}$ given by $f(z) = az + b$, where a and b are given complex numbers. We have for all z and $z_0 \in \mathbb{C}$ that

$$f(z) - f(z_0) - a(z - z_0) = 0,$$

and thus $f(z)$ is analytic in $\mathbb{C}$ with derivative $f'(z) = a$.

251.4 The Function $f(z) = z^2$ Is Analytic

If $f(z) = z^2$, then

$$f(z) - f(z_0) - 2z_0(z - z_0) = z^2 - z_0^2 - 2z_0z + 2z_0^2 = (z - z_0)^2,$$

and thus $f'(z_0) = 2z_0$ for $z_0 \in \mathbb{C}$.

251.5 The Function $f(z) = z^n$ Is Analytic for $n = 1, 2, ...$

Consider the function $f : \mathbb{C} \to \mathbb{C}$, where $f(z) = z^n$ and $n = 1, 2, ...$, is a natural number, which may be viewed as an extension of the function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^n$. By a direct extension of the proof in the case $f(x) = x^n$, we find that

$$f'(z) = nz^{n-1}. \tag{251.3}$$

We conclude that z^n is differentiable in the whole of $\mathbb{C}$, with derivative nz^{n-1}. We just gave the proof in the case $n = 1, 2$.

251.6 Rules of Differentiation

As we said, the usual rules for differentiation of sums, products and quotients valid for functions $f : \mathbb{R} \to \mathbb{R}$ extend to functions $f : \mathbb{C} \to \mathbb{C}$. In particular we have if $f(z_0) \neq 0$ and $g(z) = \frac{1}{f(z)}$, that

$$g'(z_0) = -\frac{f'(z_0)}{f^2(z_0)}. \tag{251.4}$$

Further, the composition of two analytic functions is also analytic and the Chain rule for differentiation holds: if $g(z)$ is differentiable at z_0 and $f(z)$ is differentiable at $g(z_0)$, then the composite function $h(z) = f(g(z))$ is differentiable at z_0 with derivative $h'(z_0) = f'(g(z_0))g'(z_0)$. The proof is a direct extension of the corresponding proof for real-valued functions of a real variable.

251.7 The Function $f(z) = z^{-n}$

Applying the rule (251.4), we see that $f(z) = z^{-n}$ with $n = 1, 2, ...$, is differentiable for $z \neq 0$ and

$$f'(z) = -nz^{-n-1}, \quad \text{for } z \neq 0. \tag{251.5}$$

We can summarize by stating that if $f(z) = z^n$ with $n = \pm 1, \pm 2, ...$, then $f'(z) = nz^{n-1}$, where we assume that $z \neq 0$ if $n < 0$.

251.8 The Cauchy-Riemann Equations

We shall now derive the so-called *Cauchy-Riemann equations* connecting the partial derivatives of the real part $u(x,y)$ and the imaginary part $v(x,y)$ of a complex-valued function $f(z) = u(x,y)+iv(x,y)$ of a complex variable $z = x + iy$ at a point $z_0 = x_0 + iy_0$ with $x_0, y_0 \in \mathbb{R}$, such that (251.1) is satisfied. Writing $f'(z_0) = a + ib$, with $a, b \in \mathbb{R}$, we can express (251.1) as

$$|u(x,y) + iv(x,y) - u(x_0,y_0) - iv(x_0,y_0) - (a+ib)(x - x_0 + i(y-y_0))| \\ \leq K_f(z_0)|z - z_0|^2.$$

Separating into real and imaginary parts, we conclude (recalling (91.7)) that

$$\begin{aligned} |u(x,y) - u(x_0,y_0) - a(x-x_0) + b(y-y_0)| &\leq K_f(z_0)|z-z_0|^2, \\ |v(x,y) - v(x_0,y_0) - a(y-y_0) - b(x-x_0)| &\leq K_f(z_0)|z-z_0|^2. \end{aligned} \tag{251.6}$$

Recalling the definition of the partial derivatives of $u(x,y)$ and $v(x,y)$ at (x_0,y_0) from Chapter *Vector-valued functions of several variables*, we conclude that

$$\begin{aligned} a &= \frac{\partial u}{\partial x}(x_0,y_0), \quad b = -\frac{\partial u}{\partial y}(x_0,y_0), \\ a &= \frac{\partial v}{\partial y}(x_0,y_0), \quad b = \frac{\partial v}{\partial x}(x_0,y_0), \end{aligned}$$

and we thus find that

$$\frac{\partial u}{\partial x}(x_0,y_0) = \frac{\partial v}{\partial y}(x_0,y_0), \qquad \frac{\partial u}{\partial y}(x_0,y_0) = -\frac{\partial v}{\partial x}(x_0,y_0). \tag{251.7}$$

These are the *Cauchy-Riemann equations* for $u(x,y)$ and $v(x,y)$ at the point (x_0,y_0).

We conclude that if $f(z) = u(x,y) + iv(x,y)$ is analytic in the open domain Ω of the complex plane, then the real and imaginary parts $u(x,y)$ and $v(x,y)$ satisfy the Cauchy-Riemann equations in Ω, that is,

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}, \qquad \text{in } \Omega. \tag{251.8}$$

Note that we can write the Cauchy-Riemann equations in the form $\nabla v = -\nabla \times u$, recalling that $\nabla \times u = (\frac{\partial u}{\partial y}, -\frac{\partial u}{\partial x})$.

EXAMPLE 251.1. The analytic function $f(z) = z^2 = (x+iy)^2 = x^2 - y^2 + 2ixy$ with $u(x,y) = x^2 - y^2$ and $v(x,y) = 2xy$ satisfies $\frac{\partial u}{\partial x} = 2x = \frac{\partial v}{\partial y}$ and $\frac{\partial u}{\partial y} = -2y = -\frac{\partial v}{\partial x}$.

We have seen that the Cauchy-Riemann equations (251.8) follow from the analyticity of the complex valued function $f = u + iv$. In other words, the Cauchy-Riemann equations represents a *necessary condition* for the analyticity of $f = u + iv$. The Cauchy-Riemann equations also represent a *sufficient condition*: given a pair of functions $u(x, y)$ and $v(x, y)$ satisfying the Cauchy-Riemann equations, the function $f = u + iv$ is analytic. To see this, we note that if $u(x, y)$ and $v(x, y)$ are differentiable functions satisfying (251.8), then (251.6), and thus also (251.1) holds. That is, $f = u + iv$ is analytic.

EXAMPLE 251.2. We consider the functions $u(x, y) = x + 2xy$ and $v(x, y) = y - x^2 + y^2$ and find that $\frac{\partial u}{\partial x} = 1 + 2y = \frac{\partial v}{\partial y}$ and $\frac{\partial u}{\partial y} = 2x = -\frac{\partial v}{\partial x}$, that is, u and v satisfy the Cauchy-Riemann equations and we thus conclude that the function $f(z) = u(x, y) + iv(x, y)$ must be analytic. In fact, $f(z) = u + iv = x + 2xy + i(y - x^2 + y^2) = x + iy - i(x + iy)^2 = z - iz^2$, and the analyticity is obvious.

We may summarize as follows:

Theorem 251.1 *The function $f(z) = u(x, y) + iv(x, y)$ is analytic if and only if the Cauchy-Riemann equations (251.8) are satisfied.*

251.9 The Cauchy-Riemann Equations and the Derivative

Using the limit definition (251.2) of the derivative $f'(z_0)$, we can write, varying first only x:

$$f'(z_0) = \lim_{x \to x_0} \frac{f(z) - f(z_0)}{x - x_0} = \frac{\partial u}{\partial x}(x_0, y_0) + i\frac{\partial v}{\partial x}(x_0, y_0), \tag{251.9}$$

where $z = x + iy_0$, and then only y:

$$f'(z_0) = \lim_{y \to y_0} \frac{f(z) - f(z_0)}{i(y - y_0)} = \frac{1}{i}\frac{\partial u}{\partial y}(x_0, y_0) + \frac{\partial v}{\partial y}(x_0, y_0),$$

where $z = x_0 + iy$, from which the Cauchy-Riemann equations follow by equating the real and imaginary parts of $f'(z_0)$ using that $\frac{1}{i} = -i$.

EXAMPLE 251.3. In the last example we found that $u(x, y) = x + 2xy$ and $v(x, y) = y - x^2 + y^2$ satisfy the Cauchy-Riemann equation. According to (251.9), the derivative of the corresponding analytic function $f = u + iv$ is given by $f' = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} = 1 + 2y + i(-2x)$ which agrees with our observation that $f(z) = z - iz^2$ with $f'(z) = 1 - 2iz = 1 - 2i(x + iy) = 1 + 2y - 2ix$.

EXAMPLE 251.4. By direct verification using the Cauchy-Riemann equations one finds that $f(z) = e^z = e^x(\cos(y) + i\sin(y))$ is analytic in $\mathbb{C}$, and $\frac{d}{dz}e^z = e^z$. It follows that also $\sin(z) = \frac{1}{2i}(e^{iz} - e^{-iz})$ and $\cos(z) = \frac{1}{2}(e^{iz} + e^{-iz})$ are analytic in $\mathbb{C}$, and $\frac{d}{dz}\sin(z) = \cos(z)$ and $\frac{d}{dz}\cos(z) = -\sin(z)$. See Problem 251.1

251.10 The Cauchy-Riemann Equations in Polar Coordinates

The Cauchy-Riemann equations take the following form in polar coordinates $z = re^{i\theta}$:

$$\frac{\partial u}{\partial r} = \frac{1}{r}\frac{\partial v}{\partial \theta}, \quad \frac{\partial v}{\partial r} = -\frac{1}{r}\frac{\partial u}{\partial \theta}. \tag{251.10}$$

EXAMPLE 251.5. The function $\mathrm{Log}(z) = \log(|z|) + i\,\mathrm{Arg}\ z$ is analytic in $\{z \in \mathbb{C} : z \neq 0, 0 \leq \arg z < 2\pi\}$. This follows from the Cauchy-Riemann equations in polar coordinates. We recall that $\log(z) = \log(|z|) + i\arg z$ is multi-valued since $\arg z$ is multivalued. The function $\log(z)$ with $\arg z$ restricted to $0 \leq \arg z < 2\pi$, however, is single-valued analytic.

251.11 The Real and Imaginary Parts of an Analytic Function

We shall now prove that the Cauchy-Riemann equations (251.8) imply that both $u(x, y)$ and $v(x, y)$ are harmonic in Ω, that is,

$$\Delta u = 0 \quad \text{and} \quad \Delta v = 0 \quad \text{in } \Omega.$$

In fact, this follows directly by differentiating (251.8) with respect to x and y, if we assume that $u(x, y)$ and $v(x, y)$ are twice differentiable, since

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 v}{\partial x \partial y} = \frac{\partial^2 v}{\partial y \partial x} = -\frac{\partial^2 u}{\partial y^2} \quad \text{in } \Omega \tag{251.11}$$

and thus $\Delta u = 0$ in Ω , and similarly $\Delta v = 0$ in Ω.

Now, one can show that solutions of the Cauchy-Riemann equations indeed must be twice differentiable, and thus the real and imaginary parts of an analytic function are harmonic. We sum up in the following theorem:

Theorem 251.2 *If $f : \Omega \to \mathbb{C}$ is analytic, where Ω is an open domain of the complex plane $\mathbb{C}$, then the real part $u(x, y) = Re\ f(z)$ and the imaginary part $v(x, y) = Im\ f(z)$ are harmonic in Ω.*

251.12 Conjugate Harmonic Functions

Suppose $u(x,y)$ is harmonic in a simply connected domain Ω in $\mathbb{R}^2$. We shall now prove that there exists a harmonic function $v(x,y)$, uniquely determined up to a constant, such that $f(z) = u(x,y) + iv(x,y)$ is analytic in Ω. We say that the function $v(x,y)$ is *conjugate* to $u(x,y)$. To prove this, we simply solve the Cauchy-Riemann equations $\nabla v = -\nabla \times u$ with u given using the basic result of the Chapter Potential fields, noting that $\nabla \times (-\nabla \times u) = \Delta u = 0$, that is, $-\nabla \times u$ is irrotational, and thus is the gradient of some function v. See also Problem 251.10.

EXAMPLE 251.6. For the harmonic function $u(x_1,x_2) = x_1x_2$, the conjugate $v(x_1,x_2)$ satisfies $\frac{\partial v}{\partial x_2} = \frac{\partial u}{\partial x_1} = x_2$, that is, $v = \frac{1}{2}x_2^2 + C(x_1)$ for some function C, and from $\frac{\partial v}{\partial x_1} = -\frac{\partial u}{\partial x_2} = -x_1$, that is $C'(x_1) = -x_1$, we conclude that $C(x_1) = -\frac{1}{2}x_1^2 + D$, for some arbitrary constant D. We note that also v is harmonic, and conclude that u and its conjugate v are the real and imaginary parts of the analytic function $f(z) = x_1x_2 + i(\frac{1}{2}x_1^2 - \frac{1}{2}x_2^2) = -\frac{1}{2}z^2$.

251.13 The Derivative of an Analytic Function Is Analytic

Assume that $f(z)$ is *analytic* in the open domain Ω of the complex plane. This means that the derivative $f'(z)$ exists as a complex-valued function for $z \in \Omega$, and one may ask if $f'(z)$ itself has a derivative in Ω, that is, if $f'(z)$ is analytic in Ω. The plain answer is YES, which we prove below. Thus, if $f(z)$ is analytic in Ω, then also $f'(z)$ is analytic in Ω, and thus also the derivative of $f'(z)$, that is the second derivative $f''(z)$ is analytic, and so on. We conclude that an analytic function has derivatives of all orders. This is a remarkable property of an analytic function.

To answer the question posed, it is sufficient to notice that if $u(x,y)$ and $v(x,y)$ satisfy the Cauchy-Riemann equations, then so do all derivatives of $u(x,y)$ and $v(x,y)$, in particular $\frac{\partial u}{\partial x}$ and $\frac{\partial v}{\partial x}$, and thus $f' = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}$ is analytic in Ω. We state this important result as a theorem:

Theorem 251.3 *If $f : \Omega \to \mathbb{C}$ is analytic, where Ω is an open domain of the complex plane $\mathbb{C}$, then all the derivatives $f^{(n)}(z)$, $n = 1, 2, ...$, of $f(z)$ are analytic in Ω.*

We recall that if $f : \mathbb{R} \to \mathbb{R}$ is a real-valued function of a real variable, then the analogous statement may be wrong: even if $f'(x)$ exists, it is not clear that $f''(x)$ exists.

251.14 Curves in the Complex Plane

Let Ω be an open domain in the complex plane $\mathbb{C}$, and let $\gamma : I \to \Omega$, where $I = [a, b]$ is an interval of $\mathbb{R}$, be a Lipschitz continuous function. We say that $\Gamma =$ Range of $\gamma = \{\gamma(t) : t \in I\}$, is a *curve* in $\mathbb{C}$ parameterized by $\gamma(t)$. For example $\gamma(t) = \exp(it)$ where $0 \leq t < 2\pi$ is a parametrization of the unit circle.

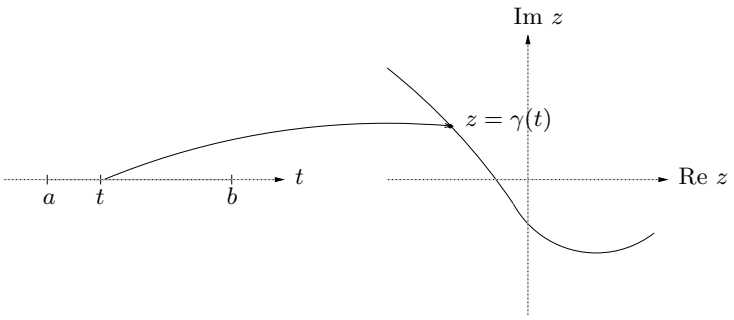


FIGURE 251.2. A curve $z = \gamma(t)$

We say that Γ is a *differentiable curve* if the corresponding parametrization $\gamma : I \to \mathbb{C}$ is differentiable on I in the sense that the related function $\gamma : I \to \mathbb{R}^2$ is differentiable. In other words, decomposing $\gamma(t) = x(t) + iy(t)$ into real and imaginary parts $x : I \to \mathbb{R}$ and $y : I \to \mathbb{R}$, we have that $\gamma(t)$ is differentiable on I if $x(t)$ and $y(t)$ are differentiable on I. We also say that $\gamma : I \to \mathbb{C}$ is Lipschitz continuous on I if the corresponding function $\gamma : I \to \mathbb{R}^2$ is Lipschitz continuous. There are thus no surprises in this context.

A curve Γ with parametrization $\gamma : [a, b] \to \mathbb{C}$ is said to be *closed and simple* if $\gamma(s) \neq \gamma(t)$ for $s < t$, unless $s = a$ and $t = b$.

We say that a domain Ω in $\mathbb{C}$ which is bounded by a simple closed curve, is *simply connected*. A simply connected domain does not have any "holes".

251.15 Conformal Mappings

Let $f : \Omega \to \mathbb{C}$ be analytic where Ω is an open domain in $\mathbb{C}$. We shall now prove that the mapping $z \to w = f(z)$ is *conformal* in Ω in the sense that angles are preserved under the mapping $w = f(z)$. This is a direct consequence of the Chain rule and the analyticity of $f(z)$, as we now show. Let then $\gamma : I \to \mathbb{C}$, where $I = [-\delta, \delta]$ with $\delta > 0$, be a curve through $z_0 \in \Omega$ with $\gamma(0) = z_0$. Consider the curve $\kappa(t) = f(\gamma(t))$ which is the image of $\gamma(t)$ under the transformation $w = f(z)$. By the Chain rule we

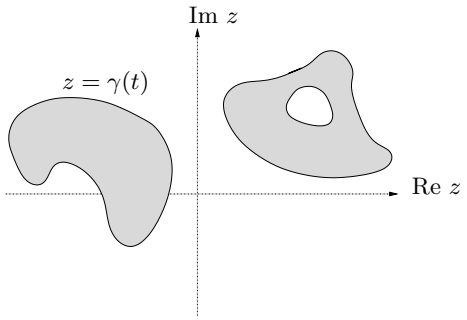


FIGURE 251.3. A simply connected domain with boundary curve $z = \gamma(t)$, to the left, and a multiply connected domain with *one* hole to the right with the boundary consisting of *two* simple closed curves

have

$$\frac{d\kappa}{dt} = \frac{df}{dz}\frac{d\gamma}{dt},$$

and we thus see, recalling that the argument of the product of two complex numbers is the sum of the arguments of the numbers, that

$$\arg \frac{d\kappa}{dt}(0) = \text{Arg } f'(z_0) + \text{Arg } \frac{d\gamma}{dt}(0),$$

where we assume that $f'(z_0) \neq 0$. Since $f(z)$ is analytic at z_0, we have that $f'(z_0)$ is independent of the curve γ, and thus the tangent direction $\frac{d\kappa}{dt}(0)$ differs from that of $\frac{d\gamma}{dt}(0)$ by the constant value Arg $f'(z_0)$, independent of γ. We conclude that the angle between two curves passing through z_0 is the same as the angle between the corresponding transformed curves passing through $f(z_0)$. This means that the mapping $w = f(z)$ is *conformal at* z_0: angles are preserved locally.

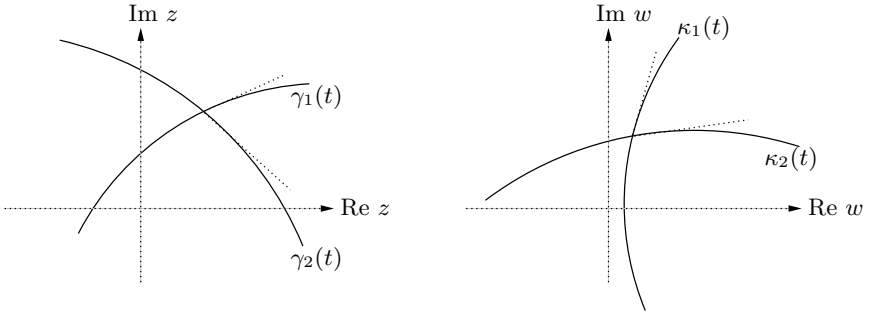


FIGURE 251.4. An analytic mapping conforms angles.

Note that since

$$\lim_{z\to z_0} |\frac{f(z)-f(z_0)}{z-z_0}| = |f'(z_0)|,$$

the mapping $w = f(z)$ changes the length scale locally by the factor $|f'(z_0)| \neq 0$. Thus although the mapping $w = f(z)$ is locally conformal, the image of a large figure may be considerably distorted because of the change of scale.

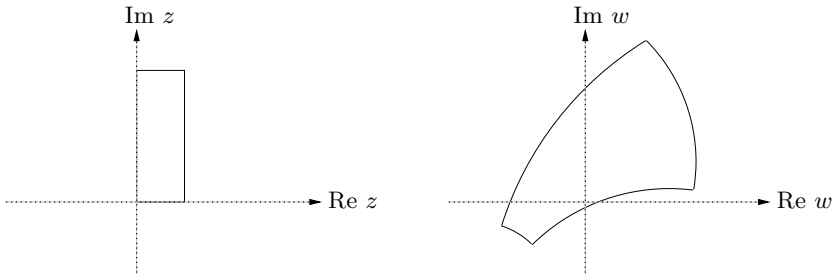


FIGURE 251.5. A conformal mapping with large deformations.

We now present some basic analytic functions $w = f(z)$ and the corresponding conformal mappings of $f : \mathbb{C} \to \mathbb{C}$.

251.16 Translation-rotation-expansion/contraction

The linear transformation:

$$w = f(z) = az + b$$

where $a, b \in \mathbb{C}$, corresponds to a rotation with Arg a and an expansion/contraction with $|a|$, and a translation with b, see Fig. 251.6

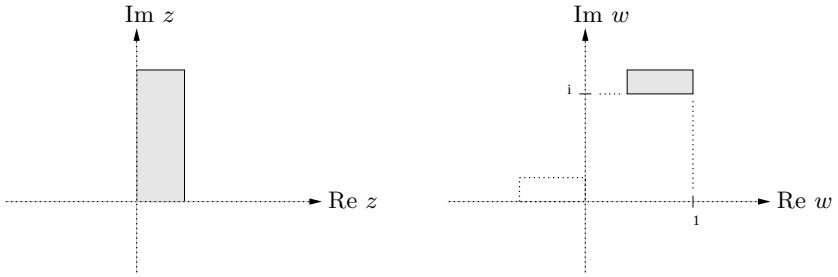


FIGURE 251.6. The mapping $w = az + b$ with $a = \frac{1}{2}i$ and $b = 1 + i$.

251.17 Inversion

The mapping

$$w = f(z) = \frac{1}{z},$$

is referred to as *inversion*. We now prove that an inversion maps every straight line or circle in the complex plane into a circle or straight line. Indeed, a circle or straight line in $\mathbb{R}^2$ can be written

$$A(x^2 + y^2) + Bx + Cy + D = 0$$

with A, B, C, D real, and $A = 0$ corresponding to a straight line. In terms of $z = x + iy$ and $\bar{z} = x - iy$, the equation takes the form

$$Az\bar{z} + B\frac{z + \bar{z}}{2} + C\frac{z - \bar{z}}{2i} + D = 0,$$

and substitution of $z = \frac{1}{w}$ gives (after multiplication with $w\bar{w}$)

$$A + B\frac{\bar{w} + w}{2} + C\frac{\bar{w} - \bar{w}}{2i} + Dw\bar{w} = 0,$$

which represents a circle or straight line.

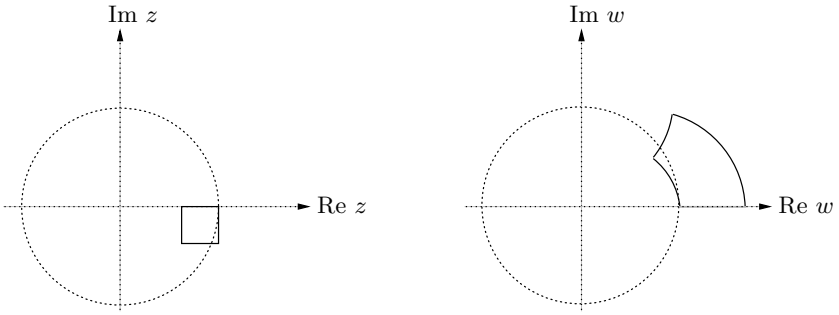


FIGURE 251.7. The mapping $w = 1/z$

251.18 Möbius transformations

A mapping of the form

$$w = f(z) = \frac{az + b}{cz + d},$$

where $a, b, c, d \in \mathbb{C}$, is said to be a *Möbius transformation*. We have

$$f'(z) = \frac{ad - bc}{(cz + d)^2},$$

and we are thus led to assume that $ad - bc \neq 0$ to guarantee conformity. Evidently, the inversion $w = \frac{1}{z}$ is a special case of a Möbius transformation. One can prove that a Möbius transformation maps every straight line or circle in the complex plane into a circle or straight line, see Problem 251.6.

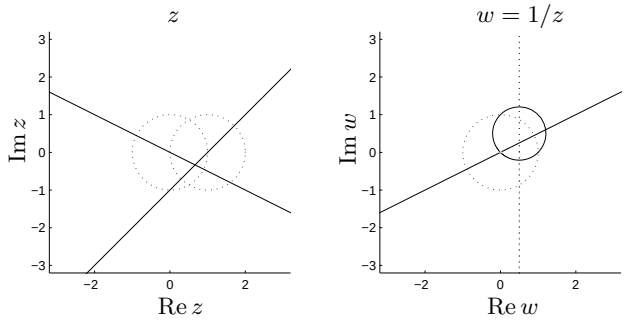


FIGURE 251.8. Further illustration of the map $f(z) = 1/z$. Note that the unit circle is mapped onto itself, while a circle through the origin is mapped onto a straight line. Note also that the straight line $y = ax$ with $a \neq 0$ passing through the origin is mapped onto its conjugate line $y = -ax$, while other lines are mapped onto circles.

EXAMPLE 251.7. (**Disc onto disc**) The function

$$w = f(z) = e^{i\alpha} \frac{z - z_0}{1 - \bar{z}_0 z}$$

where $\alpha \in \mathbb{R}$ and $z_0 \in \mathbb{C}$ with $|z_0| < 1$, maps the closed unit disc $\{|z| \leq 1\}$ onto the closed unit disc $\{|w| \leq 1\}$ in a one-to-one fashion with $f(z_0) = 0$. For the verification it suffices to verify that the circle $\{|z| = 1\}$ is mapped onto the circle $\{|w| = 1\}$.

EXAMPLE 251.8. (**Half-plane onto unit disc**) The function

$$w = f(z) = e^{i\alpha} \frac{z - z_0}{z - \bar{z}_0}$$

where $\alpha \in \mathbb{R}$ and $\operatorname{Im} z_0 > 0$, maps the upper half-plane $\{\operatorname{Im} z > 0\}$ onto the open unit disc $\{|w| < 1\}$ with $f(z_0) = 0$.

251.19 $w = z^{1/2}$, $w = e^z$, $w = \log(z)$ and $w = \sin(z)$

We describe in a couple of examples basic aspects of the mapping properties of some elementary functions.

EXAMPLE 251.9. The function

$$w = f(z) = z^{1/2} = \sqrt{|z|}e^{\frac{i}{2}\text{Arg z}},$$

maps the wedge $\{0 \le \arg z < \theta\}$ where $0 \le \theta < 2\pi$ onto the wedge $\{0 \le \arg w < \frac{\theta}{2}\}$.

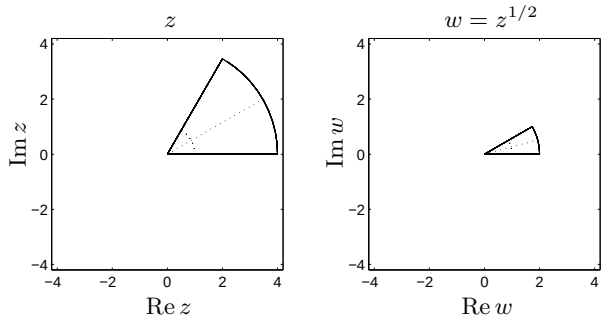


FIGURE 251.9. Illustration of the map $f(z) = z^{1/2}$.

EXAMPLE 251.10. The function $w = e^z$ maps the strip $\{z = x + iy : x \in \mathbb{R}, 0 \le y < 2\pi\}$ onto the complex plane C minus the origin. The line $\{x+iy : x \in \mathbb{R}\}$ with y fixed is mapped onto the halfline $\{(r, \theta) : r > 0\}$ with $\theta = y$ using polar coordinates.

EXAMPLE 251.11. The function $w = \text{Log}(z)$ maps $\mathbb{C}$ minus the origin onto the strip $\{w \in \mathbb{C} : 0 \le \text{Im}(w) < 2\pi\}$.

EXAMPLE 251.12. The function

$$\begin{aligned} w = f(z) = \sin(z) &= \frac{1}{2i}(e^{i(x+iy)} - e^{-i(x+iy)}) \\ &= \sin(x)\cosh(y) + i\,\cos(x)\sinh(y) = u(x,y) + iv(x,y), \end{aligned}$$

maps the strip $\{z = x + iy : \frac{-\pi}{2} < x < \frac{\pi}{2},\ y \in \mathbb{R}\}$ onto $\{w = u + iv : v \ne 0 \text{ if } |u| > 1\}$, which is the whole plane minus the two half-lines $\{u + iv : |u| > 1, v = 0\}$. The level curves of u and v are hyperbolas and ellipses, respectively. See Fig. 251.10.

251.20 Complex Integrals: First Shot

We make a direct extension of the integral of a differentiable function $F : \mathbb{R} \to \mathbb{R}$ to the integral of an analytic function $F : \mathbb{C} \to \mathbb{C}$, paralleling closely the presentation in Chapter *The Integral*.

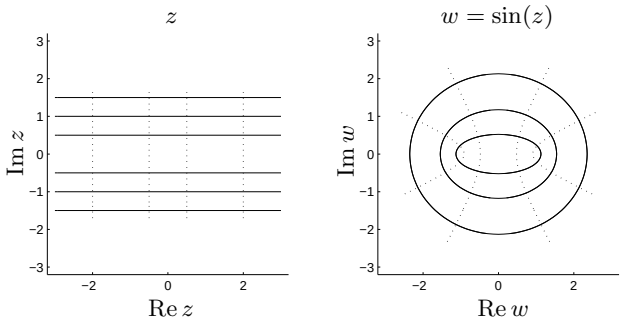


FIGURE 251.10. Illustration of the map $f(z) = \sin(z)$.

Let $F(z)$ be analytic in the domain Ω of the complex plane, with Lipschitz continuous derivative $f(z) = F'(z)$. Let Γ be a differentiable curve in Ω parameterized by $\gamma : [a, b] \to \mathbb{C}$, connecting the point $z_a = \gamma(a)$ with the point $z_b = \gamma(b)$, and let $z_a = z_0, z_1, ..., z_n = z_b$ be a sequence of points on Γ connecting z_a and z_b , see Fig. 251.11. We assume that $z_k \neq z_{k-1}$ for $k = 1, ..., n$.

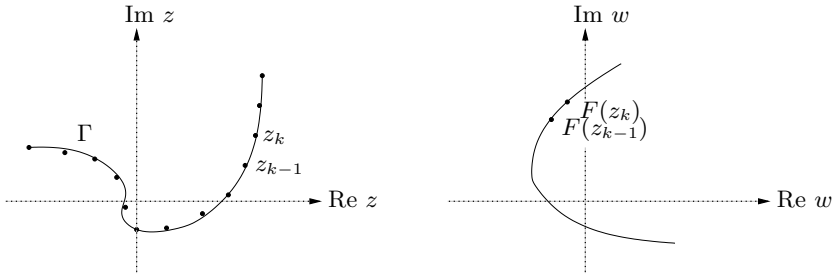


FIGURE 251.11. A curve Γ with sample points z_k and corresponding function values $F(z_k)$.

We can write

$$F(z_b) - F(z_a) = \sum_{k=1}^{n} (F(z_k) - F(z_{k-1})) = \sum_{k=1}^{n} \frac{F(z_k) - F(z_{k-1})}{z_k - z_{k-1}} (z_k - z_{k-1}). \tag{251.12}$$

Letting $\max_{k=1,...,n} |z_k - z_{k-1}|$ tend to zero, we are led to write

$$F(z_b) - F(z_a) = \int_\Gamma f(z)\, dz, \tag{251.13}$$

where we replace $\frac{F(z_k)-F(z_{k-1})}{z_k - z_{k-1}}$ by the derivative $F'(z_{k-1}) = f(z_{k-1})$ and $z_k - z_{k-1}$ by dz.

We note that the integral $\int_\Gamma f(z)\,dz$, being equal to $F(z_b) - F(z_a)$, is thus independent of the choice of the curve Γ connecting z_a and z_b. As a special case we note that if Γ is closed, corresponding to choosing $z_b = z_a$, then

$$\int_\Gamma f(z)\,dz = 0. \tag{251.14}$$

Recalling that $f(z)$ is analytic if $F(z)$ is analytic, we have found a reason to believe in *Cauchy's theorem* stating that the integral of an analytic function $f : \Omega \to \mathbb{C}$ around a simple closed curve in Ω enclosing a region contained in Ω, is zero. This is a corner-stone of the theory of analytic functions. Below we give a proof of Cauchy's theorem using a Green's formula.

251.21 Complex Integrals: General Case

Let Ω be an open domain in the complex plane and let Γ be a differentiable curve in $\mathbb{C}$ parameterized by $\gamma = (x, y) : [a, b] \to \mathbb{C}$. Let $f = u + iv : \Gamma \to \mathbb{C}$ be Lipschitz continuous and define

$$\int_\Gamma f(z)\,dz = \int_a^b \big(u(x(t), y(t)) + iv(x(t), y(t))\big)(\dot{x}(t) + i\dot{y}(t))\,dt, \tag{251.15}$$

where thus formally $dz = dx + idy = \dot{x}dt + i\dot{y}dt = (\dot{x} + i\dot{y})\,dt$. The integral is defined if $u(x, y)$ and $v(x, y)$ are Lipschitz continuous in (x, y) and $\dot{x}(t)$ and $\dot{y}(t)$ are Lipschitz continuous in t. As in the Chapter Curve Integrals, we see that the integral is independent of the parametrization.

We can express the integral as a limit of Riemann sums in the usual way:

$$\int_\Gamma f(z)\,dz = \lim_{n\to\infty} \sum_{k=1}^{n} f(z_{k-1})(z_k - z_{k-1}), \tag{251.16}$$

where $z_a = z_0, z_1, ..., z_n = z_b$ is a sequence of points along Γ with $\max_{k=1,...,n} |z_k - z_{k-1}|$ tending to zero as n tends to infinity.

Below we use also ζ as a complex variable and thus write in particular

$$\int_\Gamma f(z)\,dz = \int_\Gamma f(\zeta)\,d\zeta.$$

EXAMPLE 251.13. Let Γ be a circle around the origin of radius one oriented counter-clockwise and parameterized by $\gamma(t) = e^{it} = \cos(t) + i\sin(t) = (\cos(t), \sin(t))$ with $0 \le t < 2\pi$. Let $f(z) = z^n$ with n an

integer. We have for $n \neq -1$, since $dz = (-\sin(t) + i\cos(t))\,dt = ie^{it}\,dt$,

$$\begin{aligned}\int_\Gamma f(z)\,dz = \int_\Gamma z^n\,dz &= \int_0^{2\pi} e^{int}(-\sin(t) + i\cos(t))\,dt \\ &= i\int_0^{2\pi} e^{int}e^{it}\,dt = i\int_0^{2\pi} e^{i(n+1)t}\,dt \\ &= \frac{i}{n+1}[\sin((n+1)t) - i\cos((n+1)t)]_0^{2\pi} = 0.\end{aligned}$$

This conforms with Cauchy's theorem for $n = 0, 1, 2, ...$, since then $f(z)$ is analytic in $\mathbb{C}$. For $n = -1$ with $f(z) = \frac{1}{z}$, we get

$$\int_\Gamma \frac{1}{z}\,dz = \int_\Gamma \frac{dz}{z} = \int_0^{2\pi} \frac{ie^{it}}{e^{it}}\,dt = 2\pi i. \tag{251.17}$$

Note the counter-clockwise orientation of Γ. The function $f(z) = \frac{1}{z}$ is not analytic in the domain enclosed by Γ, since $\frac{1}{z}$ is not differentiable for $z = 0$, and thus the integral $\int_\Gamma \frac{dz}{z}$ may be non-zero. We shall see below that the derivative of $\log(z)$ is equal to $\frac{1}{z}$, but $\log(z)$ is not uniquely defined for $z \neq 0$, and thus $\int_\Gamma \frac{dz}{z}$ may be non-zero.

The functions $f(z) = z^n$ with $n = -2, -3, ..$ are all derivatives of analytic functions and thus $\int_\Gamma f(z)\,dz = 0$ if Γ is a closed curve which does not pass through 0.

251.22 Basic Properties of the Complex Integral

The complex integral has properties analogous to those of the usual real integral such as linearity, additivity over subintervals and integration by parts. For example, we have if $|f(z)| \leq M$ for $z \in \Gamma$:

$$|\int_\Gamma f(z)\,dz| \leq M\int_\Gamma ds = ML(\Gamma), \tag{251.18}$$

where $L(\Gamma)$ is the length of Γ:

$$L(\Gamma) = \int_a^b (\dot{x}^2(t) + \dot{y}^2(t))^{1/2}\,dt.$$

This follows by taking absolute values in (251.16) and then passing to the limit:

$$|\int_\Gamma f(z)\,dz| \leq \int_\Gamma |f(z)|\,|dz| = \int_\Gamma |f(z)|\,ds \leq ML(\Gamma),$$

where formally $|dz| = ds$, and thus the estimate may be viewed as a generalized triangle inequality.

251.23 Taylor's Formula: First Shot

If $f : \Omega \to \mathbb{C}$ is analytic and Γ is a straight line in Ω connecting z_0 and z, then we can write

$$f(z) = f(z_0) + \int_\Gamma f'(\zeta)\,d\zeta = f(z_0) + \int_\Gamma f'(\zeta)\frac{d}{d\zeta}(\zeta - z_0)\,d\zeta,$$

and thus by partial integration (the usual rules hold)

$$f(z) = f(z_0) + f'(z_0)(z - z_0) - \int_\Gamma f''(\zeta)(\zeta - z_0)\,d\zeta.$$

Continuing, writing $(\zeta - z_0) = \frac{1}{2}\frac{d}{d\zeta}(\zeta - z_0)^2$, we get

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + \frac{f''(z_0)}{2}(z - z_0)^2 + \int_\Gamma f^{(3)}(\zeta)\frac{(\zeta - z_0)^2}{2}\,d\zeta.$$

We conclude that for z in a neighborhood of z_0, we have

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + \frac{f''(z_0)}{2}(z - z_0)^2 + E_f(z, z_0), \quad (251.19)$$

where

$$|E_f(z, z_0)| \le K\int_\Gamma \frac{|\zeta - z_0|^2}{2}\,|d\zeta| = \frac{K}{6}|z - z_0|^3,$$

and we assume that $|f^{(3)}(\zeta)| \le K$ for $\zeta \in \Gamma$. More generally, we have the following Taylor's formula:

Theorem 251.4 *If $f : \Omega \to \mathbb{C}$ is analytic in Ω with $|f^{(n+1)}(z)| \le K$ for $z \in \Omega$, then we have for $z, z_0 \in \Omega$ (with the straight line connecting z and z_0 contained in Ω):*

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + ... + \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n + R_n(z, z_0), \quad (251.20)$$

where $|R_n(z, z_0)| \le \frac{K}{(n+1)!}|z - z_0|^{n+1}$.

251.24 Cauchy's Theorem

We shall now prove that if $f(z)$ is analytic in Ω and Γ is a simple closed curve in Ω enclosing a domain Ω_Γ contained in Ω, then

$$\int_\Gamma f(z)\,dz = 0.$$

To see this we write

$$\int_\Gamma f(z)\,dz = \int_a^b (u(x(t),v(t)) + iv(x(t),y(t)))(\dot{x}(t) + i\dot{y}(t))\,dt,$$

where $\gamma(t) = (x(t), y(t))$ with $a \le t \le b$ a parametrization of Γ. Taking the real part, we get

$$\begin{aligned}\mathrm{Re}(\int_C f(z)\,dz) &= \int_a^b (u(x(t),y(t))\dot{x}(t) - v(x(t),y(t))\dot{y}(t)\,dt \\ &= \int_a^b (u(x,y), -v(x,y)) \cdot (\dot{x},\dot{y}))\,dt.\end{aligned}$$

By the Cauchy-Riemann equations, we have

$$\nabla \times (u,-v) = \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} = 0 \quad \text{in } \Omega_\Gamma,$$

which proves, recalling Stokes' theorem (57.13), that

$$\begin{aligned}\int_a^b (u(x,y), -v(x,y)) \cdot (\dot{x},\dot{y})\,dt &= \int_\Gamma (u(x,y), -v(x,y)) \cdot ds \\ &= \int_{\Omega_\Gamma} \nabla \times (u,-v) dxdy = 0.\end{aligned}$$

We conclude that $\mathrm{Re}(\int_\Gamma f(z)\,dz) = 0$, and similarly we see that $\mathrm{Im}(\int_\Gamma f(z)\,dz) = 0$ and we have thus proved *Cauchy's theorem*:

Theorem 251.5 (Cauchy's theorem) *If $f(z)$ is analytic in Ω and Γ is a simple closed curve in Ω enclosing a domain contained in Ω, then*

$$\int_\Gamma f(z)\,dz = 0.$$

Note that Γ is not allowed to enclose "holes" of Ω where $f(z)$ is not analytic. For example, we saw above that $\int_\Gamma \frac{1}{z}\,dz = 2\pi i \neq 0$, where Γ is a circle around the origin. This is because Γ encloses the point $z = 0$ where $\frac{1}{z}$ is not analytic.

251.25 Cauchy's Representation Formula

We prove that if $f(z)$ is analytic in an open domain Ω, and Γ is a simple closed curve in Ω oriented counter-clockwise and bounding the open domain Ω_Γ contained in Ω, then for $z_0 \in \Omega_\Gamma$,

$$f(z_0) = \frac{1}{2\pi i}\int_\Gamma \frac{f(z)}{z - z_0}\,dz, \tag{251.21}$$

which is *Cauchy's representation formula*. Note the counter-clockwise orientation. Further, note that z_0 is not allowed to lie on the curve Γ; we assume that z_0 lies *inside* Γ. Cauchy's formula (251.21) shows that the values of $f(z)$ on Γ alone, determine the values of $f(z)$ in all of Ω_Γ. This shows that an analytic function is not allowed to bring surprises: if we know $f(z)$ on Γ, then we know $f(z)$ in the whole domain Ω_Γ bounded by Γ. The proof follows from realizing that the function

$$g(z) = \frac{f(z) - f(z_0)}{z - z_0} \quad \text{for } z \neq z_0, \quad g(z_0) = f'(z_0),$$

is analytic in Ω, because $g(z)$ is clearly differentiable if $z \neq z_0$ and using a Taylor expansion of $f(z)$, it follows that $g(z)$ is differentiable also at $z = z_0$ with derivative $g'(z_0) = \frac{f''(z_0)}{2}$. Indeed, recalling (251.19) we have

$$g(z) - g(z_0) = \frac{f(z) - f(z_0) - f'(z_0)(z - z_0)}{(z - z_0)} = \frac{f''(z_0)}{2}(z - z_0) + E_f(z, z_0)$$

with $|E_f(z, z_0)| \leq \frac{K}{6}|z - z_0|^2$ and K a bound for $|f^{(3)}(z)|$, which proves the desired result. We conclude that

$$\int_\Gamma \frac{f(z) - f(z_0)}{z - z_0} dz = 0,$$

and using that

$$\int_\Gamma \frac{f(z_0)}{z - z_0} dz = f(z_0) \int_\Gamma \frac{1}{z - z_0} dz = 2\pi i\, f(z_0),$$

we obtain the desired result (251.21). We summarize:

Theorem 251.6 (Cauchy's representation formula) *If $f(z)$ is analytic in an open domain Ω, and Γ is a simple closed curve in Ω oriented counter-clockwise and enclosing the open domain Ω_Γ contained in Ω, then for $z_0 \in \Omega_\Gamma$,*

$$f(z_0) = \frac{1}{2\pi i} \int_\Gamma \frac{f(z)}{z - z_0}\, dz. \tag{251.22}$$

Differentiating with respect to z_0 we obtain the following generalized representation formula:

Theorem 251.7 (Cauchy's generalized representation formula) *If $f(z)$ is analytic in an open domain Ω, and Γ is a simple closed curve in Ω oriented counter-clockwise and enclosing an open domain Ω_Γ in Ω, then for $z_0 \in \Omega_\Gamma$ and $n = 0, 1, 2, ...,$*

$$f^{(n)}(z_0) = \frac{n!}{2\pi i} \int_\Gamma \frac{f(z)}{(z - z_0)^{n+1}}\, dz. \tag{251.23}$$

We note that if z_0 lies outside the region bounded by Γ, then

$$\frac{1}{2\pi i}\int_\Gamma \frac{f(z)}{z-z_0}\,dz = 0,$$

simply because $\int_\Gamma \frac{1}{z-z_0}\,dz = 0$ in this case as a consequence of the fact that $\frac{1}{z-z_0}$ is analytic in a domain containing Γ. Choosing $z_0 \in \Gamma$ leads to a divergent integral because of the singularity of the factor $\frac{1}{z-z_0}$, and to define a proper value of the integral in this case leads to the so called *Cauchy principal value*, which we discuss below.

251.26 Taylor's Formula: Second Shot

By using Cauchy's formula we now give another version of Taylor's formula for a function $f(z)$ which is analytic in a neighborhood Ω of a point $z_0 \in \mathbb{C}$. We start writing Cauchy's formula in the form

$$f(z) = \frac{1}{2\pi i}\int_\Gamma \frac{f(\zeta)}{\zeta - z}\,d\zeta, \tag{251.24}$$

where for definiteness we choose Γ to be a counter-clockwise oriented circle around z_0 of radius r contained in Ω. Using the identity

$$\frac{1}{1-q} = 1 + q + q^2 + ... + q^n + \frac{q^{n+1}}{1-q},$$

where $q \in \mathbb{C}$ satisfies $|q| < 1$, setting $q = \frac{z-z_0}{\zeta - z_0}$ with $z \in \Omega$ and $\zeta \in \Gamma$, we can write

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - z_0}\big[1 + \frac{z-z_0}{\zeta - z_0} + ... + (\frac{z-z_0}{\zeta - z_0})^n\big] + \frac{1}{\zeta - z}(\frac{z-z_0}{\zeta - z_0})^{n+1},$$

where we used that

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - z_0 - (z - z_0)} = \frac{1}{\zeta - z_0}\ \frac{1}{1-q}.$$

Insertion into (251.24) now gives

$$f(z) = \frac{1}{2\pi i}\int_\Gamma \frac{f(\zeta)}{\zeta - z_0}\,d\zeta + \frac{z-z_0}{2\pi i}\int_\Gamma \frac{f(\zeta)}{(\zeta - z_0)^2}\,d\zeta + ...$$
$$+\frac{(z-z_0)^n}{2\pi i}\int_\Gamma \frac{f(\zeta)}{(\zeta - z_0)^{n+1}}\,d\zeta + R_n(z),$$

where

$$R_n(z) = \frac{(z-z_0)^{n+1}}{2\pi i}\int_\Gamma \frac{f(\zeta)}{(\zeta - z_0)^{n+1}(\zeta - z)}\,d\zeta. \tag{251.25}$$

Using Cauchy's representation formulas we thus obtain the following Taylor formula:

Theorem 251.8 *If $f(z)$ is analytic in a neigborhood Ω of a $z_0 \in \mathbb{C}$, then*

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + ... + \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n + R_n(z), \quad (251.26)$$

where the remainder $R_n(z)$ is given by (251.25) with Γ a circle around z_0.

If $\lim_{n\to\infty} R_n(z) = 0$ for z in a neighborhood Ω of z_0, then we obtain the following *power series representation* of $f(z)$ for $z \in \Omega$:

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n. \quad (251.27)$$

We conclude by proving that indeed $\lim_{n\to\infty} R_n(z) = 0$ for z in a neighborhood of z_0. We then assume that the disc $D_r(z_0) = \{z \in \mathbb{C} : |z - z_0| \leq r\}$ is contained in the domain Ω of analyticity of $f(z)$ and we assume that $|f(z)| \leq M$ for $z \in D_r(z_0)$. Assuming that $|z - z_0| < \frac{r}{2}$, we obtain by inserting absolute values in (251.25) using that $|\zeta - z| \geq \frac{r}{2}$ and $L(\Gamma) = 2\pi r$:

$$|R_n(z)| \leq (\frac{|z - z_0|}{r})^{n+1} 2M,$$

which proves that $\lim_{n\to\infty} R_n(z) = 0$ for $|z - z_0| \leq \frac{r}{2}$. We can extend the argument to z satisfying $|z - z_0| < r$, and we summarize as follows:

Theorem 251.9 (Taylor's formula) *If $f : \Omega \to \mathbb{C}$ is analytic and $D_r(z_0) = \{z \in \mathbb{C} : |z - z_0| \leq r\}$ is contained in Ω and f is bounded on $D_r(z_0)$, then $f(z)$ can be represented as the convergent power series (251.27) for $|z - z_0| < r$.*

Power series representations of analytic functions of the form (251.27) play an important role and we devote the next section to this topic starting with the case $z_0 = 0$.

251.27 Power Series Representation of Analytic Functions

Consider a series of the form

$$\sum_{m=0}^{\infty} a_m z^m \quad (251.28)$$

where the coefficients $a_m \in \mathbb{C}$ and we assume $z \in \mathbb{C}$. The concepts of convergence and absolute convergence for (251.28) are direct analogs of the corresponding concepts for series with a_m and z being real, see Chapter *Series*. In particular we say that $\sum_{m=0}^{\infty} a_m z^m$ is absolutely convergent if

$\sum_{m=0}^{\infty} |a_m z^m|$ is convergent, and note that an absolutely convergent series is convergent.

Each term of the series (251.28) is analytic in $\mathbb{C}$ and each partial sum

$$\sum_{m=0}^{n} a_m z^m$$

is thus analytic in $\mathbb{C}$. Suppose now that the series (251.28) is convergent for a particular $z = \hat{z}$ with $|\hat{z}| = r$. Since the terms b_m of a convergent series $\sum_{m=0}^{\infty} b_m$ must tend to zero, there is a constant C such that

$$|a_n \hat{z}^n| = |a_n| r^n \leq C \quad n = 0, 1, 2, ...$$

Suppose now that $|z| < r$. We then have

$$\sum_{n=0}^{\infty} |a_n z^n| = \sum_{n} |a_n r^n| |\frac{z}{r}|^n \leq C \sum_{n=0}^{\infty} |\frac{z}{r}|^n < \infty,$$

because $|\frac{z}{r}| < 1$. This proves that $\sum_{n=0}^{\infty} a_n z^n$ is absolutely convergent for $|z| < r$ and is thus convergent for $|z| < r$.

We say that the *radius of convergence* of $\sum_{n=0}^{\infty} a_n z^n$ is equal to r, if $\sum_{n=0}^{\infty} a_n z^n$ is convergent for $|z| < r$ but not convergent for some z with $|z| \geq r$.

One can (easily) show that inside its radius of convergence r a series $\sum_{n=0}^{\infty} a_n z^n$ is differentiable with

$$(\sum_{n=0}^{\infty} a_n z^n)' = \sum_{n=1}^{\infty} n a_n z^{n-1},$$

where the termwise differentiated series $\sum_{n=1}^{\infty} n a_n z^{n-1}$ is also convergent for $|z| < r$.

More generally, we consider power series of the form

$$\sum_{m=0}^{\infty} a_m (z - z_0)^m, \tag{251.29}$$

where we made a shift of variable from z to $z - z_0$ with $z_0 \in \mathbb{C}$ given. The notion of convergence and radius of convergence extend in the obvious way. Of course, (251.29) connects to the Taylor series of $f(z)$ at z_0 with $a_m = \frac{f^{(m)}(z_0)}{m!}$.

EXAMPLE 251.14. The series

$$\sum_{n=0}^{\infty} \frac{z^n}{n!}$$

is convergent for any fixed $z \in \mathbb{C}$, since $n! = 1 \cdot 2 \cdot 3 \cdots n$ grows much quicker than r^n for any $r > 0$. We can thus differentiate termwise and we get

$$(\sum_{n=0}^{\infty} \frac{z^n}{n!})' = \sum_{n=1}^{\infty} \frac{z^{n-1}}{(n-1)!} = \sum_{n=0}^{\infty} \frac{z^n}{n!},$$

which shows that $\sum_{n=0}^{\infty} \frac{z^n}{n!}$ satisfies the differential equation $u'(z) = u(z)$ with the "initial" condition $u(0) = 1$. It follows that

$$\exp(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}. \tag{251.30}$$

Alternatively, this follows by noting that this is the Taylor series representation of $f(z) = \exp(z)$ around $z_0 = 0$, noting that $f^{(n)}(z) = \exp(z)$ for $n = 1, 2, ..., .$

Using that $\cos(z) = \frac{1}{2}(\exp(iz) + \exp(-iz))$ and $\sin(z) = \frac{1}{2i}(\exp(iz) - \exp(-iz))$, we obtain the following Taylor series representations valid for $z \in \mathbb{C}$:

$$\cos(z) = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!}, \qquad \sin(z) = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!}.$$

Example 251.15. Another basic example is given by

$$\log(1+z) = \sum_{n=1} \frac{(-1)^{n-1}}{n} z^n \quad \text{for } |z| < 1,$$

which is readily obtained differentiating $\log(1+z)$.

251.28 Laurent Series

Consider a series of the form

$$\sum_{m=1}^{\infty} b_m z^{-m}, \tag{251.31}$$

obtained by replacing z by $\frac{1}{z}$ in a power series $\sum_{m=1}^{\infty} b_m z^m$ with radius of convergence r. The series (251.31) will thus converge for $|z| > r$. More generally we may consider a *Laurent series* of the form

$$f(z) = \sum_{m=0}^{\infty} a_m z^m + \sum_{m=1}^{\infty} b_m z^{-m}, \tag{251.32}$$

which we assume to be convergent in an annulus $\{r_1 < |z| < r_2\}$. The function $f(z)$ defined by (251.32) is analytic in the annulus $\{r_1 < |z| < r_2\}$. Conversely, if $f(z)$ is analytic in the annulus $\{r_1 < |z| < r_2\}$, then $f(z)$ admits the Laurent series expansion (251.32) with the coefficients a_m and b_m being given by

$$a_m = \frac{1}{2\pi i}\int_\Gamma \frac{f(\zeta)}{\zeta^{m+1}}\,d\zeta, \quad b_m = \frac{1}{2\pi i}\int_\Gamma f(\zeta)\zeta^{m-1}\,d\zeta, \tag{251.33}$$

where Γ is a simple closed counter-clockwise oriented curve in the annulus encircling the origin. The formula for the coefficients is obtained by multiplying by a proper power of z and integrating around Γ.

We may generalize to shifts of the origin to a given point z_0 replacing z by $z - z_0$.

Example 251.16. We have

$$\frac{1}{1-z} = \sum_{m=0}^{\infty} z^m \quad \text{for } |z| < 1,$$
$$\frac{1}{1-z} = \frac{-1}{z(1-z^{-1})} = \sum_{m=1}^{\infty} z^{-m} \quad \text{for } |z| > 1.$$

251.29 Residue Calculus: Simple Poles

Let $f(z)$ be analytic in a simply connected open domain Ω, except at an isolated point $z_0 \in \Omega$, and let Γ be a simple closed curve in Ω oriented counter-clockwise with z_0 contained in the open domain Ω_Γ bounded by Γ. We say that the simple closed curve Γ *surrounds* z_0 counter clockwise. In general the integral

$$\int_\Gamma f(z)\,dz$$

will then not be zero, but the integral will have the same value for any such simple closed curve Γ surrounding z_0 clockwise. To see this we consider two such curves Γ_1 and Γ_2 and introduce the two coinciding curves $\Gamma_3^{\pm}$ with opposite orientation joining Γ_1 and Γ_2 according to Fig. 251.12, and by joining the curves Γ_1, Γ_3^+, $-\Gamma_2$ (Γ_2 backwards) and Γ_3^- we obtain a single closed curve enclosing a domain where $f(z)$ is analytic (that is, not containing z_0 in its interior) over which the integral of $f(z)$ vanishes because of Cauchy's theorem. Thus, noting that the integrals over Γ_3^+ and Γ_3^- cancel, we have

$$0 = \int_{\Gamma_1} f(z)\,dz + \int_{-\Gamma_2} f(z)\,dz = \int_{\Gamma_1} f(z)\,dz - \int_{\Gamma_2} f(z)\,dz$$

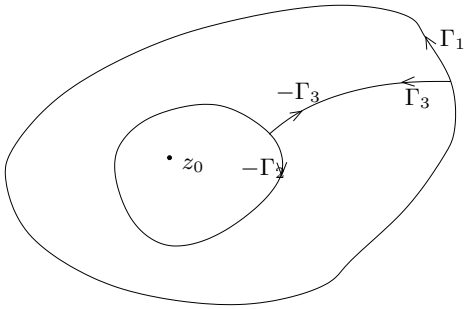


FIGURE 251.12. Two simple curves Γ_1 and $-\Gamma_2$ (Γ_2 backwards), surrounding z_0, adjoined by curves $\Gamma_3^{\pm}$ to form one simply connected curve *not* surrounding z_0.

where we used that the orientation of $-\Gamma_2$ and Γ_2 are reversed. It follows that the integral over Γ_1 is equal to the integral over Γ_2.

Suppose now that $f(z)$ has the form

$$f(z) = \frac{g(z)}{z - z_0},$$

where $g(z)$ is analytic in Ω and $z_0 \in \Omega$. We then say that $f(z)$ has a *simple pole* at $z = z_0$. We have by Cauchy's representation formula with Γ a simple closed curve surrounding z_0 counter-clockwise,

$$\int_\Gamma f(z)\,dz = \int_\Gamma \frac{g(z)}{z - z_0}\,dz = 2\pi i g(z_0).$$

The value $g(z_0)$ is called the *residue* of $f(z)$ at z_0, which we denote by $\operatorname{Res} f(z_0)$, and thus

$$\operatorname{Res} f(z_0) = g(z_0) = \lim_{z \to z_0} (z - z_0) f(z).$$

EXAMPLE 251.17. Let $f(z) = \frac{z}{z-1}$ and let Γ be the circle $\gamma(t) = (\cos(t)-1, \sin(t))$ with $0 \le t \le 2\pi$, surrounding $(1,0)$ counter-clockwise. By the Residue Theorem, we have since obviously $\operatorname{Res} f(1) = 1$

$$\int_\Gamma \frac{z}{z-1}\,dz = 2\pi i.$$

EXAMPLE 251.18. To evaluate $\int_\Gamma f(z)\,dz$, where $f(z) = \frac{1}{e^z - 1}$ and Γ is a circle centered at the origin and oriented counter-clockwise, we note that

$$f(z) = \frac{z}{e^z - 1}\frac{1}{z} = \frac{g(z)}{z}$$

with

$$\frac{1}{g(z)} = \frac{e^z - 1}{z} = h(z).$$

Since $\lim_{z\to 0} h(z) = 1$, we have Res $f(0) = g(0) = 1$, and thus $\int_\Gamma f(z)\,dz = 2\pi i$.

251.30 Residue Calculus: Poles of any Order

Suppose now $f(z)$ has a (multiple) *pole of order* $n = 2, 3, ...$, at z_0, that is, $f(z)$ is of the form

$$f(z) = \frac{g(z)}{(z - z_0)^n},$$

with $g(z)$ analytic in a neighborhood of z_0. By Cauchy's generalized representation formula we have if Γ is a simple closed curve surrounding z_0 counter-clockwise:

$$\int_\Gamma f(z)\,dz = \int_\Gamma \frac{g(z)}{(z - z_0)^n}\,dz = \frac{2\pi i}{(n-1)!}\,g^{(n-1)}(z_0).$$

We now extend the definition of the residue $\operatorname{Res} f(z_0)$ to a pole of order $n = 1, 2, ...$, by setting

$$\operatorname{Res} f(z_0) = \frac{g^{(n-1)}(z_0)}{(n-1)!},$$

and thus we have again

$$\int_\Gamma f(z)\,dz = 2\pi i \operatorname{Res} f(z_0).$$

Example 251.19. The function

$$f(z) = \frac{1}{(z-1)^2(z-3)}$$

has a pole of order 2 at $z = 1$ and order 1 at $z = 3$. We compute $\operatorname{Res} f(3) = \frac{1}{4}$, and further $\operatorname{Res} f(1) = -\frac{1}{4}$ since $\frac{d}{dz}\frac{1}{z-3} = -\frac{1}{(z-3)^2} = -\frac{1}{4}$ if $z = 1$.

251.31 The Residue Theorem

We now prove the following basic result of residue calculus:

Theorem 251.10 (The Residue Theorem) *Let $f(z)$ be analytic in a simply connected open domain Ω, except at finitely many isolated points z_1, z_2,...,z_n in Ω where $f(z)$ has simple or multiple poles, and let Γ be a simple closed curve in Ω surrounding all the z_m counter-clockwise. Then*

$$\int_\Gamma f(z)\,dz = \sum_{m=1}^{n} 2\pi i Res\, f(z_m).$$

The result follows by surrounding each of the z_m with a little circle $= \Gamma_m$ inside Γ oriented counter-clockwise. By Cauchy's theorem we then have

$$\int_\Gamma f(z)\,dz + \sum_{m=1}^{n} \int_{-\Gamma_m} f(z)\,dz = 0,$$

arguing as in the Section on Residue Calculus: simple poles, from which follows that

$$\int_\Gamma f(z)\,dz = \sum_{m=1}^{n} \int_{\Gamma_m} f(z)\,dz = 2\pi i \sum_{m=1}^{n} \mathrm{Res}\, f(z_m),$$

which proves the desired result.

Example 251.20. We compute

$$I = \int_\Gamma \frac{4-3z}{z^2-z}\,dz = \int_\Gamma \frac{4-3z}{z(z-1)}\,dz$$

where Γ is a simple closed curve surrounding counter-clockwise the two simple poles $z = 1$ and $z = 0$ of $\frac{4-3z}{z^2-z}$ and get $I = 2\pi i(-4+1) = -6\pi i$.

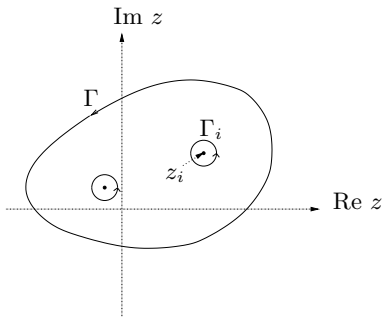


FIGURE 251.13. A curve Γ and curves Γ_i, surrounding poles of $f(z)$.

251.32 Computation of $\int_0^{2\pi} R(\sin(t),\cos(t))\,dt$

Consider an integral of the form

$$\int_0^{2\pi} R(\sin(t),\cos(t))\,dt$$

where $R(x,y)$ is a rational function of $x,y\in\mathbb{R}$. By the substitution

$$\begin{aligned} z=e^{it},\quad dz=ie^{it}dt=iz\,dt\\ \cos(t)=\frac{e^{it}+e^{-it}}{2}=\frac{1}{2}(z+\frac{1}{z})\\ i\sin(t)=\frac{e^{it}-e^{-it}}{2}=\frac{1}{2}(z-\frac{1}{z}), \end{aligned}$$

the integral is converted into

$$\int_{|z|=1} R(\frac{z^2-1}{2iz},\frac{z^2+1}{2z})\frac{dz}{iz},$$

which can be evaluated using residue calculus, provided the integrand has no poles on $|z|=1$.

Example 251.21. We compute

$$I=\int_0^{2\pi}\frac{dt}{a+\cos(t)},$$

where $a>1$ is a constant. Using the transformation just indicated we get

$$I=-2i\int_{|z|=1}\frac{dz}{z^2+2az+1}=-2i\int_{|z|=1}\frac{dz}{(z-\alpha)(z-\beta)},$$

where $\alpha=-a+\sqrt{a^2-1}$ and $\beta=-a-\sqrt{a^2-1}$. Since $|\alpha|<1$ and $|\beta|>1$, the residue at α is $\frac{1}{\alpha-\beta}$ and thus $I=\frac{2\pi}{\sqrt{a^2-1}}$.

251.33 Computation of $\int_{-\infty}^{\infty}\frac{p(x)}{q(x)}\,dx$

Integrals of the form

$$\int_{-\infty}^{\infty} f(x)\,dx \tag{251.34}$$

can be evaluated using residue calculus under the assumption that the extended function $f(z)$ with $z\in\mathbb{C}$ has no poles on the real axis and that

$|f(z)| \leq M|z|^{-2}$ for $|z|$ large. We start out showing how to use residue calculus to compute the integral

$$I = \int_{-\infty}^{\infty} \frac{1}{1+x^2}\, dx,$$

(thus without using that $\arctan(x)$ is a primitive function of $\frac{1}{1+x^2}$). We write

$$I = \lim_{R\to\infty} \int_{-R}^{R} \frac{1}{1+x^2}\, dx = \lim_{R\to\infty} \int_{\Gamma_R} f(z)\, dz,$$

where $f(z) = \frac{1}{1+z^2}$ and Γ_R is the boundary of the semi-disc $|z| \leq R$ with $\text{Re } z = x \geq 0$. This follows from the fact that

$$\lim_{R\to\infty} \int_{\Gamma_R^+} f(z)\, dz = 0,$$

where Γ_R^+ is the upper part of the semi-circle with $\text{Re } z = x > 0$. By the Residue theorem we have

$$\int_{\Gamma_R} f(z)\, dz = 2\pi i \frac{1}{2i} = \pi$$

since the residue of $f(z) = \frac{1}{(z-i)(z+i)}$ inside Γ_R is equal to $\frac{1}{z+i}$ with $z = i$. We conclude that

$$\int_{-\infty}^{\infty} f(x)\, dx = \pi,$$

which of course conforms with the result obtained using that $\frac{d}{dx}\arctan(x) = \frac{1}{1+x^2}$.

The same technique can be used if $f(x) = \frac{p(x)}{q(x)}$ is a rational function with the degree of $q(x)$ two units (or more) higher than that of the numerator $p(x)$. The same technique can be used to evaluate the Fourier transform (cf below) of $\frac{p(x)}{q(x)}$:

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{p(x)e^{i\xi x}}{q(x)}\, dx.$$

251.34 Applications to Potential Theory in $\mathbb{R}^2$

There is a strong coupling between analytic functions and potential theory in $\mathbb{R}^2$, because if $f(z) = u(x,y) + iv(x,y)$ is analytic in Ω, then the real and imaginary parts $u(x,y)$ and $v(x,y)$ are harmonic in Ω, that is, $\Delta u = \Delta v = 0$ in Ω. Conversely, as we saw above, if $u(x,y)$ is harmonic in a simply connected domain Ω in $\mathbb{R}^2$, then there exists a *harmonic conjugate* function $v(x,y)$ uniquely determined up to a constant such that $f(z) =$

$u(x,y)+iv(x,y)$ is analytic in Ω, see Problem 251.10. The Cauchy-Riemann equations state that $\nabla u = \nabla \times v$:

$$\nabla u = (\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}) = \nabla \times v = (\frac{\partial v}{\partial y}, -\frac{\partial v}{\partial x}).$$

from which follows that ∇u and ∇v are orthogonal:

$$\nabla u \cdot \nabla v = \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} = \frac{\partial v}{\partial y}\frac{\partial v}{\partial x} - \frac{\partial v}{\partial x}\frac{\partial v}{\partial y} = 0.$$

We conclude that the level curves of $u(x,y)$ and its conjugate $v(x,y)$ are orthogonal. We note that level curves of $u(x,y)$ and $v(x,y)$ in the $z=(x,y)$-plane correspond to the level lines $u =$ constant and $v =$ constant of the analytic function $w = u + iv$ in the $w = (u,v)$-plane.

In fact, much of the interest in analytic functions comes from the connection to potential theory in $\mathbb{R}^2$. Today, computational methods capable of solving also problems in $\mathbb{R}^3$, have changed this picture and analytic functions now play a less important role in areas of applications such as fluid and solid mechanics.

Applications to fluid mechanics typically concern incompressible irrotational flow in 2d with $u(x,y)$ representing a velocity potential and $v(x,y)$ an associated so called *stream function*. The velocity U of the flow is then given by $U = \nabla u = \nabla \times v$

$$U = \nabla u = (\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}) = \nabla \times v = (\frac{\partial v}{\partial y}, -\frac{\partial v}{\partial x}).$$

We have $\nabla \cdot U = \Delta u = 0$ and $\nabla \times U = -\Delta v = 0$ and thus U is incompressible and irrotational. The level curves of $u(x,y)$ with normal ∇u correspond to equi-potential curves, and the level curves of v with normal $\nabla v = -\nabla \times u$ will correspond to the *streamlines* followed by a fluid particle moving with the velocity U. We conclude that each analytic function $f(z) = u(x,y) + iv(x,y)$ may be associated to a particular stationary incompressible and irrotational fluid flow, and the level curves of u and v form a mutually orthogonal set of curves with the level curves of v describing the streamlines of the flow.

In applications to electromagnetics, $u(x,y)$ represents an electric potential with ∇u an electric field, and the level curves of $v(x,y)$ represent the curves traced by electrically charged particles in the electric field.

In applications to heat flow $u(x,y)$ may represent temperature and the level curves for u thus become isolines for temperature and ∇u is proportional to the heat flow.

EXAMPLE 251.22. **(Flow in a corner)** The function $w = u+iv = z^2$ describes a certain flow in the quarter-plane $\{z = x+iy : x,y \geq 0\}$, with corresponding potential $u(x,y) = x^2 - y^2$ and stream-function $v(x,y) = 2xy$, see Fig. 251.14.

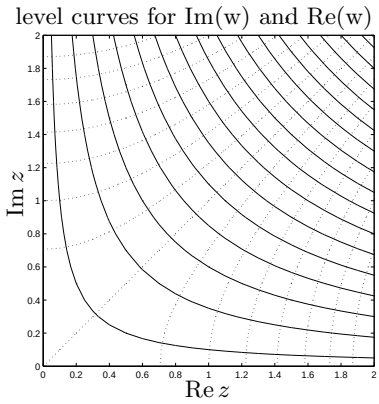


FIGURE 251.14. Level curves of $\mathrm{Im}(w) = 2xy$ (solid) and $\mathrm{Re}(w) = x^2 - y^2$ (dotted) for $w = z^2$.

The equi-potential lines $u(x, y) =$ constant and streamlines $v(x, y) =$ constant in the (x, y)-plane are the images of the lines $u =$ constant and $v =$ constant under the mapping $z = w^{1/2}$ from the halfplane $\{w = u + iv : v \geq 0\}$ onto the quarter-plane $\{z = x + iy : x, y \geq 0\}$.

EXAMPLE 251.23. **(The spinning tennis ball)** We consider two types of rotation-free incompressible flow around the unit disc $\{(x_1, x_2) : x_1^2 + x_2^2 < 1\}$ in two dimensions. The first flow is given by the function $f(z) = z + \frac{1}{z}$, which in polar coordinates with $z = re^{i\theta}$ takes the form

$$f(z) = u(r, \theta) + iv(r, \theta) = (r + \frac{1}{r})\cos(\theta) + i(r - \frac{1}{r})\sin(\theta). \quad (251.35)$$

This represents a symmetric flow with the velocity $\approx (1, 0)$ (far) away from the disc, and the level curves of $v(r, \theta)$ give the streamlines of the flow around the disc, see Fig. 251.15.

The second flow is a flow circulating around the disc given by

$$g(r, \theta) = -\frac{iK}{2\pi}\log(z) = \frac{K}{2\pi}\theta + i(-\frac{K}{2\pi}\log(r)) \quad (251.36)$$

Consider now the flow given by $f(z) + g(z)$ with stream-function $(r - \frac{1}{r})\sin(\theta) - \frac{K}{2\pi}\log(r)$. We may consider this to be the flow around a spinning tennis ball in a horisontal stream of air, see Fig. 251.16

We now recall *Bernouilli's law* stating that for steady inviscid irrotational flow, the quantity

$$p + \frac{|U|^2}{2}$$

is constant because $\nabla(p + \frac{|U|^2}{2}) = 0$, which follows by direct computation, see Problem 251.13. We conclude that high velocity implies low

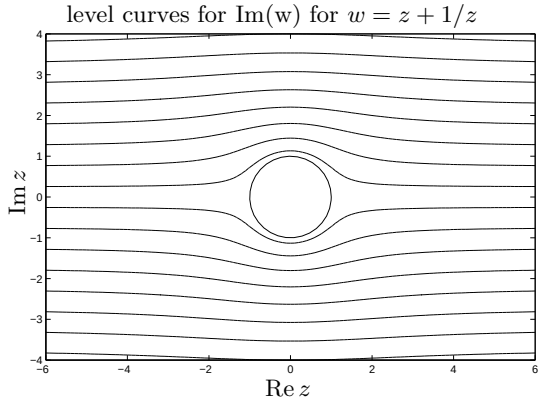


FIGURE 251.15. Level curves of $\mathrm{Im}(w)$ for $w = z + 1/z$.

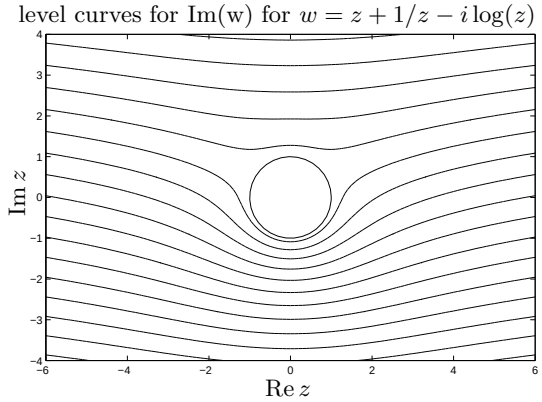


FIGURE 251.16. Level curves of $\mathrm{Im}(w)$ for $w = z + 1/z - i\log(z)$.

pressure. Now inspecting Fig. 251.16 we see that the velocity is high below the ball (dense level curves of the stream function), and thus the pressure is low below the ball and thus there will be a resulting force downward, which is referred to as *lift*. This is the reason a top-spin in tennis is so efficient in bringing down the ball inside the lines. The more top spin the more curved path of the ball! Björn Borg was one of the first to really exploit this law of mechanics. One can show that the *lift* is proportional to the *circulation* given by

$$\int_\Gamma u \cdot ds,$$

where Γ is the unit circle oriented counter-clockwise. The circulation of the flow given by (251.35) is equal to zero because of symmetry, while the circulation of the flow given by (251.36) is equal to K. The lift of the spinning tennis ball gives a hint to the mechanism behind flying. In fact, the design of an airplane wing with non-symmetric cross section and a sharp trailing edge creates a circulation around the wing which causes a lift, see Fig. 251.17.

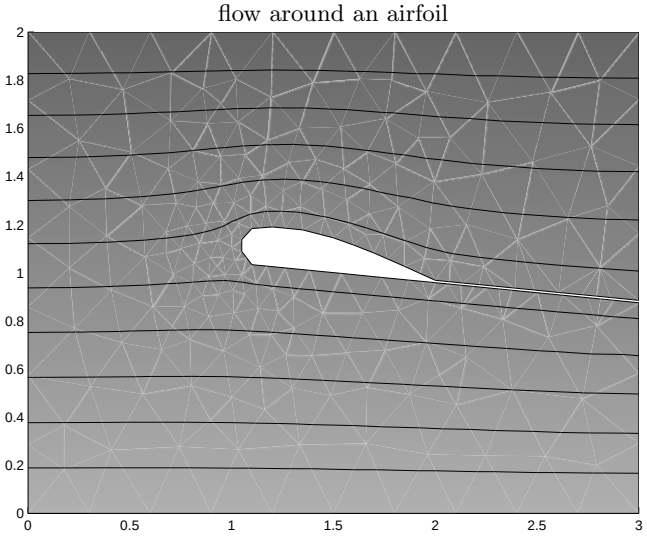


FIGURE 251.17. Potential flow around an airfoil.

The lifting clockwise circulation around the wing may be considered compensated for by a counter-clockwise vortex in the turbulent layer behind the wing, here localized to the line of discontinuity behind the wing, because the total rotation of the flow must be zero.

EXAMPLE 251.24. (**Flow through an aperture**) The function

$$\begin{aligned} z = \sin(w) &= \frac{1}{2i}(e^{i(u+iv)} - e^{-i(u+iv)}) \\ &= \sin(u)\cosh(v) + i(\cos(u)\sinh(v)), \end{aligned}$$

maps the strip $\{w = u + iv : \frac{-\pi}{2} < u < \frac{\pi}{2}, v \in \mathbb{R}\}$ onto $\{z = x + iy : y \neq 0 \text{ if } |x| > 1\}$, that is, the whole plane minus the two half-lines $\{x + iy : |x| > 1, y = 0\}$. The corresponding inverse function $w = f(z) = \arcsin(z) = \sin^{-1}(z)$ may be viewed as the potential for flow through an aperture, see Fig. 251.18.

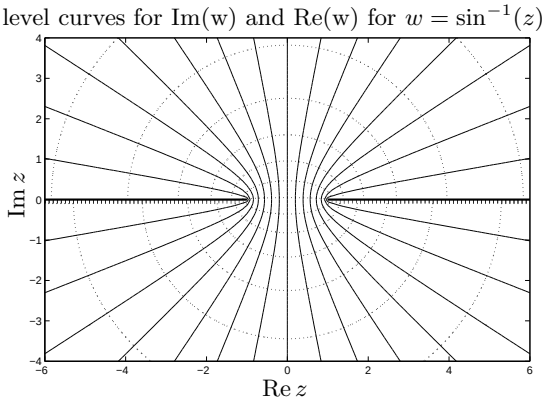


FIGURE 251.18. Level curves of $\mathrm{Im}(w)$ (solid) and $\mathrm{Re}(w)$ (dotted) for $w = \arcsin(z)$.

The streamlines are hyperbolas and the equipotential lines are ellipses.

EXAMPLE 251.25. (**Discontinuous electric potential**) Consider the function $f(z) = u(x, y) + iv(x, y) = i\log(z) = i\log(|z|) - \mathrm{Arg}\ z$ in the right half-plane $\{z \in \mathbb{C} : \mathrm{Re}\ z \geq 0\}$. We have $u(x, y) = \arctan(\frac{y}{x})$ and $v(x, y) = \log(r)$ with $r = (x^2 + y^2)^{1/2}$ and we plot the equi-potential and level curves of the curves in Fig. 251.19.

Note that the potential $u(x, y)$ approaches the value $\frac{\pi}{2}$ for x tending to zero if $y > 0$ and the value $-\frac{\pi}{2}$ for x tending to zero if $y < 0$, corresponding to discontinuous boundary values for $x = 0$.

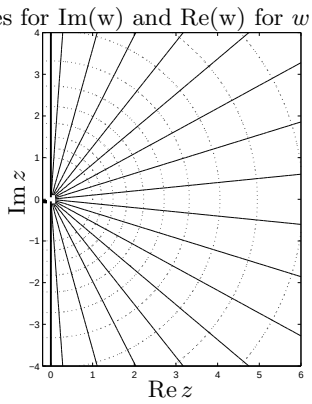


FIGURE 251.19. Level curves of Im(w) (solid) and Re(w) (dotted) for $w = i\log(z)$.

Chapter 251 Problems

251.1. (a) Prove that $f(z) = e^z$ is analytic and that $f'(z) = e^z$. (b) Prove that $\sin(z)$ and $\cos(z)$ are analytic with derivatives $\cos(z)$ and $-\sin(z)$, respectively.

251.2. It is possible to view an analytic function $f : \mathbb{C} \to \mathbb{C}$ as a function $F : \mathbb{R}^2 \to \mathbb{R}^2$ if we set $f(z) = u(x,y) + iv(x,y)$, $z = x + iy$ and $F(x,y) = (u(x,y), v(x,y))$. Explain the connection between the Jacobian F' of $F(x,y)$ and the derivative f', and motivate the Cauchy-Riemann equations this way.

251.3. What happens if we try to choose $z_0 \in \Gamma$ in Cauchy's representation formula?

251.4. Prove Liouville's theorem stating that if $f(z)$ is analytic in the whole complex plane and bounded, then $f(z)$ is constant. Hint: Use the representation formula for $f'(z)$ with Γ a circle with large radius.

251.5. Prove Morera's theorem stating that if $f : \Omega \to \mathbb{C}$ satisfies $\int : \Gamma f(z)\,dz = 0$ for all sinmple closed curves in Ω, then $f(z)$ is analytic in Ω. Hint: Define $F(z) = \int_{\Gamma_z} f(\zeta)\,d\zeta$, where Γ_z is a courve joining a fixed point z_0 with the variable point $z \in \Omega$. Show independence of the specific choice of Γ_z and then that $F'(z) = f(z)$.

251.6. Prove that a Möbius transformation maps every straight line or circle in the complex plane into a circle or straight line. Hint: write $w = \frac{az+b}{cz+d}$ in the form $w = -\frac{ad-bc}{c}\frac{1}{cz+d} + \frac{a}{c}$.

251.7. Compute (a) $\int_0^{2\pi} \frac{d\theta}{5-3\sin(\theta)}$, (b) $\int_{-\infty}^{\infty} \frac{x}{1+x^4}\,dx$.

251.8. Prove that $\int_{-\infty}^{\infty} \frac{\sin\theta}{\theta} = 2\pi$.

251.9. Prove (251.10).

251.10. Prove that if $u(x,y)$ is harmonic in a simply connected domain Ω, then there exists a function v such that $u+iv$ satisfies the Cauchy-Riemann equations in Ω. Hint use the central result of Chapter *Potential fields*.

251.11. Construct your own examples of 2d irrotational potential flow, electrostatics, and heat flow, by combining elementary functions such as z^α, e^z, $\log(z)$, $\sin(z)$, $\sinh(z)$ and Möbius transformations.

251.12. Give a different proof of Cauchy's representation theorem using that $\frac{g(z)}{z-z_0}$ is a analytic in the domain $\Omega_\epsilon = \{z \in \Omega : |z-z_0| > \epsilon\}$, so that $\int_{\Gamma_\epsilon} \frac{g(z)}{z-z_0}\,dz = 0$, where Γ_ϵ is the boundary of Ω_ϵ. Then let ϵ tend to zero.

251.13. Show that if the fluid velocity $u = (u_1, u_2)$ defined in a domain Ω in $\mathbb{R}^2$ satisfies $\nabla \cdot u = \nabla \times u = 0$ and solves the stationary momentum equation $(u\cdot\nabla)u + \nabla p - \Delta u = 0$ in Ω, then $\nabla(p + \frac{|u|^2}{2}) = 0$ in Ω. This proves Bernouilli's Law stating that $p + \frac{|u|^2}{2}$ is constant so that high velocity corresponds to low pressure.

251.14. Determine the images the circle $|z| = 1$ and the unit disc $|z| < 1$ under the mapping $w = \frac{i(1-z)}{1+z}$. Use the result to determine the elektrostatic potential $\varphi(x,y),\ (z = x+iy)$ in the unit disc $|z| < 1$ with boundary values

$$\varphi(x,y) = \begin{cases} P, & \text{om } |z| = 1, \quad x > 0,\ y > 0, \\ 0, & \text{om } |z| = 1, \quad x < 0, \text{ eller } y < 0. \end{cases}$$

251.15. Let T be a triangle with corners at $0,\ 1$ and $1+i$. Determine the image of T under the mapping $w = \frac{z}{1-z}$.

251.16. Determine a harmonic function $\varphi(x,y)$ in the domain between the hyoerbolas $x^2 - y^2 = 1$ and $x^2 - y^2 = 4$ with boundary values $\varphi(x,y) = 2xy$ on $x^2 - y^2 = 1$ and $\varphi(x,y) = 4xy$ on $x^2 - y^2 = 4$.

252
Fourier Series

Yesterday was my 21st birthday, at that age Newton and Pascal had already acquired many claims to immortality. (Fourier 1787, age 21)

252.1 Introduction

We give in the following two chapters a short account of *Fourier analysis* starting with *Fourier series* in this chapter and continuing in the next chapter to *Fourier transforms*. The basic idea is to represent (or approximate) given functions as linear combinations of trigonometric functions. We have met the same general idea in the Chapter Piecewise linear approximation, where we studied approximation of given functions as a linear combination of piecewise polynomials. Fourier representations have particular properties which are useful in for example signal/image processing with important applications to e.g. computer tomography. In recent years variants of Fourier techniques referred to as *Wavelets* have been developed with applications to for example compression of images. We touch this topic at the end of the Chapter *Fourier transforms*.

Fourier (1768-1830), see Fig. 252.1 used trigonometric series in his famous *Théorie analytique de chaleur* (1822) to study properties of solutions of the heat equation. The idea of expressing a general function as a Fourier series (or as a power series) has influenced the development of mathematical analysis profoundly with the driving force being the formidable success of these techniques for certain classes of problems, for example linear constant coefficient differential equations. However, as any highly specialized tool or

FIGURE 252.1. Fourier, Inventor of Fourier series: "Mathematics compares the most diverse of phenomena and discovers the secret analogies between them".

organism, these techniques have not been able to adapt to the needs of a changing world with computational methods for nonlinear differential equations taking over as work-horse in applications. Nevertheless, Fourier analysis still plays a fundamental role for the basic understanding of many phenomena.

We start with Fourier series in complex form and then pass to the real form as a special case. Fourier series concern functions $f : \mathbb{R} \to \mathbb{C}$ which are *periodic* with a certain period $a > 0$, that is $f(x + a) = f(x)$ for $x \in \mathbb{R}$. We often normalize to $a = 2\pi$ and thus consider 2π-periodic functions $f : \mathbb{R} \to \mathbb{C}$ satisfying $f(x + 2\pi) = f(x)$ for $x \in \mathbb{R}$. Usually we restrict attention to real-valued functions $f : \mathbb{R} \to \mathbb{R}$. Fourier transforms concern non-periodic functions $f : \mathbb{R} \to \mathbb{C}$.

We shall see that representing a given 2π-periodic function $f(x)$ as a Fourier series corresponds to expressing $f(x)$ as a linear combination of a certain set a trigonometric functions $\{e_m(x)\}$:

$$f(x) = \sum_m c_m e_m(x) \tag{252.1}$$

with certain coefficients $c_m \in \mathbb{C}$. We thus view the functions $e_m(x)$ as *basis functions* and express a general function $f(x)$ as a certain *linear combination* of basis functions. For example $f(x) = 0.5\sin(2x) - 0.8\sin(7x)$ is a linear combination of the two basis functions $\sin(2x)$ and $\sin(7x)$ with coefficients 0.5 and 0.8, see Fig. 252.2.

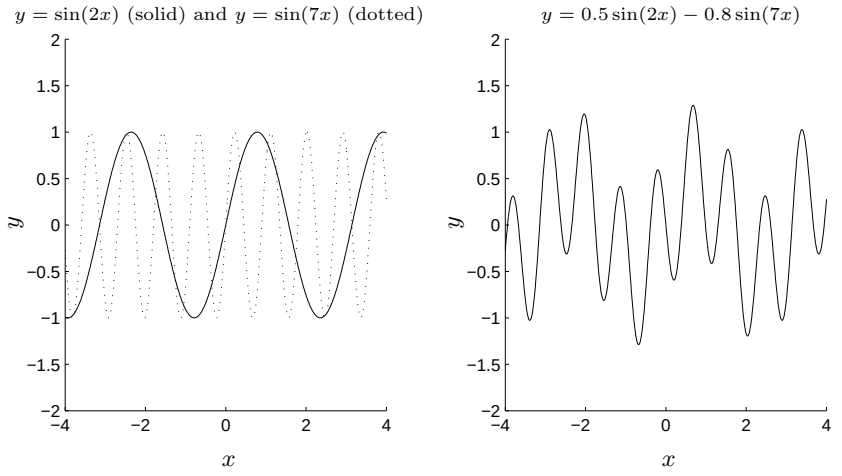


FIGURE 252.2. The functions $\sin(2x)$ and $\sin(7x)$, and the linear combination $0.5\sin(2x) + 0.8\sin(7x)$ of the two.

The trigonometric basis functions $e_m(x)$ used in Fourier series are of the form

$$\sin(mx), \quad \cos(mx), \ m = 0, 1, 2, ..., \quad \text{(real Fourier series)} \tag{252.2}$$

or

$$e^{imx} = \cos(mx) + i\sin(mx), \ m = 0, \pm 1, \pm 2, ..., \quad \text{(complex Fourier series)}. \tag{252.3}$$

Each basis function or "harmonic" $\sin(mx)$, $\cos(mx)$ or e^{imx}, is periodic with period $\frac{2\pi}{|m|}$ and (angular) *frequency* or *wave number* $|m|$. The larger $|m|$ is the higher is the frequency and the quicker do the basis functions $\sin(mx)$, $\cos(mx)$ and e^{imx} "oscillate". The series (252.1) expresses $f(x)$ as a linear combination of basis functions of increasing frequencies. Since the basis functions are all periodic with period 2π, so is their linear combination $f(x)$.

The basis functions (252.2) and (252.3) are *orthogonal* with respect to the $L_2(-\pi, \pi)$ scalar product

$$(v, w) = \int_{-\pi}^{\pi} v(x)\overline{w(x)}\, dx \tag{252.4}$$

with $\overline{w(x)}$ the complex conjugate of $w(x)$, with corresponding norm $\|v\| = (v, v)^{1/2}$. The orthogonality makes the coefficients c_m directly computable upon taking the $L_2(-\pi, \pi)$ scalar product of (252.1) with e_m to give

$$c_m = \frac{(f, e_m)}{(e_m, e_m)}.$$

.

252.2 Warm Up I: Orthonormal Basis in $\mathbb{C}^n$

To prepare we consider an analogous situation in $\mathbb{C}^n$: We recall that $\mathbb{C}^n$ is the set of ordered n-tuples $x = (x_1, ..., x_n)$ with $x_k \in \mathbb{C}$ for $k = 1, ..., n$. The scalar product (x, y) of two vectors x and y in $\mathbb{C}^n$ is defined by $x \cdot y = (x, y) = \sum_{j=1}^n x_j \overline{y_j}$, with corresponding norm $|x| = (x, x)^{1/2}$

Let now $\{g_1, ..., g_n\}$ be a set of n vectors in $\mathbb{C}^n$, that is each $g_k = (g_{k1}, ..., g_{kn})$ is a vector in $\mathbb{C}^n$ with components $g_{kj} \in \mathbb{C}$. We recall that the set $\{g_1, ..., g_n\}$ is an orthonormal basis in $\mathbb{C}^n$ if the g_k are mutually orthogonal and have norm equal to one, that is

$$(g_k, g_m) = 0 \quad \text{if } k \neq m, \quad \text{and } |g_m| = 1 \quad \text{for } m = 1, ...n.$$

If $\{g_1, ..., g_n\}$ is an orthonormal basis, then we can express a given vector $u \in \mathbb{C}^n$ as a linear combination of basis vectors in the form

$$u = \sum_{k=1}^n c_m g_m, \quad \text{where } c_m = (u, g_m) \quad \text{for } m = 1, ...n,$$

where the fact that $c_m = (u, g_m)$ follows by taking the scalar product and using the orthonormality.

252.3 Warm Up II: Series

We recall from Chapter *Series* that a series $\sum_{m=1}^\infty \alpha_m$ with coefficients $\alpha_m \in \mathbb{C}$, is said to be *convergent* if the sequence $\{s_n\}_{n=1}^\infty$ of partial sums $s_n = \sum_{m=1}^n \alpha_m$ converges as n tends to infinity. The series is said to be *absolutely convergent* if $\sum_{m=1}^\infty |\alpha_m|$ is convergent, which is the same as requiring the sequence of partial sums $\hat{s}_n = \sum_{m=1}^n |\alpha_m|$ to be bounded above, that is $\hat{s}_n \leq K$ for $n = 1, 2, ...$, where K is a positive constant. For a series with non-negative terms, the concepts of convergence and absolute convergence coincide. A typical example of a positive (absolutely) convergent series is given by $\sum_{m=1}^\infty m^{-2}$. To see that $s_n = \sum_{m=1}^n m^{-2}$ is bounded above, we use the fact that

$$s_n \leq 1 + \sum_{m=2}^n \int_{m-1}^m x^{-2}\, dx \leq 1 + \int_1^n x^{-2}\, dx \leq 1 + [-x^{-1}]_1^n \leq 2.$$

The same argument shows that $\sum_{m=1}^\infty m^{-\alpha}$ is convergent if $\alpha > 1$.

We also recall that an alternating series of the form $\sum_{m=1}^\infty (-1)^m a_m$, with $\{a_m\}$ a decreasing positive sequence tending to zero, is convergent.

252.4 Complex Fourier Series

A series of the form

$$\sum_{m=-\infty}^{\infty} c_m e^{imx} = \sum_{1}^{\infty} c_{-m} e^{-imx} + c_0 + \sum_{1}^{\infty} c_m e^{imx}, \tag{252.5}$$

where $x \in \mathbb{R}$, is said to be a *Fourier series* with *Fourier coefficients* $c_m \in \mathbb{C}$, $m = 0, \pm 1, \pm 2, ...$ The corresponding *truncated Fourier series*

$$\sum_{m=-n}^{n} c_m e^{imx} = \sum_{m=1}^{n} c_{-m} e^{-imx} + c_0 + \sum_{m=1}^{n} c_m e^{imx}, \tag{252.6}$$

where $n = 1, 2, ...$, may be viewed as a finite linear combination of the set of basis functions

$$\{1, e^{\pm ix}, e^{\pm i2x}, ..., e^{\pm inx}\}$$

with coefficients c_m.

The orthogonality of the basis functions $\{e^{imx}\}$ is expressed by:

$$\int_{-\pi}^{\pi} e^{imx} e^{-ikx} dx = \begin{cases} 0 & \text{if } k \neq m, \\ 2\pi & \text{if } k = m, \end{cases} \tag{252.7}$$

which follows by direct integration.

We shall typically consider cases with the Fourier coefficients c_m satisfying for some positive constant K,

$$|c_m| \leq K m^{-2}, \quad m = \pm 1, \pm 2, ... \tag{252.8}$$

In this case the series (252.5) converges absolutely for all x, since

$$\sum_{-\infty}^{\infty} |c_m e^{imx}| = \sum_{-\infty}^{\infty} |c_m| \leq |c_0| + 2K \sum_{m=1}^{\infty} m^{-2} < \infty,$$

and thus defines a function $f : \mathbb{R} \to \mathbb{C}$ represented by a converging Fourier series:

$$f(x) = \sum_{m=-\infty}^{\infty} c_m e^{imx}. \tag{252.9}$$

The series (252.9) gives a *spectral decomposition* of $f(x)$ into harmonics e^{imx} with different amplitudes c_m. The series (252.9) thus gives a description of the function $f(x)$ in terms of amplitudes of different harmonics included in $f(x)$. In musical terms we may think of $f(x)$ as a "chord" built by a number of "tones" $c_m e^{imx}$ of different frequencies m and amplitudes c_m. A spectral decomposition of a "chord" $f(x)$ would display the "tones" building the "chord", try *The Sound of Functions* in the *Mathematics Laboratory* for a direct experience.

We note that the basis functions $\{e^{imx}\}$ have *global support*, that is, each basis function e^{imx} is nonzero for all $x \in \mathbb{R}$. The basis functions $\{e^{imx}\}$ thus combines the following properties: orthogonality and global support. We contrast this to the 'hat functions' which are the basis functions for continuous piecewise linear approximation: the hat functions have local support but are not (quite) orthogonal. The best combination would be orthogonality together with local support. So-called *wavelets* introduced in recent years combine these properties.

Suppose now that $f(x)$ is defined by a converging Fourier series (252.9). Multiplying by e^{-imx} with $m = 0, \pm 1, \pm 2, ...$ and integrating over the interval $[-\pi, \pi]$, and using the orthogonality properties (252.7), we find that

$$c_m = c_m(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-imx}\,dx, \tag{252.10}$$

where we indicated the dependence of the Fourier coefficient $c_m = c_m(f)$ on the function $f(x)$. We thus have the *Fourier series representation*

$$f(x) = \sum_{m=-\infty}^{\infty} c_m(f)e^{imx}, \tag{252.11}$$

expressing $f(x)$ as a linear combination of different harmonics e^{imx} with different frequencies, where the Fourier coefficients $c_m(f)$ are given by (252.10).

Conversely, if $f : \mathbb{R} \to \mathbb{C}$ is a given 2π-periodic (Lipschitz continuous) function and we define $c_m(f)$ by (252.10), then we may ask if $f(x)$ can be represented by its Fourier series (252.11) for all x. We shall prove below that this is true if $f(x)$ is 2π-periodic and differentiable. This is the basic result of Fourier analysis stating that an arbitrary 2π-periodic differentiable function can be given a spectral decomposition in the form of a Fourier series. This result includes the "completeness" aspect of the basis functions $\{e^{imx}\}$, that is, the fact that *any* differentiable function can be represented as a Fourier series.

252.5 Fourier Series as an Orthonormal Basis Expansion

Normalizing the basis functions e^{imx} we obtain the orthonormal basis functions $e_m(x) = \frac{1}{\sqrt{2\pi}}e^{imx}$ satisfying

$$(e_m, e_k) = 0 \quad \text{if } k \neq m, \quad (e_m, e_m) = 1. \tag{252.12}$$

A Fourier series representation takes the following form in the normalized basis:

$$f(x) = \sum_{m=-\infty}^{\infty} \tilde{c}_m(f)\frac{1}{\sqrt{2\pi}}e^{imx}, \quad \tilde{c}_m(f) = \frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi} f(x)e^{-imx}\,dx.$$

Of course, it would be natural to work with the normalized basis functions $\{\frac{1}{\sqrt{2\pi}}e^{imx}\}$ and the corresponding renormalized Fourier coefficients $\tilde{c}_m(f)$ thus distributing the 2π-factor into two $\sqrt{2\pi}$-factors, but we follow the most common notation and include the 2π-factor in the Fourier coefficient $c_m(f)$ coupled to the basis function e^{imx}, which also simplifies notation somewhat.

252.6 Truncated Fourier Series and Best L_2-Approximation

The *truncated Fourier series*

$$S_n f(x) = \sum_{m=-n}^{n} c_m(f)e^{imx}$$

of a given function $f(x)$ is a *best approximation* of $f(x)$ in the sense that

$$\|f - S_n f\| \le \|f - g_n\|$$

for any function $g_n(x) = \sum_{m=-n}^{n} d_m e^{imx}$ with $d_m \in \mathbb{C}$, $m = 0, \pm 1, ..., \pm n$. This is because, by the definition of the Fourier coefficients,

$$(f - S_n f, e_m) = 0 \quad \text{for } m = 0, \pm 1, ..., \pm n,$$

and thus $S_n f(x)$ is the best approximation in the $L_2(-\pi, \pi)$ norm of $f(x)$ in the linear space spanned by the functions $\{1, e^{\pm ix}, e^{\pm i2x}, ..., e^{\pm inx}\}$, compare Chapter *Piecewise Linear Approximation*.

252.7 Real Fourier Series

Using that $e^{imx} = \cos(mx) + i\sin(mx)$ and $\cos(-mx) = \cos(mx)$ and $\sin(-mx) = -\sin(mx)$, we can write (252.9) in the form

$$\sum_{m=-\infty}^{\infty} c_m e^{imx} = c_0 + \sum_{m=1}^{\infty} a_m \cos(mx) + \sum_{m=1}^{\infty} b_m \sin(mx),$$

where

$$a_m = c_m + c_{-m}, \quad b_m = i(c_m - c_{-m}), \quad m = 1, 2, ...$$

If $f(x)$ is real, that is $f : \mathbb{R} \to \mathbb{R}$, then $\bar{c}_m = c_{-m}$ and thus $a_m = c_m + \bar{c}_m = 2\mathrm{Re}(c_m) \in \mathbb{R}$ and $b_m = i(c_m - \bar{c}_m) = -2\mathrm{Im}(c_m) \in \mathbb{R}$, and

$$c_m = \frac{a_m}{2} - i\frac{b_m}{2}, \quad c_{-m} = \frac{a_m}{2} + i\frac{b_m}{2}, \quad m = 0, 1, 2, ... \tag{252.13}$$

The Fourier series of a real-valued 2π-periodic function $f : \mathbb{R} \to \mathbb{R}$ can thus be written alternatively as a Sine and Cosine series of the form

$$f(x) = \frac{a_0}{2} + \sum_{m=1}^{\infty} a_m \cos(mx) + \sum_{m=1}^{\infty} b_m \sin(mx),$$

where a_m, $b_m \in \mathbb{R}$, are given by

$$a_m = a_m(f) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(mx)\,dx \quad \text{for } m = 0, 1, 2, ...,$$
$$b_m = b_m(f) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(mx)\,dx \quad \text{for } m = 1, 2, ...$$

We note that if $f(x)$ is even, that is $f(x) = f(-x)$, then $b_m = 0$ for $m = 1, 2, ...$, and thus $f(x)$ has a *Cosine series representation:*

$$f(x) = \frac{a_0(f)}{2} + \sum_{m=1}^{\infty} a_m(f)\cos(mx). \tag{252.14}$$

Correspondingly, if $f(x)$ is odd, that is $f(x) = -f(-x)$, then $a_m = 0$ for $m = 0, 1, ...$, and thus $f(x)$ has a *Sine series representation:*

$$f(x) = \sum_{m=1}^{\infty} b_m(f)\sin(mx). \tag{252.15}$$

In the applications below we usually consider Cosine and Sine series for real-valued functions $f : \mathbb{R} \to \mathbb{R}$. The complex Fourier series is useful in the analysis of convergence of Fourier series.

We now present a couple of examples with Fourier coefficients having different rates of convergence to zero (as m^{-2}, m^{-3} and m^{-1}).

EXAMPLE 252.1. Let $f : \mathbb{R} \to \mathbb{R}$ be a 2π-periodic function given by $f(x) = |x|$ for $-\pi \le x \le \pi$. The function $f(x)$ is real-valued and even, and thus has a Cosine series of the form (252.14). We compute using integration by parts if $m > 0$:

$$a_0(f) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\,dx = \frac{2}{\pi}\int_0^{\pi} x\,dx = \pi,$$
$$a_m(f) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(mx)\,dx = \frac{2}{\pi}\int_0^{\pi} x\cos(mx)\,dx$$
$$= \frac{2}{\pi}\Big[\frac{x\sin(mx)}{m}\Big]_0^{\pi} - \frac{2}{\pi}\int_0^{\pi}\frac{\sin(mx)}{m}\,dx = \frac{2}{\pi}\frac{(-1)^m - 1}{m^2}.$$

Since $(-1)^m - 1 = -2$ if m is odd and $(-1)^m - 1 = 0$ if m is even, the Fourier series representation of $f(x) = |x|$ takes the form

$$|x| = \frac{\pi}{2} - \frac{4}{\pi}\sum_{k=1}^{\infty}\frac{\cos((2k-1)x)}{(2k-1)^2}.$$

We plot the corresponding truncated series with summation over $k = 1, ..., n$ for different values of n in Fig. 252.3:

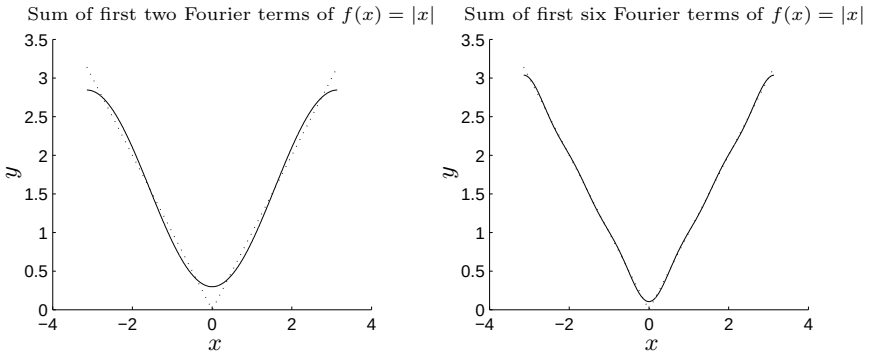


FIGURE 252.3. The sum of the first two and first six terms of the fourier series of $|x|$ (dotted).

EXAMPLE 252.2. Let $f : \mathbb{R} \to \mathbb{R}$ be an odd 2π-periodic function given by $f(x) = x(\pi - x)$ for $0 \le x \le \pi$. We compute its Sine series coefficients:

$$\begin{aligned} b_m(f) &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(mx)\,dx = \frac{2}{\pi}\int_0^{\pi} x(\pi - x)\sin(mx)\,dx \\ &= -\frac{2}{\pi}\Big[\frac{x(\pi - x)\cos(mx)}{m}\Big]_0^{\pi} + \frac{2}{\pi}\int_0^{\pi}\frac{(\pi - 2x)\cos(mx)}{m}\,dx \\ &= \frac{2}{\pi m}\Big[\frac{(\pi - 2x)\sin(mx)}{m}\Big]_0^{\pi} + \frac{2}{\pi m^2}\int_0^{\pi} 2\sin(mx)\,dx \\ &= \frac{4}{\pi m^3}(1 - (-1)^m). \end{aligned}$$

EXAMPLE 252.3. Define a 2π-periodic function $f(x)$ by setting

$$f(x) = \begin{cases} 1 & \text{for } |x| < a, \\ 0 & \text{for } a < |x| \le \pi, \end{cases}$$

where $0 < a < \pi$, see Fig. 252.4.

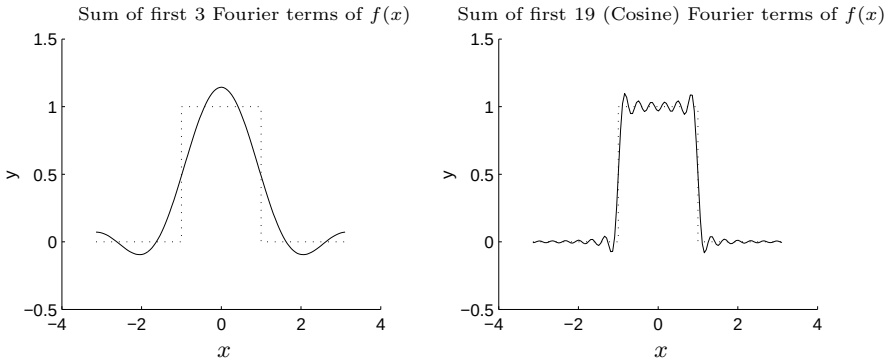


FIGURE 252.4. The sum of the first 3 and first 19 terms of the fourier series of a piecewise constant function (dotted).

This is a piecewise Lipschitz continuous 2π-periodic even function, and we can compute its Fourier coefficients. We have $b_m(f) = 0$ and, for $m > 0$, $2c_m(f)$

$$= a_m(f) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(mx)\,dx = \frac{2}{\pi}\int_0^a \cos(mx)\,dx = \frac{2\sin(ma)}{\pi m}, \tag{252.16}$$

while $a_0(f) = \frac{2a}{\pi}$. We thus expect that

$$f(x) = \frac{a}{\pi} + \frac{2}{\pi}\sum_{m=1}^{\infty}\frac{\sin(ma)}{m}\cos(mx).$$

We shall return to this equality below, with particular focus on the values $x = \pm a$ where $f(x)$ has jump discontinuities.

252.8 Basic Properties of Fourier Coefficients

We now present some basic properties of the Fourier coefficients

$$c_m(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-imx}\,dx \quad m = 0, \pm 1, \pm 2, ...,$$

of a given 2π-periodic Lipschitz continuous function $f : \mathbb{R} \to \mathbb{C}$.

Linearity

Fourier coefficients satisfy the following obvious linearity properties:

$$c_m(f + g) = c_m(f) + c_m(g), \quad c_m(\alpha f) = \alpha c_m(f),$$

where f and g are two functions with Fourier coefficients $c_m(f)$ and $c_m(g)$, and $\alpha \in \mathbb{C}$.

Fourier Coefficients of the Derivative $Df = f'$

We now couple the Fourier coefficients of the derivative $Df = \frac{df}{dx}$ of a 2π-periodic function $f : \mathbb{R} \to \mathbb{C}$ to the Fourier coefficients of f. The trick is to integrate by parts: Using the periodicity of $f(x)$, we find that

$$c_m(Df) = \frac{1}{2\pi}\int_{-\pi}^{\pi} Df(x)e^{-imx}\,dx = im\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-imx}\,dx = im\,c_m(f),$$

and we have thus proved:

Theorem 252.1 *If $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic and differentiable with derivative Df, then for $m = 0, \pm 1, \pm 2, ...$*

$$c_m(Df) = im\,c_m(f). \tag{252.17}$$

This is one of the fundamental results of Fourier analysis, and translates the operation of differentiation $D = \frac{d}{dx}$ with respect to x to multiplication of Fourier coefficients with im where m is the frequency. This opens the way of translating differential equations in the variable x to algebraic equations in the frequency m, which may be very useful and illuminating in certain applications.

We can directly generalize to

Theorem 252.2 *If $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic and k times differentiable with derivative $D^k f$, then for $m = 0, \pm 1, \pm 2, ...$*

$$c_m(D^k f) = (im)^k c_m(f). \tag{252.18}$$

EXAMPLE 252.4. Consider the differential equation $Du(x) + u(x) = f(x)$, where $f(x)$ a given 2π-periodic function and we seek a 2π-periodic solution $u(x)$. This equation models, for example, a resistor and capacitor in series, with $u(x)$ a primitive function of the current, $f(x)$ an applied voltage, and x representing time, see the Chapter Electrical circuits. Alternatively, $Du(x) + u(x) = f(x)$ models an inductor and resistance in series with $u(x)$ now the current, and again $f(x)$ an applied voltage. For the Fourier coefficients we have using Theorem 252.1

$$im\,c_m(u) + c_m(u) = c_m(f),$$

and thus

$$c_m(u) = \frac{c_m(f)}{1 + im} = \frac{(1 - im)c_m(f)}{1 + m^2}.$$

This shows that the indicated circuits act as so-called *low-pass filters*, with the property of damping high-frequency components: we view $f(x)$ as the input and $u(x)$ as the output and note that the Fourier coefficients of $u(x)$ decay quicker than those of $f(x)$.

EXAMPLE 252.5. Consider the differential equation $-D^2u(x)+u(x) = f(x)$ with $f(x)$ a given 2π-periodic function and we seek a 2π-periodic solution $u(x)$. Since $c_m(D^2u) = (im)^2c_m(u)$, we obtain the following algebraic equation for the Fourier coefficients:

$$(m^2+1)c_m(u) = c_m(f) \quad \text{for } m \neq 0.$$

We can thus express the solution $u(x)$ of $-D^2u(x) = u(x) = f(x)$ as a Fourier series

$$u(x) = \sum_{-\infty}^{\infty} \frac{c_m(f)}{m^2+1} e^{imx},$$

if the data $f(x)$ is given as a Fourier series: $f(x) = \sum_{-\infty}^{\infty} c_m(f)e^{imx}$. Again, we see that the differential equation acts as a low-pass filter with damping of high-frequency components of the data $f(x)$.

EXAMPLE 252.6. More generally, consider the following differential equation $p(D)u(x) = f(x)$, where $p(D) = \sum_{k=0}^{q} a_k D^k$ is a differential equation with constant coefficients $a_k \in \mathbb{C}$, the data $f(x)$ is 2π-periodic and we seek a 2π-periodic solution $u(x)$. Arguing as above, we get the following equation for the Fourier coefficients:

$$p(im)c_m(u) = \sum_{k=0}^{q} a_k(im)^k c_m(u) = c_m(f),$$

that is, assuming $p(im) \neq 0$ (or $c_m(f) = 0$ if $p(im) = 0$),

$$c_m(u) = \frac{c_m(f)}{p(im)},$$

which gives the Fourier series for the solution, if the Fourier series for the data $f(x)$ is given.

The Fourier coefficients $c_m(f)$ tend to zero as $|m| \to \infty$

As a direct consequence of the preceding result, we conclude that the Fourier coefficients $c_m(f)$ of a 2π-periodic differentiable function $f(x)$ with integrable derivative Df, tend to zero as $|m|$ tends to infinity: Since $|im\, c_m(f)| = |c_m(Df)|$, we have

$$|c_m(f)| = \frac{1}{|m|}|c_m(Df)| \leq \frac{1}{2\pi|m|}\int_{-\pi}^{\pi} |Df|\, dx \to 0 \quad \text{as } |m| \to \infty.$$

Similarly, if $f(x)$ is 2π-periodic with integrable derivative $D^k f$ of order $k > 1$, then for $m = \pm 1, \pm 2, ...,$

$$|c_m(f)| \leq \frac{1}{2\pi|m^k|}\int_{-\pi}^{\pi} |D^k f|\, dx.$$

We conclude that the larger k is, the more rapid is the convergence of $c_m(f)$ to zero.

We can also go in the direction of less regularity and ask if we can show that the Fourier coefficients $c_m(f)$ tend to zero as $|m| \to \infty$ under the weaker assumption that f is Lipschitz continuous only. To this end we first note that for any $-\pi < a < b < \pi$, we have

$$\int_a^b e^{-imx}dx = \frac{1}{-im}[e^{-imx}]_a^b \to 0 \quad \text{as } m \to \infty. \tag{252.19}$$

This may be seen as a consequence of the rapid oscillations of e^{-imx} with $|m|$ large, which causes a lot of cancellations in any integral of the form (252.19) with the effect that the integrals decreases to zero as m increases to infinity, see the following figure.

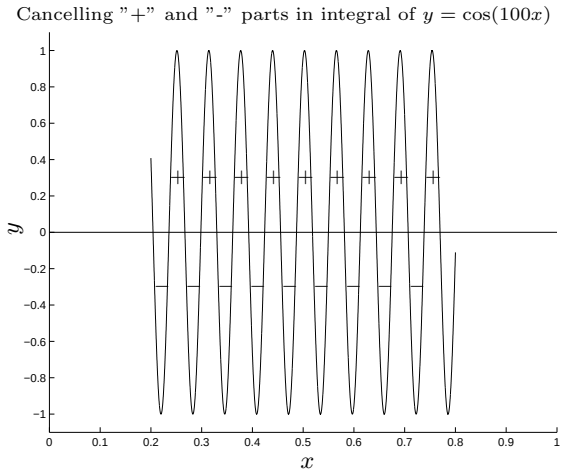


FIGURE 252.5. An illustration of the fact that $\int_a^b \cos(mx)dx$ and $\int_a^b \sin(mx)dx$ is small for m large.

The estimate (252.19) shows that if $f(x)$ is piecewise constant on $[-\pi, \pi]$, that is a linear combination sum of functions equal to one on a certain interval and zero elsewhere, then $c_m(f) \to 0$ as $|m| \to \infty$.

Finally, a given Lipschitz continuous function $f : [-\pi, \pi] \to \mathbb{C}$ can be approximated by a piecewise constant function $\tilde{f}(x)$, so that

$$\int_{-\pi}^{\pi} |f(x) - \tilde{f}(x)|$$

is as small as we please, which leads to the famous

Theorem 252.3 (Riemann-Lebesgue lemma) *If $f : [-\pi, \pi]$ is Lipschitz continuous, then $c_m(f) \to 0$ as $|m| \to \infty$.*

The assumption can be relaxed to piecewise Lipschitz continuity.

Convolution

Given two 2π-periodic functions $f(x)$ and $g(x)$, we define a new 2π-periodic function $f * g$ by

$$(f * g)(x) = \int_{-\pi}^{\pi} f(x-y)g(y)\,dy \quad x \in \mathbb{R}.$$

We say that $f * g$ is the *convolution* of f and g. Changing variables, setting $y = x - t$, we find that

$$(f * g)(x) = \int_{-\pi}^{\pi} f(t)g(x-t)\,dt = \int_{-\pi}^{\pi} f(y)g(x-y)\,dy \quad x \in \mathbb{R},$$

and thus the integrand can take the form $f(x-y)g(y)$ or $f(y)g(x-y)$.

We shall now prove that

$$c_m(f * g) = 2\pi\, c_m(f)c_m(g). \tag{252.20}$$

By direct computation, changing order of integration and using the change of variable $t = x - y$, we have

$$\begin{aligned}
c_m(f * g) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} (f * g)(x)e^{-imx}\,dx \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x-y)g(y)\,dy\,e^{-imx}\,dx \\
&= \int_{-\pi}^{\pi} g(y)e^{-imy}\Big(\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-y)\,e^{-im(x-y)}\,dx\Big)dy \\
&= \int_{-\pi}^{\pi} g(y)e^{-imy}\Big(\frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\,e^{-imt}\,dt\Big)dy \\
&= c_m(f)\int_{-\pi}^{\pi} g(y)e^{-imy}\,dy = 2\pi c_m(f)c_m(g).
\end{aligned}$$

EXAMPLE 252.7. Let $g : \mathbb{R} \to \mathbb{R}$ be a 2π-periodic function defined by

$$g(x) = \frac{1}{2a} \quad \text{for } -a \le x \le a,$$

where $0 < a < \pi$. For a small, we may view $g(x)$ as an approximate delta function. The convolution

$$(f * g)(x) = \int_{-\pi}^{\pi} f(x-y)g(y)\,dy = \frac{1}{2a}\int_{-a}^{a} f(x-y)\,dy$$

is an average of $f(x)$ over the interval $[x-a, x+a]$. Recalling (252.16), and using (252.20), we get

$$c_m(f * g) = c_m(f)\frac{\sin(ma)}{ma}.$$

We conclude that $c_m(f*g)$ is close to $c_m(f)$ if ma is small, and $c_m(f*g)$ is much smaller than $c_m(f)$ if ma is large. The Fourier coefficients of the average $f*g$ thus decay quicker than those of f, and thus $f*g$ is a smoothed version of f: taking the average increases smoothness reflected by quickly decreasing Fourier coefficients.

252.9 The Inversion Formula

We shall now prove that if $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic and differentiable, then for all $x \in \mathbb{R}$

$$\lim_{n\to\infty} \sum_{-n}^{n} c_m(f) e^{imx} = f(x).$$

In other words, the function $f(x)$ can be represented as a convergent Fourier series:

$$f(x) = \sum_{m=-\infty}^{\infty} c_m(f) e^{imx} \quad \text{for } x \in \mathbb{R}.$$

We have

$$\begin{aligned} \sum_{-n}^{n} c_m e^{imx} &= \sum_{-n}^{n} \frac{1}{2\pi} \int_{-\pi}^{\pi} f(y) e^{-imy}\, dy\, e^{imx} \\ &= \int_{-\pi}^{\pi} f(y) \frac{1}{2\pi} \sum_{-n}^{n} e^{im(x-y)}\, dy = \int_{-\pi}^{\pi} f(y) D_n(x-y)\, dy, \end{aligned} \tag{252.21}$$

where, setting $\theta = x - y$,

$$\begin{aligned} D_n(\theta) &= \frac{1}{2\pi} \sum_{-n}^{n} e^{im\theta} = \frac{1}{2\pi} e^{-in\theta} \sum_{m=0}^{2n} e^{im\theta} \\ &= \frac{1}{2\pi} e^{-in\theta} \frac{1 - e^{i(2n+1)\theta}}{1 - e^{i\theta}} = \frac{1}{2\pi} \frac{e^{-i\frac{\theta}{2}}}{e^{-i\frac{\theta}{2}}} \, \frac{e^{-in\theta} - e^{i(n+1)\theta}}{1 - e^{i\theta}} = \frac{1}{2\pi} \frac{\sin(n\theta + \frac{\theta}{2})}{\sin(\frac{\theta}{2})} \end{aligned}$$

is the so-called *Dirichlet kernel*. We here used that $\sum_{m=0}^{2n} e^{im\theta}$ is a finite geometric series with factor $e^{i\theta}$. Using the convolution notation we can write (252.21) in the compact form

$$\sum_{-n}^{n} c_m e^{imx} = f * D_n(x).$$

In order for $f * D_n(x)$ to approximate $f(x)$, we expect D_n to somehow behave like the identity. We look at a plot of $D_n(\theta)$:

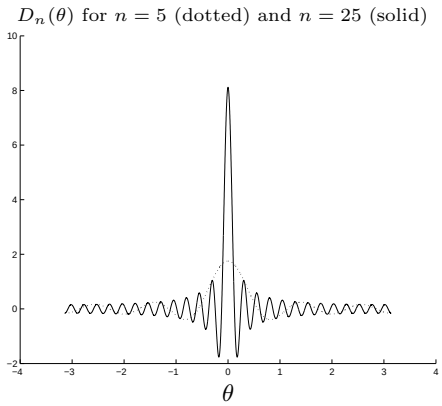


FIGURE 252.6. A plot of $D_n(\theta)$.

We see that $D_n(\theta)$ oscillates and has a peak at $\theta = 0$. Integrating $D_n(\theta) = \frac{1}{2\pi}\sum_{-n}^{n} e^{im\theta}$ term by term over $[-\pi, \pi]$ noting that all integrated terms vanish but one, we see that the total area (with sign) under the graph of D_n is equal to one, that is

$$\int_{-\pi}^{\pi} D_n(\theta)\, d\theta = 1, \tag{252.22}$$

which expresses one aspect of the idea that D_n behaves like the identity. The other aspect of the approximate identity nature of D_n is the increasing focussing of the peak of D_n at 0 as n increases.

Using (252.22), we can write

$$f(x) - f * D_n(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} (f(x) - f(y))D_n(x-y)\, dy$$
$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} g(x,y)\sin((n+\frac{1}{2})(x-y))\, dy$$

where

$$g(x,y) = \frac{f(x) - f(y)}{\sin(\frac{x-y}{2})}.$$

Now if $f(x)$ is twice differentiable, then $g(x,y)$ is differentiable with respect to y for all $y \in \mathbb{R}$ with derivative $Dg(x,y)$ (see the corresponding argument in the proof of Cauchy's formula). Integrating by parts we thus have

$$f(x) - D_n * f(x) = -\frac{1}{2\pi}\frac{1}{n+\frac{1}{2}}\int_{-\pi}^{\pi} Dg(x,y)\cos((n+\frac{1}{2})(x-y))\, dy \to\ 0$$

as $n \to \infty$. In case $f(x)$ is differentiable with piecewise Lipschitz continuous derivative, then $Dg(x,y)$ is Lipschitz continuous in y, and by Riemann-Lebesgue' lemma, we find the same conclusion. We summarize in the following basic theorem:

Theorem 252.4 *If $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic with piecewise Lipschitz continuous derivative, then $f(x)$ may be represented by a convergent Fourier series:*

$$f(x) = \sum_{m=-\infty}^{\infty} c_m(f) e^{imx} \quad \text{for } x \in \mathbb{R},$$

where the coefficients c_m are given by (252.10).

The assumption on $f(x)$ can be relaxed: it suffices to assume that $f(x)$ is piecewise differentiable with piecewise Lipschitz continuous derivative. At a point x of discontinuity, the Fourier series converges to the mean value of the left hand limit $f^-(x) = \lim_{y \to x, y<x} f(y)$ and the right hand limit $f^+(x) = \lim_{y \to x, y>x} f(y)$:

$$\sum_{m=-\infty}^{\infty} c_m(f) e^{imx} = \frac{f^-(x) + f^+(x)}{2}. \tag{252.23}$$

EXAMPLE 252.8. We have

$$\sum_{m=1}^{\infty} \frac{\sin(ma)}{\pi\, m} \cos(mx) = \begin{cases} 1 & \text{if } |x| < a, \\ \frac{1}{2} & \text{if } |x| = a, \\ 0 & \text{if } |x| > a. \end{cases}$$

252.10 Parseval's and Plancherel's Formulas

Suppose $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic with a convergent Fourier series representation:

$$f(x) = \sum_{m=-\infty}^{\infty} c_m(f) e^{imx}, \tag{252.24}$$

where

$$c_m(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-imx}\, dx.$$

Using the orthogonality (252.7) of the functions $\{e^{imx}\}$, we find that

$$\begin{aligned} \int_{-\pi}^{\pi} |f(x)|^2\, dx = \int_{-\pi}^{\pi} f(x)\overline{f(x)}\, dx &= \int_{-\pi}^{\pi} \Big(\sum_{m=-\infty}^{\infty} c_m(f) e^{imx} \Big) \Big(\sum_{k=-\infty}^{\infty} \overline{c_k(f)} e^{-ikx} \Big)\, dx \\ &\sum_{m,k=-\infty}^{\infty} c_m(f)\overline{c_k(f)} \int_{-\pi}^{\pi} e^{imx} e^{-ikx}\, dx \\ &= 2\pi \sum_{m=-\infty}^{\infty} |c_m(f)|^2. \end{aligned}$$

We have now proved the celebrated:

Theorem 252.5 (Parseval's formula) *If $f(x)$ has a convergent Fourier series representation, then*

$$\int_{-\pi}^{\pi} |f(x)|^2\,dx = 2\pi \sum_{m=-\infty}^{\infty} |c_m(f)|^2.$$

We can in an obvious way generalize to obtain:

Theorem 252.6 (Plancherel's formula) *If $f(x)$ and $g(x)$ have convergent Fourier series representations, then*

$$\int_{-\pi}^{\pi} f(x)\overline{g(x)}\,dx = 2\pi \sum_{m=-\infty}^{\infty} c_m(f)\overline{c_m(g)}.$$

252.11 Space Versus Frequency Analysis

We are now ready to lean back and reflect a bit about the nature of Fourier series. Suppose that $f(x)$ is a given 2π-periodic function. If we want to describe the nature of the function $f(x)$, that is the variation of $f(x)$ with x, we can try to give some kind of list of $f(x)$ values for different values of x. We may call this a physical description where we think of x as a space or time variable. Now using Fourier series we can instead express the function $f(x)$ as a Fourier series, determined by the Fourier coefficients $\{c_m(f)\}$. Describing $f(x)$ through its Fourier coefficients, may be viewed as a frequency-description. In the physical description, we describe the function f in terms of its function values $f(x)$ for different values of x. In the frequency description, we describe f in terms of the Fourier coefficients $c_m(f)$ as a sum $f(x) = \sum_m c_m(f)e^{imx}$.

To describe a given function $f(x)$ we may thus look at the variation of $f(x)$ with x, or the variation of $c_m(f)$ with m.

We have noted that the decay of $c_m(f)$ with m couples to the regularity of $f(x)$: if $f(x)$ is highly regular with many derivatives, then the Fourier coefficients $c_m(f)$ decay quickly with increasing m, and vice versa. If the Fourier coefficients decay quickly, then only a few terms in the Fourier series suffices to represent the function to high accuracy.

252.12 Different Periods

Suppose $f : \mathbb{R} \to \mathbb{C}$ is periodic with period $\frac{2\pi}{\omega}$ with $\omega > 0$. We considered above the case $\omega = 1$, and we now generalize to $\omega > 0$. For example: the functions $\sin(\omega x)$, $\sin(2\omega x)$, $\sin(3\omega x)$..., are periodic with period $\frac{2\pi}{\omega}$.

Defining $g(x) = f(\frac{x}{\omega})$, we have that $g(x)$ is 2π-periodic since $g(x+2\pi) = f(\frac{x+2\pi}{\omega}) = f(\frac{x}{\omega} + \frac{2\pi}{\omega}) = f(\frac{x}{\omega}) = g(x)$, and a Fourier series representation

of $g(x)$:

$$g(x) = \sum_{m=-\infty}^{\infty} c_m(g)e^{imx}, \quad c_m(g) = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(y)e^{-imy}\,dy$$

translates into the following Fourier series representation of $f(\frac{x}{\omega})$:

$$f(\frac{x}{\omega}) = \sum_{m=-\infty}^{\infty} c_m(g)e^{imx}, \quad c_m(g) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\frac{y}{\omega})e^{-imy}\,dy$$

which takes the following form changing variables from $\frac{x}{\omega}$ to x and $\frac{y}{\omega}$ to y:

$$f(x) = \sum_{m=-\infty}^{\infty} c_m(f)e^{im\omega x}, \quad c_m(f) = \frac{\omega}{2\pi}\int_{-\frac{\pi}{\omega}}^{\frac{\pi}{\omega}} f(y)e^{-im\omega y}\,dy. \quad (252.25)$$

252.13 Weierstrass Functions

Consider a series of the form

$$\sum_{m=1}^{\infty} a^{-m}\sin(b^m x), \quad (252.26)$$

where $a > 1$, $b > a$. This type of series was presented by Weierstrass as an example of a Lipschitz continuous function that is not differentiable at any point, see Fig. 252.7, where we plot the corresponding truncated series $\sum_{m=1}^{n} a^{-m}\sin(b^m x)$ with $n = 10$. We see that as n increases the series oscillates increasingly wildly, and gives an irregular "chaotic" impression.

Since $a > 1$ the series (252.26) is absolutely convergent, and defines a function $f(x) = \sum_{m=1}^{\infty} a^{-m}\sin(b^m x)$, but the series

$$\sum_{m=1}^{\infty} a^{-m}b^m\cos(b^m x)$$

obtained by termwise differentiation, does not converge since $\frac{b}{a} > 1$, which indicates that $f(x)$ is nowhere differentiable. The Weierstrass function, or the corresponding truncated series, is an example of a function with a sequence of "microscales" $\frac{2\pi}{b^m}$, $m = 1, 2, ...$ corresponding to the different basis functions $\sin(b^m x)$. The function $f(x)$ thus has the same oscillating nature on all scales and thus has a "fractal" nature. It is believed that phenomena like turbulence also have a fractal nature, which may be useful in attempts to model microscales which are not possible to model numerically.

Choosing $b = 2$ (or b any natural number > 1), gives a series of the form $\sum_{m=1}^{\infty} a^{-m}\sin(2^m x)$, which is an example of a *lacunary Fourier series*, with just very few Fourier coefficients being non-zero. A Weierstrass function with b a natural number is thus a lacunary Fourier series.

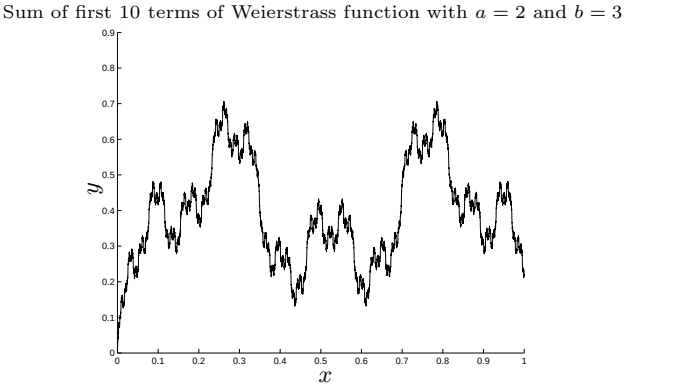


FIGURE 252.7. Plots of a truncated Weierstrass function.

252.14 Solving the Heat Equation Using Fourier Series

We consider the 1d homogeneous heat equation:

$$\begin{aligned} \dot{u}(x,t) - u''(x,t) &= 0 \quad \text{for } 0 < x < \pi,\ t > 0, \\ u(0,t) = u(\pi,t) &= 0 \quad \text{for } t > 0, \\ u(x,0) &= u_0(x) \quad \text{for } 0 < x < \pi, \end{aligned} \tag{252.27}$$

where u_0 is a given initial value. We observe that for $m = 1, 2, ...,$, the function $v(x,t) = v_m(x,t) = e^{-m^2 t}\sin(mx)$ satisfies

$$\dot{v}(x,t) - v''(x,t) = 0 \quad \text{for } 0 < x < \pi, \quad v(0,t) = v(\pi,t) = 0 \quad \text{for } t > 0,$$

and thus any finite linear combination

$$u(x,t) = \sum_{m=1}^{J} b_m e^{-m^2 t}\sin(mx)$$

with coefficients $b_m \in \mathbb{R}$, satisfies (252.27) with corresponding initial data $u_0 = \sum_{m=1}^{J} b_m \sin(mx)$. Each term $e^{-m^2 t}\sin(mx)$ has the form of a product of a function of x only, namely $\sin(mx)$ with *frequency* m, and a function of t only, namely $e^{-m^2 t}$. We see that the factor $e^{-m^2 t}$ decays with increasing t and the rate of decay increases quickly with increasing frequency m. We illustrate this in Fig. 252.8.

More generally, if the initial data u_0 has a convergent Sine series (with $u_0(x)$ extended as an odd function to $[-\pi, \pi]$)

$$u_0(x) = \sum_{m=1}^{\infty} b_m(u_0)\sin(mx), \tag{252.28}$$

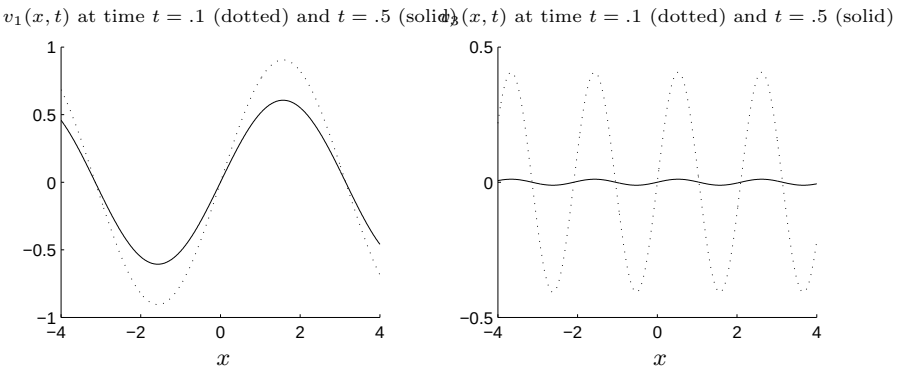


FIGURE 252.8. The solutions $v_j(x,t)$ of the heat equation corresponding to frequencies $j = 1$ and $j = 3$.

with Fourier coefficients

$$b_m(u_0) = \frac{2}{\pi}\int_0^\pi u_0(x)\sin(mx)\,dx, \tag{252.29}$$

then the function defined by

$$u(x,t) = \sum_{m=1}^{\infty} b_m(u_0)e^{-m^2 t}\sin(mx), \tag{252.30}$$

solves the initial value problem (252.27).

252.15 Computing Fourier Coefficients with Quadrature

To compute the Fourier coefficients

$$c_m(f) = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-imx}\,dx, m = 0, \pm 1, \pm 2, ...$$

of a given 2π-periodic function $f : \mathbb{R} \to \mathbb{C}$, we will in general have to use quadrature. Using the quadrature points $x_n = \frac{2\pi n}{N}$, $n = 0, ..., N-1$, with weights $\omega_n = \frac{2\pi}{N}$, corresponding to a left end-point quadrature formula with N uniformly distributed points, we would approximate $c_m(f)$ by

$$c_m(f) = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-imx}\,dx \approx \frac{1}{2\pi}\sum_{n=0}^{N-1} f(x_n)e^{-imx_n}\,\omega_n \equiv \widehat{f}(m),$$

We cannot expect this quadrature formula to be accurate for $m > N$ since then the variation of e^{imx} would not be captured by the quadrature points

$\frac{2\pi n}{N}$. We note that $\widehat{f}(m)$ is periodic with period N: that is $\widehat{f}(m) = \widehat{f}(m+N)$, and it is thus natural to consider $\widehat{f}(m)$ for $m = 0, ..., N-1$, or equivalently with $|m| \leq (N-1)/2$. We call $(N-1)/2$ the *Nyquist cut-off frequency* which corresponds to at least 2 quadrature points on each period for frequencies m with $|m| \leq (N-1)/2$. According to the inversion formula, we could hope, assuming that the Fourier coefficients $c_m(f)$ are small enough for m larger than cut-off, that

$$f(x_n) \approx \sum_{m=0}^{N-1} \widehat{f}(m)e^{imx_n} \quad \text{for } n = 0, ..., N-1, \tag{252.31}$$

which thus would represent an approximate discrete Fourier decomposition for the selected values x_n based on computing the Fourier coefficients $c_m(f)$ by quadrature for $m = 0, ..., N-1$. This leads us directly into the *discrete Fourier transform*, which we now discuss.

252.16 The Discrete Fourier Transform

Suppose $\{f_n\}_{n=0}^{N-1}$ is a set of N given complex numbers. We define a corresponding sequence $\{\widehat{f}_m\}_{m=0}^{N-1}$ by

$$\widehat{f}_m = \frac{1}{N}\sum_{n=0}^{N-1} f_n e^{-2\pi imn/N}, \quad \text{for } m = 0, ..., N-1.$$

We say that the sequence $\{\widehat{f}_m\}_{m=0}^{N-1}$ is the *discrete Fourier transform* of the sequence $\{f_n\}_{n=0}^{N-1}$. In the setting of the previous section we have $f_n = f(\frac{2\pi n}{N})$ and $\widehat{f}_m \approx c_m(f)$.

We find from the definitions

$$\sum_{m=0}^{N-1} \widehat{f}(m)e^{2\pi imn/N} = \sum_{m=0}^{N-1}\frac{1}{N}\sum_{k=0}^{N-1} f_k e^{-2\pi imk/N}e^{2\pi imn/N}$$
$$= \sum_{k=0}^{N-1} f_k \frac{1}{N}\sum_{m=0}^{N-1} e^{2\pi im(n-k)/N}$$

and using that

$$\frac{1}{N}\sum_{m=0}^{N-1} e^{2\pi im(n-k)/N} = \begin{cases} 1 & \text{if } k = n, \\ 0 & \text{else,} \end{cases}$$

we obtain the following inversion formula, to be compared with (252.31),

$$f_n = \sum_{m=0}^{N-1} \widehat{f}(m)e^{2\pi imn/N}, \quad \text{for } n = 0, ..., N-1. \tag{252.32}$$

To compute the discrete Fourier transform of $\{f_n\}_{n=0}^{N-1}$, we would need on the order of N^2 operations (multiplications or additions). If $N = 2^k$ for some natural number k, it is possible to organize the computation of the discrete Fourier transform so that required operations would be of the order N up to a logarithm. The corresponding transform referred to as the *Fast Fourier Transform FFT* developed by Cooley and Tukey in the 1960s, is one of the highlights of applied mathematics of modern time.

Chapter 252 Problems

252.1. Complete the details of the proof of (252.17) and (252.18).

252.2. Prove (252.23).

252.3. Show that the Sine series coefficients for the odd function $f(x) = x^3 - \pi^2 x$ for $-\pi \le x \le \pi$, are given by $b_m(f) = 12\frac{(-1)^m}{m^3}$.

252.4. Show that the Cosine series coefficients for the even function $f(x) = x^4 - 2\pi^2 x^2$ for $-\pi \le x \le \pi$, are given by $a_0 = \frac{14\pi^4}{15}$, $a_m(f) = 48\frac{(-1)^{m+1}}{m^4}$, $m = 1, 2, ...$

252.5. Prove that $\sum_{m=1}^{\infty} \frac{1}{m^4} = \frac{\pi^4}{90}$.

252.6. Define a 2-periodic function $f(x)$ by $f(x) = (x+1)^2$ for $-1 < x < 1$. Expand $f(x)$ in a complex Fourier seriea. Find a 2-periodic solution to the differential equation $2y'' - y' - y = f$.

252.7. Expand the function $\cos x$ as a π-periodic Fourier sine series on the interval $(0, \frac{\pi}{2})$. Use the result to compute $\sum_{n=1}^{\infty} \frac{n^2}{(4n^2-1)^2}$.

252.8. Determine the discrete Fourier transform $\hat{f}_m$ of

$$f_n = \left\{ \begin{array}{ll} 1, & 0 \le n \le k-1, \\ 0, & k \le n \le N-1. \end{array} \right\},$$

and use a Parseval formula to compute

$$\sum_{\mu=1}^{N-1} \frac{1 - \cos\frac{2\pi\mu k}{N}}{1 - \cos\frac{2\pi\mu}{N}}.$$

252.9. Determine the discrete Fourier transform of $f_n = \sin\frac{n\pi}{N}$, $n = 0, \ldots, N-1$.

253
Fourier Transforms

> As the natural ideas of equality developed it was possible to conceive the sublime hope of establishing among us a free government exempt from kings and priests, and to free from this double yoke the long-usurped soil of Europe. I readily became enamoured of this cause, in my opinion the greatest and most beautiful which any nation has ever undertaken. (Fourier 1793, joining a Revolutionary Committee of the French Revolution)

Fourier series concern function $f : \mathbb{R} \to \mathbb{C}$ which are periodic. We now consider functions $f : \mathbb{R} \to \mathbb{C}$ which are non-periodic and the analogous concept is then the *Fourier transform*, which we will study in this chapter. For a given (piecewise Lipschitz continuous) function $f : \mathbb{R} \to \mathbb{C}$ such that $f(x)$ is integrable over $\mathbb{R}$, that is,

$$\int_{\mathbb{R}} |f(x)|\,dx < \infty, \tag{253.1}$$

we define for $\xi \in \mathbb{R}$

$$\widehat{f}(\xi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)e^{-i\xi x}\,dx, \tag{253.2}$$

noting that the integral is absolutely convergent and thus well defined under the assumption (253.1). We say that the function $\widehat{f} : \mathbb{R} \to \mathbb{C}$ defined by (253.2) is the *Fourier transform* of $f(x)$.

We shall now develop a calculus for the Fourier transform which is analogous to that developed for Fourier series in the previous chapter. In par-

ticular we shall prove the inversion formula:

$$f(x) = \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{i\xi x}\, d\xi \quad \text{for } x \in \mathbb{R},$$

under the assumption that $f(x)$ is differentiable on $\mathbb{R}$. As we go along the analogy between Fourier series and Fourier transforms will be uncovered.

We compute the Fourier transform of some basic functions.

EXAMPLE 253.1. If $f(x) = e^{-|x|}$ for $x \in \mathbb{R}$, then

$$\begin{aligned}\widehat{f}(\xi) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-|x|-i\xi x}\, dx \\ &= \frac{1}{2\pi}\int_{-\infty}^{0} e^{x-i\xi x}\, dx + \frac{1}{2\pi}\int_{0}^{\infty} e^{-x-i\xi x}\, dx \\ &= \frac{1}{2\pi}\frac{1}{1-i\xi} + \frac{1}{2\pi}\frac{1}{1+i\xi} = \frac{1}{\pi}\frac{1}{1+\xi^2}.\end{aligned}$$

EXAMPLE 253.2. If $f(x) = e^{-\frac{ax^2}{2}}$ for $x \in \mathbb{R}$, where $a > 0$ is a constant then

$$\widehat{f}(\xi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\frac{ax^2}{2}} e^{-i\xi x}\, dx = \frac{1}{2\pi} e^{-\frac{\xi^2}{2a}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\sqrt{a}x + i\frac{\xi}{\sqrt{a}})^2}\, dx.$$

To evaluate the integral, we shall use Cauchy's theorem for analytic functions as follows: We note that the function $g(z) = e^{-\frac{1}{2}(\sqrt{a}z + i\frac{\xi}{\sqrt{a}})^2}$ is analytic in z, and we may thus shift the line of integration to obtain

$$\widehat{f}(\xi) = \frac{1}{2\pi} e^{-\frac{\xi^2}{2a}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\sqrt{a}x)^2}\, dx,$$

and recalling (224.25) we thus have

$$\widehat{f}(\xi) = \frac{1}{2\pi} e^{-\frac{\xi^2}{2a}} \int_{-\infty}^{\infty} e^{-\frac{ax^2}{2}}\, dx = \frac{1}{\sqrt{2\pi a}} e^{-\frac{\xi^2}{2a}}$$

We note that as a tends to zero, the function $f(x)$ tends to 1 for all x, and $\widehat{f}(\xi)$ tends to $\delta(0)$, the delta function at 0, see Fig. 253.1

EXAMPLE 253.3. Defining $f(x)$ by

$$f(x) = 1 \quad \text{for } -a \le x \le a$$

where $a > 0$, we obtain (see Fig. 253.2

$$\widehat{f}(\xi) = \frac{1}{2\pi}\int_{-a}^{a} e^{-i\xi x}\, dx = \frac{1}{\pi}\frac{\sin(\xi a)}{\xi}.$$

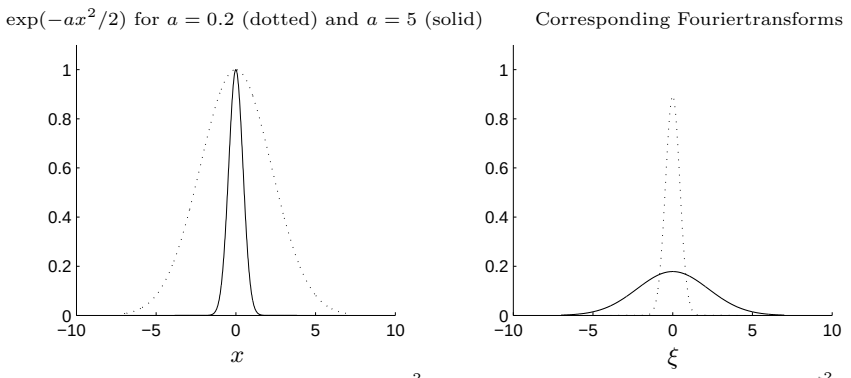


FIGURE 253.1. The functions $e^{-\frac{ax^2}{2}}$ and its Fourier transform $\frac{1}{\sqrt{2\pi a}}e^{-\frac{\xi^2}{2a}}$ for different $a > 0$.

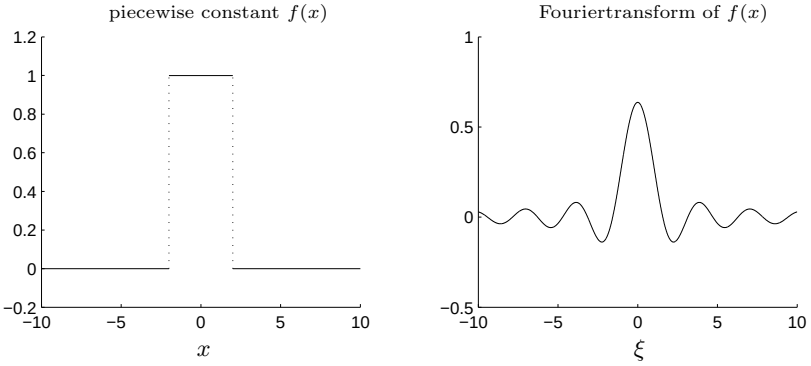


FIGURE 253.2. The above piecewise constant function and its Fourier transform $\frac{1}{\pi}\frac{\sin(\xi a)}{\xi}$.

253.1 Basic Properties of the Fourier Transform

We now present some basic properties of the Fourier transform

$$\widehat{f}(\xi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)e^{-i\xi x}\,dx, \quad \xi \in \mathbb{R},$$

of given functions $f : \mathbb{R} \to \mathbb{R}$ which are integrable over $\mathbb{R}$.

Linearity

The Fourier transform satisfies the following obvious linearity properties:

$$\widehat{(f+g)}(\xi) = \widehat{f}(\xi) + \widehat{g}(\xi), \quad \widehat{(\alpha f)}(\xi) = \alpha\widehat{f}(\xi).$$

where f and g are two functions with Fourier transforms $\widehat{f}$ and $\widehat{g}$, and $\alpha \in \mathbb{C}$.

253.1.1 Scaling

Let $g : \mathbb{R} \to \mathbb{R}$ be integrable and define $f(x) = g(ax)$, where $a > 0$ is a constant. Then, changing variables setting $y = ax$, we have

$$\hat{f}(\xi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} g(ax)e^{-i\frac{\xi}{a}ax}\,dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} g(y)e^{-i\frac{\xi}{a}y}\,\frac{1}{a}dy = \frac{1}{a}\,\hat{g}(\frac{\xi}{a}).$$

We conclude that if $f(x) = g(ax)$, then $\hat{f}(\xi) = \frac{1}{a}\,\hat{g}(\frac{\xi}{a})$.

The Fourier Transform of the Derivative $Df = \frac{df}{dx}$

We now couple the Fourier transform of the derivative $Df = \frac{df}{dx}$ of a function f to the Fourier transform of f. The trick is to integrate by parts:

$$\widehat{Df}(\xi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} Df(x)e^{-i\xi x}\,dx = i\xi\frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)e^{-i\xi x}\,dx = i\xi\widehat{f}(\xi).$$

We summarize in the following theorem:

Theorem 253.1 *If $f : \mathbb{R} \to \mathbb{C}$ is integrable with integrable derivative Df, then for $\xi \in \mathbb{R}$,*

$$\widehat{Df}(\xi) = i\xi\widehat{f}(\xi) \tag{253.3}$$

This is one of the fundamental results of Fourier analysis, and translates the operation of differentiation $D = \frac{d}{dx}$ with respect to x to multiplication of Fourier transforms with $i\xi$ where ξ is the frequency. More generally we have

$$\hat{D^k f}\xi = (i\xi)^k\hat{f}(\xi). \tag{253.4}$$

This opens the way of translating differential equations in the variable x to algebraic equations in the frequency ξ, which may be very useful and illuminating in certain applications.

EXAMPLE 253.4. Consider the differential equation $-D^2u(x)+u(x) = f(x)$ on $\mathbb{R}$ with $f(x)$ a given integrable function and seeking an integrable solution $u(x)$. Since $\widehat{D^2u}(\xi) = (i\xi)^2\widehat{u}(\xi)$, we obtain the algebraic equation

$$(\xi^2 + 1)\widehat{u}(\xi) = \widehat{f}(\xi) \quad \text{for } \xi \neq 0$$

and we can thus express the solution $u(x)$ as a Fourier integral

$$u(x) = \int_{\mathbb{R}} \frac{\widehat{f}(\xi)}{\xi^2+1}e^{i\xi x}\,d\xi,$$

in terms of the Fourier transform $\widehat{f}(\xi)$ of the data $f(x)$.

253.2 The Fourier Transform $\widehat{f}(\xi)$ Tends to 0 as $|\xi| \to \infty$

As a direct consequence of the preceding result, we conclude that the Fourier transform $\widehat{f}(\xi)$ of a differentiable function $f(x)$ with integrable derivative $Df(x)$, tends to zero as $|\xi|$ tends to infinity. This is simply because

$$|\widehat{f}(\xi)| = \frac{1}{|\xi|}\widehat{Df}(\xi) \leq \frac{1}{2\pi|\xi|}\int_{-\infty}^{\infty} |Df|\,dx \to 0 \quad \text{as } |\xi| \to \infty.$$

This result can be extended to the case of $f(x)$ being integrable, as in the corresponding case of Fourier series.

253.3 Convolution

Given two functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, we define a new function $f * g : \mathbb{R} \to \mathbb{R}$ by

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-y)g(y)\,dy.$$

We say that $f * g$ is the *convolution* of f and g. We shall prove that

$$\widehat{f * g} = 2\pi \widehat{f}\,\widehat{g}$$

By direct computation, changing order of integration and using the change of variable $t = x - y$, we have

$$\begin{aligned}
\widehat{f * g}(\xi) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} (f * g)(x)e^{-i\xi x}\,dx = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x-y)g(y)\,dy\,e^{-i\xi x}\,dx \\
&= \int_{-\infty}^{\infty} g(y)e^{-i\xi y}\Big(\frac{1}{2\pi}\int_{-\infty}^{\infty} f(x-y)\,e^{-i\xi(x-y)}\,dx\Big)dy \\
&= \int_{-\infty}^{\infty} g(y)e^{-i\xi y}\Big(\frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\,e^{-i\xi t}\,dt\Big)dy \\
&= \widehat{f}(\xi)\int_{-\infty}^{\infty} g(y)e^{-i\xi y}\,dy = 2\pi\widehat{f}(\xi)\widehat{g}(\xi).
\end{aligned}$$

We summarize:

Theorem 253.2 *We have* $\widehat{f * g}(\xi) = 2\pi\widehat{f}(\xi)\widehat{g}(\xi)$ *for* $\xi \in \mathbb{R}$.

253.4 The Inversion Formula

We shall now prove that if $f(x)$ is differentiable, then for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} f_n(x) = f(x),$$

where

$$f_n(x) = \int_{-n}^{n} \widehat{f}(\xi) e^{i\xi x}\, d\xi$$

and thus for all $x \in \mathbb{R}$, the function $f(x)$ can be represented as a convergent Fourier integral:

$$f(x) = \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{i\xi x}\, d\xi.$$

We have

$$\begin{aligned} f_n(x) = \int_{-n}^{n} \widehat{f}(\xi) e^{i\xi x}\, d\xi &= \frac{1}{2\pi} \int_{-\infty}^{\infty} f(y) \int_{-n}^{n} e^{i\xi(x-y)}\, d\xi dy \\ &= \int_{-\infty}^{\infty} f(y) D_n(x-y)\, dy, \end{aligned} \tag{253.5}$$

where, setting $\theta = x - y$,

$$D_n(\theta) = \frac{1}{2\pi} \int_{-n}^{n} e^{i\xi\theta}\, d\xi = \frac{1}{\pi} \frac{\sin(n\theta)}{\theta}$$

is the *Dirichlet kernel* for the Fourier transform. Using the convolution notation we can write (253.5) in the compact form

$$f_n(x) = f * D_n(x).$$

With experience from the Dirichlet kernel for Fourier series, we expect D_n to be an approximate identity. Looking at a plot of $D_n(\theta)$ in Fig. 253.3, we see that $D_n(\theta)$ oscillates with a peak at $\theta = 0$, which gets sharper with increasing n. One can prove that, see Problem 253.5,

$$\int_{-\infty}^{\infty} D_n(\theta)\, d\theta = 1 \tag{253.6}$$

and we can thus write

$$\begin{aligned} f(x) - f_n(x) &= \int_{-\infty}^{\infty} (f(x) - f(y)) D_n(x-y)\, dy \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} g(y) \sin((n(x-y))\, dy \end{aligned}$$

where

$$g(y) = \frac{f(x) - f(y)}{x - y}.$$

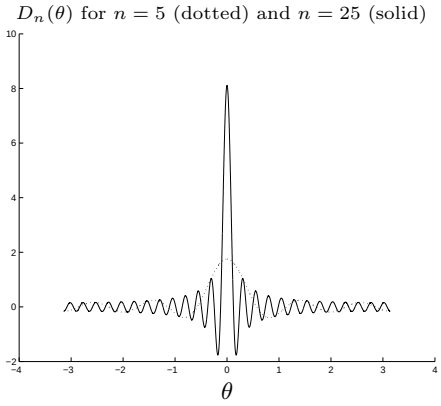


FIGURE 253.3. The function $D_n(\theta)$.

Now if $f(x)$ is differentiable with integrable derivative, it follows with an argument similar to that used in the case of Fourier series (integrating by parts), that

$$f(x) - f_n(x) \to 0$$

as $n \to \infty$, which proves the following basic theorem.

Theorem 253.3 *If $f(x)$ is differentiable, then $f(x)$ is given by a convergent Fourier integral:*

$$f(x) = \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{i\xi x}\, d\xi \quad \textit{for } x \in \mathbb{R}.$$

253.5 Parseval's Formula

Parseval's formula takes the following form for the Fourier transform (for a proof see Problem 253.3):

Theorem 253.4 *If $f(x)$ has a convergent Fourier series representation, then*

$$\int_{-\infty}^{\infty} |f(x)|^2\, dx = 2\pi \int_{-\infty}^{\infty} |\widehat{f}(\xi)|^2\, d\xi.$$

253.6 Solving the Heat Equation Using the Fourier Transform

We consider the 1d homogeneous heat equation on $\mathbb{R}$

$$\begin{aligned} \dot{u}(x,t) - u''(x,t) &= 0 \quad \text{for } x \in \mathbb{R},\ t > 0, \\ u(x,0) &= u_0(x) \quad \text{for } x \in \mathbb{R}, \end{aligned} \tag{253.7}$$

where the initial value $u_0(x)$ is integrable over $\mathbb{R}$ and we seek a solution $u(x,t)$ which is integrable over $\mathbb{R}$ for all $t > 0$. Taking Fourier transforms with respect to x, we are led to the following initial value problem for each $\xi \in \mathbb{R}$

$$\frac{d}{dt}\widehat{u}(\xi,t) + \xi^2\widehat{u}(\xi,t) = 0 \quad \text{for } t > 0, \quad \widehat{u}(\xi,0) = \widehat{u}_0(\xi)$$

with solution

$$\widehat{u}(\xi,t) = e^{-t\xi^2}\widehat{u}_0(\xi).$$

We thus obtain the following solution formula

$$u(x,t) = \frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} e^{-\frac{(x-y)^2}{4t}}u_0(y)\,dy.$$

Here we used that $\widehat{u}$ is the product of the Fourier transforms $e^{-t\xi^2}$ of $\sqrt{\pi/t}e^{-\frac{x^2}{4t}}$ and $\widehat{u}_0$ of u_0, the inverse transform of which thus is the convolution of $\frac{1}{4\pi t}e^{-\frac{x^2}{4t}}$ and u_0.

253.7 Fourier Series and Fourier Transforms

Suppose $f : \mathbb{R} \to \mathbb{C}$ is periodic with period $\frac{2\pi}{\omega}$ with $\omega > 0$ and has the Fourier series representation

$$f(x) = \sum_{m=-\infty}^{\infty} c_m(f)e^{im\omega x}, \quad c_m(f) = \frac{\omega}{2\pi}\int_{-\frac{\pi}{\omega}}^{\frac{\pi}{\omega}} f(y)e^{-im\omega y}\,dy,$$

which we write in the form

$$f(x) = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi}\Big(\int_{-\frac{\pi}{\omega}}^{\frac{\pi}{\omega}} f(y)e^{-im\omega y}\,dy\Big)\,e^{im\omega x}\omega. \tag{253.8}$$

We now compare with a Fourier transform representation of a non-periodic function $f : \mathbb{R} \to \mathbb{C}$ according to the previous section:

$$f(x) = \int_{-\infty}^{\infty} \frac{1}{2\pi}\Big(\int_{-\infty}^{\infty} f(y)e^{-i\xi y}\,dy\Big)\,e^{i\xi x}\,d\xi. \tag{253.9}$$

We formally obtain (253.9) from (253.8) by replacing $m\omega$ by ξ and ω by $d\xi$ viewing the sum over m as a Riemann sum and letting ω tend to 0.

Note the normalization used in the definition of the Fourier transform $\widehat{f}(\xi)$ with the factor $\frac{1}{2\pi}$, and the factor $\frac{\omega}{2\pi}$ in the definition of the Fourier coefficients $c_m(f)$ of a function f with period $\frac{2\pi}{\omega}$.

253.8 The sampling theorem

Let $f : \mathbb{R} \to \mathbb{C}$ be a given function with Fourier transform $\widehat{f}(\xi)$, and suppose that $\hat{f}(\xi) = 0$ for $|\xi| \geq \pi$. By Fourier's inversion formula, we have

$$f(x) = \int_{-\pi}^{\pi} \widehat{f}(\xi) e^{ix\xi}\, d\xi.$$

We now expand $\widehat{f}(\xi)$ in a Fourier series:

$$\widehat{f}(\xi) = \sum_{m=-\infty}^{\infty} c_m(\widehat{f}) e^{im\xi}$$

with Fourier coefficients

$$c_m(\widehat{f}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{f}(\eta) e^{-im\eta}\, d\eta.$$

Using that $\widehat{f}(\eta) = 0$ for $|\eta| \geq \pi$, we can write

$$c_m(\widehat{f}) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\eta) e^{-im\eta}\, d\eta = \frac{1}{2\pi} f(-m)$$

where we used Fourier's inversion formula. We thus obtain the representation formula:

$$\begin{aligned} f(x) &= \int_{-\pi}^{\pi} \frac{1}{2\pi} \sum_{m=-\infty}^{\infty} f(-m) e^{im\xi} e^{ix\xi}\, d\xi \\ &= \frac{1}{2\pi} \sum_{m=-\infty}^{\infty} f(-m) \int_{-\pi}^{\pi} e^{im\xi} e^{ix\xi}\, d\xi \\ &= \sum_{m=-\infty}^{\infty} f(-m) \frac{\sin(x+m)}{\pi(x+m)} = \sum_{m=-\infty}^{\infty} f(m) \frac{\sin(x-m)}{\pi(x-m)} \end{aligned}$$

which gives a representation of $f(x)$ for any value of x in terms of the values $\{f(m)\}$ with m integer. We have now prove the famous:

Theorem 253.5 (Sampling theorem) *If $f : \mathbb{R} \to \mathbb{C}$ has a Fourier transform $\widehat{f}(\xi)$ such that $\hat{f}(\xi) = 0$ for $|\xi| \geq \pi$, then*

$$f(x) = \sum_{m=-\infty}^{\infty} f(m) \frac{\sin(x-m)}{\pi(x-m)}$$

We conclude that *sampling* the values $f(m)$ of the function $f(x)$ for the integer values $m = 0, \pm 1, \pm 2, ...$, gives information of all the values of $f(x)$ for any $x \in \mathbb{R}$, under the assumption that $\hat{f}(\xi) = 0$ for $|\xi| \geq \pi$.

EXAMPLE 253.5. The Sampling theorem takes the following form for a function $f(x)$ such that $\hat{f}(\xi) = 0$ for $|\xi| \geq a\pi$, where $a > 0$ is a constant:

$$f(x) = \sum_{m=-\infty}^{\infty} f(\frac{m}{a})\frac{\sin(ax-m)}{\pi(ax-m)}.$$

This follows by applying the Sampling theorem to $g(x) = f(\frac{x}{a})$, recalling that $\widehat{g}(\xi) = a\widehat{f}(a\xi)$ and noting that $\widehat{g}(\xi) = 0$ if $|\xi| \geq \pi$ since $\hat{f}(\xi) = 0$ for $|\xi| \geq a\pi$. We see that the larger the factor a gets, the closer the sampling points $\frac{m}{a}$ will be distributed. Of course this couples to the Nyquist cut-off frequency.

253.9 The Laplace Transform

We give a brief account of the *Laplace transform*, which is closely related to the Fourier transform. The Laplace transform is useful in solving certain constant-coefficient linear initial value problems analytically with classical applications in e.g. control theory.

For a given function $f : [0, \infty) \to \mathbb{R}$, we define the *Laplace transform* $Lf : [0, \infty)$ by

$$Lf(s) = \int_0^\infty e^{-st} f(t)\, dt \quad \text{for } s \in [0, \infty).$$

We denote here the independent variable by t indicating typical applications with t representing time.

EXAMPLE 253.6. If $f(t) = e^{-at}$, then $Lf(s) = \frac{1}{s+a}$.

EXAMPLE 253.7. If $f(t) = \frac{t^n}{n!}$ then $Lf(s) = \frac{1}{s^{n+1}}$. This follows by repeated integration by parts.

EXAMPLE 253.8. If $f(t) = \sin(mt)$ then $Lf(s) = \frac{m}{m^2+s^2}$. If $f(t) = \cos(mt)$ then $Lf(s) = \frac{s}{m^2+s^2}$.

We note the following connection between the Laplace transform of $Df = f'$ and f:

$$Lf'(s) = sLf(s) - f(0) \tag{253.10}$$

which follows by integration by parts.

Laplace Transforms and Constant-Coefficient Linear Initial Value Problems

The typical application goes as follows: Consider the initial value problem $u'(t) + u(t) = f(t)$ for $t > 0$ with $u(0) = 0$. Taking Laplace transforms of

both sides we get

$$sLu(s) + Lu(s) = Lf(s), \quad \text{or } Lu(s) = \frac{Lf(s)}{s+1}$$

For example, if $f(t) = 1$, then $Lf(s) = \frac{1}{s}$ and thus $Lu(s) = \frac{1}{s(s+1)} = \frac{1}{s} - \frac{1}{s+1}$ and we conclude that $u(s) = 1 - e^{-t}$. Having a catalogue of Laplace transforms we may expect to be able to solve constant-coefficient linear initial value problems.

253.10 Wavelets and the Haar Basis

We give a short introduction to wavelets in the simplest setting of 1d piecewise constant approximation using the *Haar basis*, which combines the features of orthogonality and local support. We thus consider functions defined on the unit interval $[0, 1]$ and we let $0 = x_0 < x_1 < ... < x_N = 1$ be a uniform subdivision with $x_j = jh_n$, $h_n = 2^{-n}$ and $N = 2^n$ for some natural number n. A natural orthogonal basis for the space V_n of piecewise constant functions on the subdivision $0 = x_0 < x_1 < ... < x_N = 1$ consists of the set of functions $\{\varphi_{n,k}\}_{k=0}^N$, where $\varphi_{n,k}(x) = 1$ for $x \in I_{n,k} = (kh_n, (k+1)h_n)$ and $\varphi_{n,k}(x) = 0$ else, that is, each basis function $\varphi_{n,k}(x)$ is equal to 1 on the subinterval $I_{n,k}$ and vanishes elsewhere. We can express these functions through scaling and translation of one single function in the form

$$\varphi_{n,k} = \varphi(2^n x - k) \quad \text{for }, \quad k = 0, ..., N-1,$$

where $\varphi(x) = 1$ for $x \in (0, 1)$, and $\varphi(x) = 0$ else. We note that V_{n-1} is a subspace of V_n since the space V_n is built on a finer subdivision than V_{n-1}.

We shall now present a different orthogonal basis for V_n which displays the "difference" between V_n and V_{n-1}, and which carries useful information on the various scales in V_n. More precisely, we shall express each $u \in V_n$ in the form $u = v + w$ with $v \in V_{n-1}$ and $w \in W_{n-1}$, where W_{n-1} is spanned by the functions $\psi_{n-1,k} = \psi(2^{n-1}x - k)$ for $k = 1, ..., 2^{n-1}$, expressed through scaling and translation of the single function $\psi(x)$ given by

$$\psi(x) = \begin{cases} 1 & \text{for } 0 < x < \frac{1}{2}, \\ -1 & \text{for } \frac{1}{2} < x < 1, \end{cases}$$

and $\psi(x) = 0$ else. We note that $(v, w) = \int_0^1 v(x)w(x)\,dx = 0$ if $v \in V_{n-1}$ and $w \in W_{n-1}$. Further, the two functions $\varphi_{n-1,k}$ and $\psi_{n-1,k}$ obviously span the two-dimensional space of functions on the interval $I_{n,k}$ which are piecewise constant on the two subintervals $kh_n < x < kh_n + h_{n+1}$ and $kh_n + h_{n+1} < x < kh_n + h_n$ of $I_{n,k}$. We thus have the following orthogonal decomposition

$$V_n = V_{n-1} \oplus W_{n-1},$$

stating that each function $u \in V_n$ can be expressed in the form $u = v + w$ with $v \in V_{n-1}$, $w \in W_{n-1}$ and $(v, w) = 0$, see Fig. 253.4.

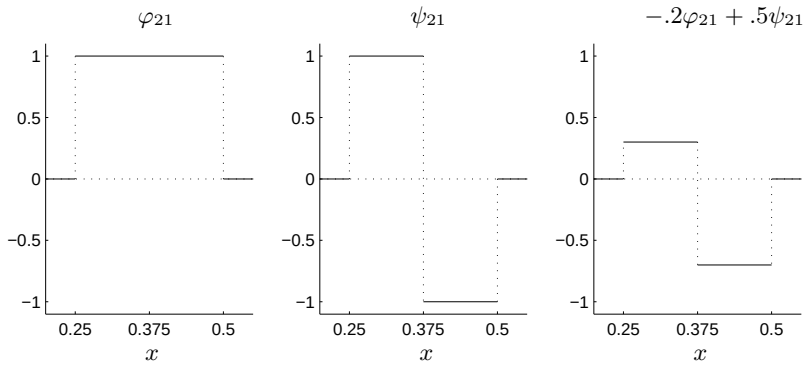


FIGURE 253.4. Illustration of the orthogonal decomposition $V_n = V_{n-1} \oplus W_{n-1}$.

We can thus express V_n as an orthogonal sum:

$$V_n = V_0 \oplus W_0 \oplus W_1 \oplus ... \oplus W_{n-1}$$

where each space $\oplus W_j$ measures variations on the scale 2^{-j}. The corresponding basis functions comprise the so-called *Haar basis* for V_n:

$$\{\varphi_0 = \varphi, \psi_{1,1} = \psi, \psi_{2,1}, \psi_{2,2}, \psi_{3,1}, \psi_{3,2}, \psi_{3,3}, \psi_{3,4}, ..., \psi_{n-1,1}, ..., \psi_{n-1,2^{n-1}}\}$$

combining orthogonality and local support.

Chapter 253 Problems

253.1. Solve using Fourier series the differential equation $-D^2u(x) = f(x)$ with $f(x)$ a given 2π-periodic function with zero mean value and we seek a 2π-periodic solution $u(x)$ with zero mean value.

253.2. Model the following electrical circuits: (i) resistor 1 and inductor in series over applied voltage (ii) resistor 1 and capacitor in series over applied voltage (iii) resistor 2 coupled in series with resistor 1 and inductor in parallel over applied voltage. Output voltage drop over resistor 1. Solve using Fourier series. Show that (i) and (ii) correspond to low-pass filters and (iii) to high-pass filter.

253.3. Prove Parseval's formula for the Fourier transform. Hint: Set $\widehat{g}(\xi) = \overline{\widehat{f}}(\xi)$, which is the same as setting $g(-x) = \overline{f}(x)$, and integrate over ξ:

$$\int_{-\infty}^{\infty} f(x)\overline{f}(x)\,dx = f * g(0) = \int_{-\infty}^{\infty} \widehat{f * g}(\xi)\,d\xi = 2\pi \int_{-\infty}^{\infty} |\widehat{f}(\xi)|^2\,d\xi,$$

and use Theorem 253.2.

253.4. Prove that for $a \in \mathbb{R}$, we have (i) $\widehat{g}(\xi) = e^{-ia\xi}\widehat{f}(\xi)$ if $g(x) = f(x-a)$, (ii) $\widehat{g}(\xi) = \widehat{f}(\xi - a)$ if $g(x) = e^{iax}f(x)$.

253.5. Prove (253.6).

253.6. Compute the Fourier transform of the functions a) $\frac{x}{(x^2+a^2)^2}$, b) $\frac{1}{(x^2+a^2)^2}$, c) $\frac{x}{(x^2+1)(x^2+2x+5)}$, d) $e^{-a|x|}\sin xt \quad (a>0, b>0)$.

253.7. The function $f(x)$ has the Fourier transform $\frac{1-i\xi}{1+i\xi}\,\frac{\sin\xi}{\xi}$. Compute $\int_{-\infty}^{\infty}|f(x)|^2dx$.

253.8. Compute $\int_{-\infty}^{\infty}\frac{\sin x}{x(x^2+1)}dx$ using the Fourier transform.

253.9. A function $f(x)$ has the Fourier transform $\frac{1}{|\xi|^3+1}$. Compute $\int_{-\infty}^{\infty}|f * f'|^2\,dx$.

253.10. Compute the Fourier transform of the function $f(x) = \int_0^2 \frac{\sqrt{\xi}}{1+\xi}e^{i\xi x}d\xi$. Then compute a) $\int_{-\infty}^{\infty} f(x)\cos x dx$, b) $\int_{-\infty}^{\infty}|f(x)|^2dx$.

253.11. Determine the solution $f(t)$, $t>0$, to the initial value problem

$$f''(t) - f'(t) + f(t) + 6\int_0^t f(\tau)d\tau = 2e^t \quad \text{for } t>0,$$

with initial values $f(0)=1, f'(0)=0$.

254
Analytic Functions Tool Bag

254.1 Differentiability and analyticity

A function $f : \Omega \to \mathbb{C}$ is *differentiable* at $z_0 \in \Omega$ with derivative $f'(z_0) \in \mathbb{C}$, if for z close to z_0, we have

$$|f(z) - f(z_0) - f'(z_0)(z - z_0)| \leq K_f(z_0)|z - z_0|^2,$$

where $K_f(z_0)$ is a non-negative real constant depending on f and z_0.

A function $f : \Omega \to \mathbb{C}$ is *analytic* in the open domain Ω of the complex plane if $f(z)$ is differentiable at all $z_0 \in \Omega$ with derivative $f'(z_0)$. If $f : \Omega \to \mathbb{C}$ is analytic, then also $f' : \Omega \to \mathbb{C}$ is analytic with derivative $f'' : \Omega \to \mathbb{C}$, which is also analytic, and so on. An analytic function $f : \Omega \to \mathbb{C}$ thus has derivatives of all orders $f^{(n)} : \Omega \to \mathbb{C}$, $n = 1, 2, ...$, which are all analytic.

The usual rules for differentiation of sums, products and quotients valid for functions $f : \mathbb{R} \to \mathbb{R}$ extend to functions $f : \mathbb{C} \to \mathbb{C}$.

The function $f(z) = z^n$ is analytic in $\mathbb{C}$ for $n = 1, 2, ...$

The function $f(z) = z^{-n}$ is analytic for $z \neq 0$ if $n = 1, 2, ...$

254.2 The Cauchy-Riemann Equations

If $f(z) = u(x, y) + iv(x, y)$ is analytic in the open domain Ω of the complex plane, then the real and imaginary parts $u(x, y)$ and $v(x, y)$ satisfy the

Cauchy-Riemann equations in Ω:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x},$$

or in polar coordinates $z = re^{i\theta}$:

$$\frac{\partial u}{\partial r} = \frac{1}{r}\frac{\partial v}{\partial \theta}, \quad \frac{\partial v}{\partial r} = -\frac{1}{r}\frac{\partial u}{\partial \theta}.$$

254.3 The Real and Imaginary Parts of an Analytic Function

If $f : \Omega \to \mathbb{C}$ is analytic, where Ω is an open domain of the complex plane $\mathbb{C}$, then the real part $u(x, y) = \text{Re } f(z)$ and the imaginary part $v(x, y) = \text{Im } f(z)$ are harmonic in Ω.

254.4 Conjugate Harmonic Functions

If $u(x, y)$ is harmonic in a simply connected domain Ω in $\mathbb{R}^2$, then there exists a harmonic function $v(x, y)$, uniquely determined up to a constant, such that $f(z) = u(x, y) + iv(x, y)$ is analytic in Ω. The function $v(x, y)$ is *conjugate* to $u(x, y)$.

254.5 Curves in the Complex Plane

A set $\Gamma = \text{Range of } \gamma = \{\gamma(t) : t \in I\}$ in an open domain Ω in the complex plane $\mathbb{C}$ parameterized by a Lipschitz continuous mapping $\gamma : I \to \Omega$, where $I = [a, b]$ is an interval of $\mathbb{R}$, is said to be a curve. The unit circle is a curve parameterized by the function $\gamma(t) = \exp(it)$ with $0 \leq t < 2\pi$. Γ is a *differentiable curve* if the corresponding parametrization $\gamma : I \to \mathbb{C}$ is differentiable on I, that is, decomposing $\gamma(t) = x(t) + iy(t)$ into real and imaginary parts,that is, if $x : I \to \mathbb{R}$ if $x(t)$ and $y(t)$ are differentiable on I.

A curve Γ with parametrization $\gamma : [a, b] \to \mathbb{C}$ is said to *closed and simple* if $\gamma(a) = \gamma(b)$ and $\gamma(s) = \gamma(t)$ only if $a = b$. A domain Ω in $\mathbb{C}$ which is bounded by a simple closed curve, is *simply connected*. A simply connected domain does not have any "holes".

254.6 An Analytic Function Defines a Conformal Mapping

An analytic function $f : \Omega \to \mathbb{C}$, where Ω is an open domain in $\mathbb{C}$, is conformal in Ω in the sense that angles are preserved under the mapping $w = f(z)$.

254.7 Complex Integrals

We define

$$\int_\Gamma f(z)\,dz = \int_a^b \big(u(x(t),y(t)) + iv(x(t),y(t))\big)(\dot{x}(t) + i\dot{y}(t))\,dt,$$

where Ω is an open domain in the complex plane, Γ is a differentiable curve in $\mathbb{C}$ parameterized by $\gamma = (x,y) : [a,b] \to \mathbb{C}$, and $f = u + iv : \Gamma \to \mathbb{C}$ is Lipschitz continuous. Formally we have $dz = dx + idy = \dot{x}dt + i\dot{y}dt = (\dot{x} + i\dot{y})\,dt$.

254.8 Cauchy's Theorem

If $f(z)$ is analytic in Ω and Γ is a simple closed curve in Ω enclosing a domain contained in Ω, then

$$\int_\Gamma f(z)\,dz = 0.$$

254.9 Cauchy's Representation Formula

If $f(z)$ is analytic in an open domain Ω, and Γ is a simple closed curve in Ω oriented counter-clockwise and enclosing the open domain Ω_Γ contained in Ω, then for $z_0 \in \Omega_\Gamma$,

$$f(z_0) = \frac{1}{2\pi i}\int_\Gamma \frac{f(z)}{z - z_0}\,dz,$$

and for $n = 1, 2, ...,$

$$f^{(n)}(z_0) = \frac{n!}{2\pi i}\int_\Gamma \frac{f(z)}{(z - z_0)^{n+1}}\,dz.$$

254.10 Taylor's formula

If $f(z)$ is analytic in a neighborhood Ω of a $z_0 \in \mathbb{C}$, then

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + ... + \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n + R_n(z),$$

where

$$R_n(z) = \frac{(z - z_0)^{n+1}}{2\pi i} \int_\Gamma \frac{f(\zeta)}{(\zeta - z_0)^{n+1}(\zeta - z)}\, d\zeta.$$

254.11 The Residue Theorem

If $f(z)$ is analytic in a simply connected open domain Ω, except at finitely many isolated points z_1, z_1,...,z_n in Ω, where $f(z)$ has simple or multiple poles, and Γ is a simple closed curve in Ω surrounding all the z_m counterclockwise, then

$$\int_\Gamma f(z)\, dz = \sum_{m=1}^{n} 2\pi i \mathrm{Res}\, f(z_m).$$

255
Fourier Analysis Tool Bag

255.1 Properties of Fourier Coefficients

The Fourier coefficients $c_m(f)$ of a given 2π-periodic Lipschitz function$f : \mathbb{R} \to \mathbb{C}$ are defined by

$$c_m(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-imx}\,dx \quad m = 0, \pm 1, \pm 2, ...,$$

and satisfy

$$c_m(f+g) = c_m(f) + c_m(g), \quad c_m(\alpha f) = \alpha c_m(f), \quad \text{for } \alpha \in \mathbb{C},$$
$$c_m(D^k f) = (im)^k c_m(f) \quad \text{for } k = 0, 1, 2, ...,$$

If $f : [-\pi, \pi]$ is Lipschitz continuous, then $c_m(f) \to 0$ as $|m| \to \infty$ (Riemann-Lebesgue' Lemma).

255.2 Convolution

Defining for 2π-periodic functions $f(x)$ and $g(x)$, the convolution $f * g$ by

$$(f * g)(x) = \int_{-\pi}^{\pi} f(x-y)g(y)\,dy \quad x \in \mathbb{R},$$

we have

$$c_m(f * g) = 2\pi\, c_m(f)c_m(g).$$

255.3 Fourier Series Representation

If $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic with piecewise Lipschitz continuous derivative, then $f(x)$ may be represented by a convergent Fourier series:

$$f(x) = \sum_{m=-\infty}^{\infty} c_m(f) e^{imx} \quad \text{for } x \in \mathbb{R}.$$

255.4 Parseval's Formula

If $f(x)$ has a convergent Fourier series representation, then

$$\int_{-\pi}^{\pi} |f(x)|^2 \, dx = 2\pi \sum_{m=-\infty}^{\infty} |c_m(f)|^2.$$

255.5 Discrete Fourier Transforms

If $\{f_n\}_{n=0}^{N-1}$ is a sequence of N given complex numbers, then we may define a corresponding sequence $\{\widehat{f}_m\}_{m=0}^{N-1}$ by

$$\widehat{f}_m = \frac{1}{N} \sum_{n=0}^{N-1} f_n e^{-2\pi imn/N}, \quad \text{for } m = 0, ..., N-1,$$

and we say that the sequence $\{\widehat{f}_m\}_{m=0}^{N-1}$ is the *discrete Fourier transform* of the sequence $\{f_n\}_{n=0}^{N-1}$. We have the following inversion formula:

$$f_n = \sum_{m=0}^{N-1} \widehat{f}(m) e^{2\pi imn/N}, \quad \text{for } n = 0, ..., N-1.$$

255.6 Fourier Transforms

For $f : \mathbb{R} \to \mathbb{C}$ piecewise Lipschitz continuous and integrable over $\mathbb{R}$, we define the Fourier transform of $f(x)$ for $\xi \in \mathbb{R}$ by

$$\widehat{f}(\xi) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x) e^{-i\xi x} \, dx.$$

We have the inversion formula:

$$f(x) = \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{i\xi x} \, d\xi \quad \text{for } x \in \mathbb{R},$$

under the assumption that $f(x)$ is differentiable on $\mathbb{R}$ with integrable derivative.
If $f(x) = e^{-|x|}$ for $x \in \mathbb{R}$, then

$$\widehat{f}(\xi) = \frac{1}{\pi}\frac{1}{1+\xi^2}.$$

If $f(x) = e^{-\frac{ax^2}{2}}$ for $x \in \mathbb{R}$, where $a > 0$ is a constant, then

$$\widehat{f}(\xi) = \frac{1}{2\sqrt{a}}e^{-\frac{\xi^2}{2a}}.$$

If $f(x) = 1 \quad \text{for } -a \le x \le a$ and $f(x) = 0$ else, where $a > 0$, then

$$\widehat{f}(\xi) = \frac{\sin(\xi a)}{\xi}.$$

255.7 Properties of Fourier Transforms

If f and g are two functions with Fourier transforms $\widehat{f}$ and $\widehat{g}$, and $\alpha \in \mathbb{C}$, then

$$\widehat{(f+g)}(\xi) = \widehat{f}(\xi) + \widehat{g}(\xi), \quad \widehat{(\alpha f)}(\xi) = \alpha\widehat{f}(\xi).$$

If $g : \mathbb{R} \to \mathbb{R}$ is integrable and $f(x) = g(ax)$, then $\hat{f}(\xi) = \frac{1}{a}\hat{g}(\frac{\xi}{a})$.
If $f : \mathbb{R} \to \mathbb{C}$ is integrable with integrable derivative, then

$$\widehat{Df}(\xi) = i\xi\widehat{f}(\xi).$$

Defining for two integrable functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, the convolution $f * g$ by

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-y)g(y)\,dy,$$

we have

$$\widehat{f * g}(\xi) = 2\pi\widehat{f}(\xi)\widehat{g}(\xi) \quad \text{for } \xi \in \mathbb{R}.$$

Parseval's formula:

$$\int_{-\infty}^{\infty} |f(x)|^2\,dx = 2\pi\int_{-\infty}^{\infty} |\widehat{f}(\xi)|^2\,d\xi.$$

255.8 The Sampling Theorem

If $f : \mathbb{R} \to \mathbb{C}$ has a Fourier transform $\widehat{f}(\xi)$ such that $\hat{f}(\xi) = 0$ for $|\xi| \ge a\pi$, where $a > 0$ is a constant, then

$$f(x) = \sum_{m=-\infty}^{\infty} f(\frac{m}{a})\frac{\sin(ax-m)}{\pi(ax-m)}.$$

Part XVI

Brain Storm

1958

256

Lorenz and the Essence of Chaos*

> I am convinced that chaos, along with its many associated concepts - strange attractors, basin boundaries, period-doubling bifurcations and the like - can readily be understood and relished by readers who have no special mathematical or other scientific background...
> (E. Lorenz, in Foreword to *The Essence of Chaos*)

256.1 Introduction

On December 29, 1972, the meteorologist Edward Lorenz presented in a session on the Global Atmospheric Research Program at the 139th meeting of the American Association for the Advancement of Science in Washington D.C., a talk with the title *Predictability: Does the Flap of a Butterfly's Wings in Brazil Set off a Tornado in Texas?* The talk by Lorenz with its "Butterfly effect" rocketed to fame a decade later during the development of "Chaos Theory" that became a fashion in mathematics and physics during the 80s, with the pretention of explaining a variety of phenomena from turbulent fluid flow to collapsing stock markets sharing qualities of *unpredictability*. A decade earlier, "Catastrophe Theory" played a similar role, while today very few remember this intriguing subject. Of course, unpredictability or "chaos" is a phenomenon that has long been familiar to mankind. The word "chaos" comes from early Greek cosmology and signifies the complete lack of order of the Universe before the creation of Gaea and Eros (Earth and Desire).

Lorenz' question is connected to the obvious difficulty of making reasonably reliable predictions of the daily weather over longer time than a week. A weather forecast is made by numerically solving an IVP modeling the evolution of the atmosphere, including variables such as temperature, wind speed and pressure. There are many sources of errors in a weather forecast made this way: errors in the initial value, modeling errors and numerical errors, and it seems that these errors are magnified at a rate that limits the predictions, depending on the scale from a few hours in very local models to weeks in global circulation models.

Lorenz' Butterfly analogy indicates that in certain dynamical systems, very small causes may have large effects after some time. We have already met such a problem in the form of a pendulum being released starting from the unstable top position: depending on the initial perturbation the position of the pendulum will be vastly different after some time (one side or the other). In meteorology, this corresponds to a situation where the weather-man can't say if a certain low pressure will take this way or that way, and thus can't be sure if it will rain in Göteborg tomorrow or not. In his book, Lorenz gives other examples of unstable systems such as a pinball machine, where very small changes in the action of the player can change the outcome of the game completely. Of course there are many other examples from real life of "small" causes having large effects, from soccer games to the assassination of Archduke Francis Ferdinand by the Serb nationalist Gavrilo Princip in Sarajevo on June 28 1914, initiating the First World War.

256.2 The Lorenz System

Lorenz formulated an IVP of the form $\dot{u} = f(u)$ with $f : \mathbb{R}^3 \to \mathbb{R}^3$ given by

$$f(u) = (-10u_1 + 10u_2, 28u_1 - u_2 - u_1u_3, -\frac{8}{3}u_3 + u_1u_2),$$

which is the famous *Lorenz system*. Lorenz found that the solution of this system is very sensitive to perturbations. The system has some vague connection to a very simple model for fluid flow and has been given the role of explaining properties of fluid motion, such as turbulence. This was not Lorenz' original idea, who just wanted to make a connection to the apparent unpredictability and supposed sensitivity to perturbations of common meteorological models. If the seemingly very harmless and innocent Lorenz system could have unpredictable solutions, then there should be no surprise that also the weather could be unpredictable.

More precisely, Lorenz found that two solutions of the Lorenz system with very close initial data will stay close for some time but will eventually move apart completely. The Lorenz system is therefore very difficult to

solve accurately using a numerical method over times longer than say 30 units. The numerical solution will stay close the the exact solution for some time, but will eventually move apart significantly. Of course there are many IVP:s sharing this property of instability. Even the simple pendulum has this property if the pendulum reaches the top position with small velocity. It is thus remarkable that the Lorenz system seemed to present some kind of surprise to the scientific world. But it did, and it has become quite popular to explain all sorts of phenomena, from turbulence to politics, by referring to the "strange attractor" supposedly being displayed in plots of solutions of the Lorenz system.

The Lorenz system in component form reads:

$$\begin{cases} \dot{u_1} = -10u_1 + 10u_2, \\ \dot{u_2} = 28u_1 - u_2 - u_1u_3, \\ \dot{u_3} = -\frac{8}{3}u_3 + u_1u_2, \\ u_1(0) = u_{01},\ u_2(0) = u_{02},\ u_3(0) = u_{03}, \end{cases} \tag{256.1}$$

and u_0 is a given initial condition. The system (256.1) has three equilibrium points $\bar{u}$ with $f(\bar{u}) = 0$: $\bar{u} = (0, 0, 0)$ and $\bar{u} = (\pm 6\sqrt{2}, \pm 6\sqrt{2}, 27)$. The equilibrium point $\bar{u} = (0, 0, 0)$ is unstable with the corresponding Jacobian $f'(\bar{u})$ having one positive (unstable) eigenvalue and two negative (stable) eigenvalues. The equilibrium points $(\pm 6\sqrt{2}, \pm 6\sqrt{2}, 27)$ are slightly unstable with the corresponding Jacobians having one negative (stable) eigenvalue and two eigenvalues with very small positive real part (slightly unstable) and also an imaginary part. More precisely, the eigenvalues at the two non-zero equilibrium points are $\lambda_1 \approx -13.9$ and $\lambda_{2,3} \approx .0939 \pm 10.1i$.

In Fig. 256.1, we present two views of a solution $u(t)$ that starts at $u(0) = (1, 0, 0)$ computed to time 30 with an error tolerance of $TOL = 0.5$ using an adaptive IVP-solver of the form presented in Chapter *Adaptive IVP-solvers*. We can think of $u(t) = (x(t), y(t), z(t))$ as the position at time t of a particle that moves according to the equation $\dot{u} = f(u)$. In Fig. 256.1 we thus plot the trajectory or path followed by the particle as the particle moves with increasing time. The plotted trajectory is typical: the particle is kicked away from the unstable point $(0, 0, 0)$ and moves towards one of the non-zero equilibrium points. It then slowly orbits away from that point and at some time decides to cross over towards the other non-zero equilibrium point, again slowly orbiting away from that point and coming back again, orbiting out, crossing over, and so on. This pattern of some orbits around one non-zero equilibrium point followed by a transition to the other non-zero equilibrium point is repeated with a seemingly random number of revolutions around each equilibrium point.

As noted by Lorenz, a close inspection of the trajectory in Fig. 256.1 reveals quite a bit of structure in the behavior of the solution. From the path of the trajectory, it seems that, roughly speaking, there are two flat "lobes" in which the orbits around the non-zero equilibrium points are

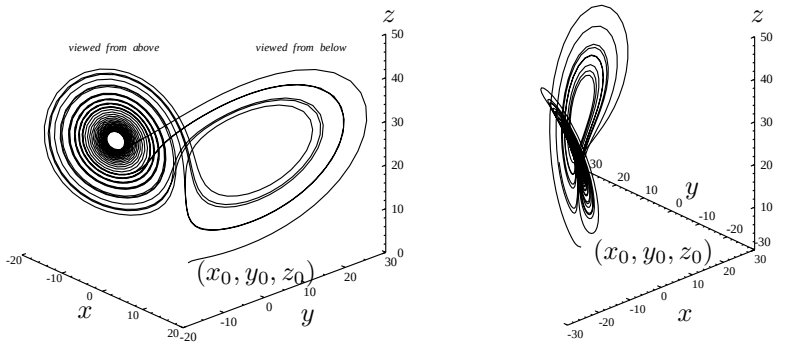


FIGURE 256.1. Two views of a numerical trajectory of the Lorenz system over the time interval $[0, 30]$ starting at $(1, 0, 0)$ computed with absolute error tolerance 0.5.

located. In each lobe, the spiraling segments of the trajectory seem to be grouped in "bands" that are made up of parts of the trajectory that are spiralling out from the equilibrium point and parts of the trajectory that have just crossed over from the other lobe. Only the trajectories in the outer band switch to the other equilibrium point. This causes a sharp separation between trajectories located in the outer band and those located in the next band inside as the trajectories approach the z-axis. We refer to this as *cutting* through the action of a "razor" separating the trajectories in the outer band. The trajectories in the outer band expand in width as they approach the other equilibrium point, with trajectories near the outside of the band ending up nearer to the fixed point. We refer to this as *expansion* and *flipping* respectively. The position of initial approach of the trajectory to an equilibrium point determines the number of orbits the trajectory makes in that lobe before returning to the other equilibrium point. Finally, we see that the orbits in one band come close to the next outer band after one revolution, this repeats with every band of the trajectory, until eventually they all end up in the outer band and leave towards the other equilibrium point. We refer to this as *interlacing*. In short, we can describe the dynamics of the Lorenz system as a never-ending process of cutting, expansion, flipping, and interlacing.

256.3 The Accuracy of the Computations

The first task is to measure the reliability of the computed error bound based on an a posteriori error estimate of the form presented in Chapter

Adaptive IVP-solvers. Since we do not have the exact solution, we perform the following experiment: Using the initial data $(0, 1, 0)$, we compute twice using residual tolerances 10^{-5} and 10^{-9} and approximate the error in the less accurate computation by taking the difference between the values of the less accurate and more accurate computations. In Fig. 256.2, we plot the computed error bound and the approximate error. The error bound predicts the size of the error quite well in spite of the sensitivity of the solution to perturbations. Similar results are obtained for a variety of initial data.

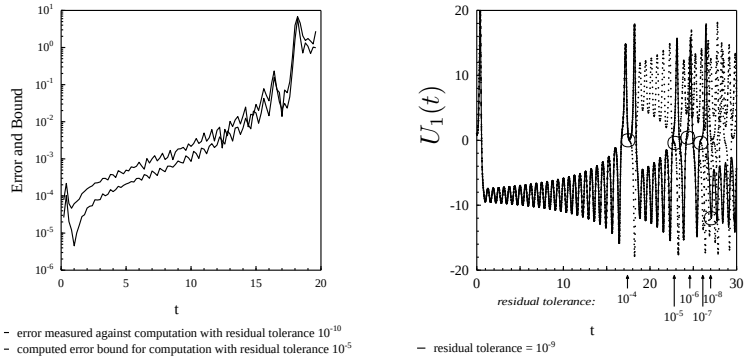


FIGURE 256.2. On the left, we plot the results of the reliability test for the computed error bound with initial condition is $(0, 1, 0)$. On the right, we plot the effect of changing the residual tolerance on the accuracy in $U_1(t)$ component with initial condition is $(1, 0, 0)$.

To give some idea of the behavior of the error control, we plot the step sizes used in a computation with absolute error tolerance 0.75 in Fig. 256.3. The step sizes vary roughly by a factor of 6 over the interval of computation. In Fig. 256.3, we also plot the product of the time step and the residual for this computation. We note that these values are kept within 10% of a constant value. With more computational work, the size of the variations can be reduced, which produces a more smoothly-varying error bound.

256.4 Computability of the Lorenz System

Encouraged by these results we decrease the tolerance or, equivalently, the time step, and try to compute an accurate solution to the Lorenz system on an even longer time interval. Using the cG(1) method as described in Chapter *Adaptive IVP-solvers*, we compute solutions with smaller and smaller time steps, $k = 0.01$, $k = 0.001$ and $k = 0.0001$, and expect to

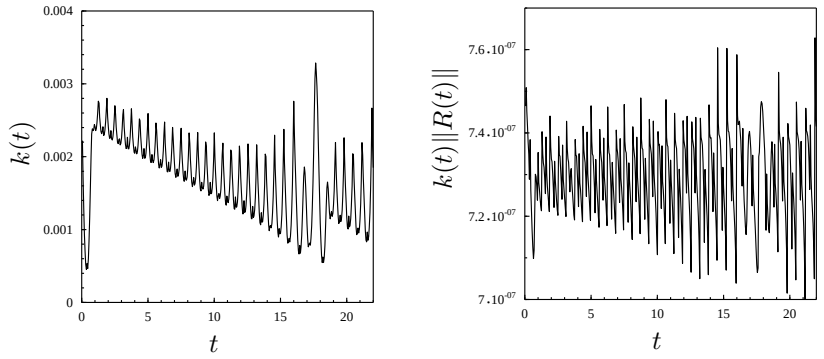


FIGURE 256.3. Time steps and residual $\times$ time step versus time for a computation beginning at $(1, 0, 0)$ with absolute error tolerance 0.75.

produce more and more accurate solutions. We plot the U_1-component of the solution in Fig. 256.4 where we also indicate the points at which the solutions are no longer accurate. We see that even with $300,000$ time steps the solution is not accurate beyond $t = 26$. Decreasing the time step with a factor 10 or 100 will take us only a little further, but the computation will take 10 or 100 times longer. We conclude that it is difficult to compute the solution to the Lorenz system over long time intervals.

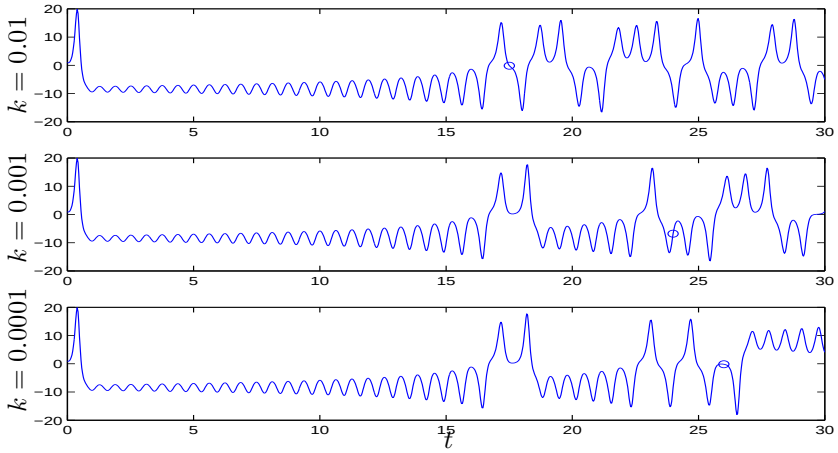


FIGURE 256.4. The U_1-component of the cG(1) solution for different time steps. The small circles indicate the points at which the solutions are no longer accurate.

To examine in detail the computability of the Lorenz system we return to the error estimate that we derived for the error $e(t)$ of the cG(1) method:

$$\|e(t)\| \le S_c(T) \max_{0\le t\le T} \|k(t)R(t)\|. \tag{256.2}$$

Remember that the stability factor $S_c(T)$ for the Lorenz system is defined in terms of the solution to the linearized dual problem as

$$S_c(T) = \max_{\varphi_0\in\mathbb{R}^3} \frac{\int_0^T \|\dot\varphi(t)\|\,dt}{\|\varphi_0\|}.$$

Judging by the error estimate we should be able to reach as far as we want if only the time step $k(t)$ and the residual $R(t)$ are small enough. However, a little more careful analysis reveals an additional error contribution, which is often ignored. Including also this term into our error estimate, we find:

$$\|e(t)\| \le S_c(T) \max_{0\le t\le T} \|k(t)R(t)\| + S_0(T) \max_{0\le t\le T} \epsilon/k(t), \tag{256.3}$$

where ϵ is the *machine precision* of the computer, i.e. the smallest number for which $1+\epsilon \neq 1$ (in computer arithmetic) and $S_0(T)$ is a new stability factor. For a standard computer (in 2002) with so-called double-precision arithmetic, the machine precision is $\epsilon \approx 10^{-16}$. The stability factor $S_0(T)$ is defined in terms of the dual solution as

$$S_0(T) = \max_{\varphi_0\in\mathbb{R}^3} \frac{\int_0^T \|\varphi(t)\|\,dt}{\|\varphi_0\|}.$$

The additional term in our refined error estimate (256.3) accounts for the round-off error made at every time step in the computation; when the new value $U(t_n)$ for the cG(1) solution is computed in every time step, it is unavoidable that we make a round-off error of size ϵ. As we shall see, it is the second term that sets the limit for the computability of the Lorenz system; the second term in (256.3) can be large even though the first term is small.

The difficulty of computing accurate solutions to the Lorenz system becomes obvious if we plot the size of the stability factors. In Fig. 256.5 we plot the size of the stability factor $S_0(T)$ associated with round-off errors as function of the final time T. Notice the logarithmic scale in this figure. A simple approximation of the growth of this stability factor is

$$S_0(T) \approx 10^{T/3},$$

and so the round-off error grows as $E_r = 10^{T/3}\cdot 10^{-16}/k = 10^{T/3-16}/k$. Notice how the error grows larger if we decrease the time step! This is natural (although unusual), since with a smaller time step we will have to take a larger number of time steps and thus make a larger number of round-off errors. To make the influence from round-off errors small we specify a

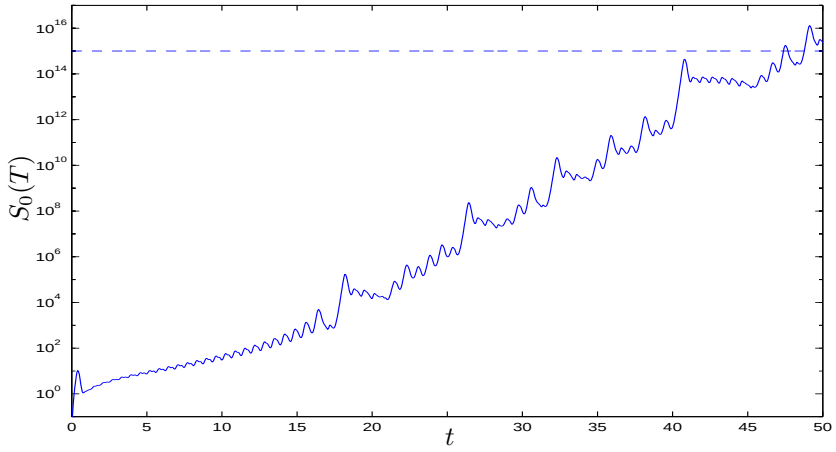


FIGURE 256.5. The growth of the stability factor $S_0(T)$ for the Lorenz problem.

large time step, say $k = 0.1$, for which the round-off error now grows as $10^{T/3-15}$. At time $T = 3 \cdot 15 = 45$ the accumulated round-off error is then $E_r = 1$, which means that we cannot expect to compute much beyond time $T = 45$, since then the round-off error will dominate anyway. Using the cG(1) method, we will not even reach $T = 45$, since we have to use a time step much smaller than $k = 0.1$ (as seen in Fig. 256.4) to make the first term in the error estimate small.

256.5 The Lorenz Challenge

From the previous discussion it is now clear that the mysterious unpredictability and "chaotic" behavior of the Lorenz system only means that the stability factors grow quickly, making it difficult to compute accurate solutions over long time intervals. The obvious challenge is now, using the method of choice, *to compute an accurate solution to the Lorenz system over as long a time interval* $[0, T]$ *as possible.*

We saw in the previous section that brute force is not the way to go. It is not enough to use a very fast computer with very small and very many time steps. Using the cG(1) method we cannot reach much further than $T = 30$, no matter how small time steps we use since then the accumulated round-off error will grow large. A solution to this problem would be if we could design a method, similar to cG(1), which can be used with larger time steps than what is possible with cG(1). As one can expect, there exist corresponding methods cG(2), cG(3) and so on, which can be used with larger time steps. It can be proved that for these cG(q) methods, the error

grows as k^{2q}, i.e. we have so called *a priori* error estimates of the type

$$\|e(T)\| \leq C(T)k^{2q},$$

where $C(T)$ is a constant (unknown!) depending on the exact solution $u(t)$. We say that the cG(q) method is of *order* $2q$. The standard cG(1) method is thus a second order method. (This is in agreement with (256.2) since one factor $k(t)$ is hidden inside $R(t)$.) With a higher order method, i.e. $q > 1$, we can thus obtain a smaller error with the same time step, which makes it possible to compute the solution with larger time steps. This in turn implies that with a higher order method, we can keep the round-off error smaller and thus reach further than what is possible with the cG(1) method. In Fig. 256.6 we plot the U_1-components of solutions to the Lorenz system, computed with time step $k = 0.1$, with a sequence of higher order methods. We see that with high enough order, the solutions agree to a point just beyond $T = 45$ as we predicted; the first term in our error estimate (256.3) has been reduced by increasing the order of the method and so the second term dominates. It is possible to reach beyond time $T = 50$, perhaps to $T = 100$, but to do this we have to go from double-precision arithmetic to quadruple-precision.

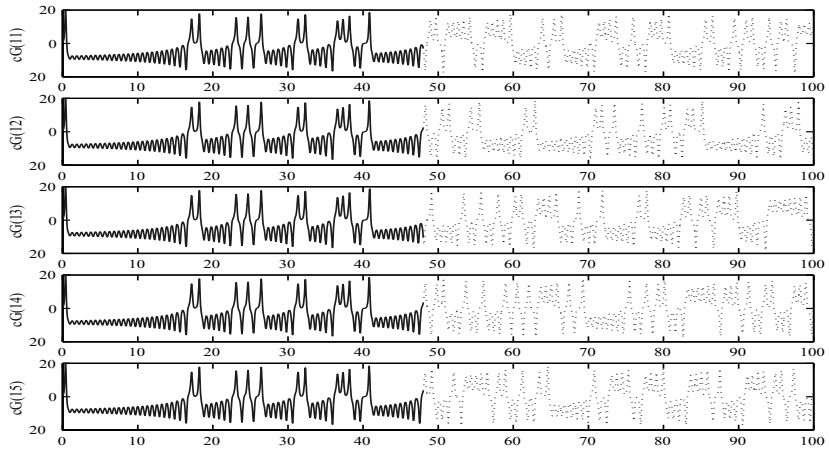


FIGURE 256.6. The U_1-component of cG(q) solutions for $q = 11, 12, 13, 14, 15$ with time step $k = 0.1$. Dashed lines indicate where the solution is no longer accurate.

256.6 The Lorenz World Record

Benjamin Kehlet and Anders Logg present in Long-Time Computation of the Lorenz System accurate solutions for $T = 1000$ using $cg(100)$ of order 200 computing with

420 correct decimals. This is the present World record. The contest continues...

Chapter 256 Problems

256.1. Verify that the three equilibrium points given in the text satisfy $f(u) = 0$. Linearize the system around these equilibrium points, i.e. compute the eigenvalues (and eigenvectors) for the Jacobian of f at the three equilibrium points.

256.2. Compute a solution to the Lorenz system and plot the orbit $(x(t), y(t), z(t))$ for $t \in [0, T]$. Do you agree with the description of the dynamics of the Lorenz system as never-ending process of cutting, expansion, flipping, and interlacing?

256.3. Repeat the experiment outlined in Section 256.4, i.e. compute solutions to the Lorenz system using the cG(1) method with a sequence of smaller and smaller time steps and examine the accuracy of the solutions (by comparing them to each other). Can you reach beyond $T = 25$?

256.4. Try the same experiment as in the previous problem but now with the lower order methods explicit Euler and implicit Euler. How far do you reach now?

256.5. Implement a simple version of the fourth-order cG(2) method given by

$$\begin{aligned} U(t_{n-1/2}) &= U(t_{n-1}) + \textstyle\int_{t_{n-1}}^{t^n} f(U(t),t)\cdot(5-6(t-t_{n-1})/k_n)/4\ dt, \\ U(t_n) &= U(t_{n-1}) + \textstyle\int_{t_{n-1}}^{t^n} f(U(t),t)\ dt, \end{aligned}$$

where $U(t)$ is the quadratic polynomial on $[t_{n-1}, t_n]$ determined by the three values $U(t_{n-1})$, $U(t_{n-1/2})$ and $U(t_n)$. How much further can you reach with this method?

256.6. Give a motivation for the additional term in the refined error estimate (256.3), starting from the estimate containing errors caused by using the wrong initial condition as presented in Chapter *Adaptive IVP-solvers*.

256.7. Take on the Lorenz Challenge, i.e. compute an accurate solution over $[0, T]$ with T as large as possible. No rules, all is allowed!

> But Aristarchus has brought out a book consisting of certain hypotheses, wherein it appears, as a consequence of the assumptions made, that the universe is many times greater than the 'universe' just mentioned. His hypotheses are that the fixed stars and the sun remain unmoved, that the earth revolves about the sun on the circumference of a circle, the sun lying in the middle of the orbit, and that the sphere of fixed stars, situated about the same centre as the sun, is so great that the circle in which he supposes the earth to revolve bears such a proportion to the distance of the fixed stars as the centre of the sphere bears to its surface. (Archimedes about Aristharcus of Samos)

257
The Solar System*

> There is talk of a new astrologer who wants to prove that the earth moves and goes around instead of the sky, the sun, the moon, just as if somebody were moving in a carriage or ship might hold that he was sitting still and at rest while the earth and the trees walked and moved. But that is how things are nowadays: when a man wishes to be clever he must needs invent something special, and the way he does it must needs be the best! The fool wants to turn the whole art of astronomy upside-down. However, as Holy Scripture tells us, so did Joshua bid the sun to stand still and not the earth.
> (Sixteenth century reformist M. Luther in his table book *Tischreden*, in response to Copernicus' pamphlet *Commentariolus*, 1514.)

257.1 Introduction

The problem of mathematical modeling of our Solar System including the Sun, the 9 planets Venus, Mercury, Tellus (the Earth), Mars, Jupiter, Saturn, Uranus, Neptune and Pluto together with a large number of moons and asteroids and occasional comets, has been of prime concern for humanity since the dawn of culture. The ultimate challenge concerns mathematical modeling of the Universe consisting of billions of galaxies each one consisting of billions of stars, one of them being our own Sun situated in the outskirts of the Milky Way galaxy.

According to the *geocentric* view presented by Aristotle (384-322 BC) in *The Heavens* and further developed by Ptolemy (87-150 AD) in *The Great*

System dominating the scene over 1800 years, the Earth is the center of the Universe with the Sun, the Moon, the other planets and the stars moving around the Earth in a complex pattern of circles upon circles (so-called epicycles). Copernicus (1473–1543) changed the view in *De Revolutionibus* and placed the Sun in the center in a new *heliocentric* theory, but kept the complex system of epicycles (now enlarged to a very complex system of 80 circles upon circles). Johannes Kepler (1572–1630) discovered, based on the extensive accurate observations made by the Swedish/Danish scientist Tycho Brahe (1546–1601), that the planets move in elliptic orbits with the Sun in one of the foci following *Kepler's laws*, which represented an enormous simplification and scientific rationalization as compared to the system of epicycles.

In fact, already Aristarchus (310-230 BC) of Samos understood that the Earth rotates around its axis and thus could explain the (apparent) motion of the stars, but these views were rejected by Aristotle arguing as follows: if the Earth is rotating, how is it that an object thrown upwards falls on the same place? How come this rotation does not generate a very strong wind? No one until Copernicus could question these arguments. Can you?

Newton (1642–1727) then cleaned up the theory by showing that the motion of the planets could be explained from one single hypothesis: the inverse square law of gravitation, see Chapter *Newton's nightmare* below. In particular, Newton derived Kepler's laws for the *two-body problem* with one (small) planet in an elliptic orbit around a (large) sun, see Chapter *Lagrange and the Principle of Least Action*. Leibniz criticized Newton for not giving any explanation of the inverse square law, which Leibniz believed could be derived from some basic fact, beyond one of "mutual love" which was quite popular. A sort of explanation was given by Einstein (1879–1955) in his theory of *General Relativity* with gravitation arising as a consequence of space-time being "curved" by the presence of mass. Einstein revolutionized *cosmology*, the theory of the Universe, but relativistic effects only add small corrections to Newton's model for our Solar System based on the inverse square law. Einstein gave no explanation why space-time gets curved by mass, and still today there is no convincing theory of gravitation with its mystical feature of "action at a distance" through some mechanism yet to be discovered. In Chapter *Laplacian Models* below we give a derivation of the inverse square law using a mathematical argument presented by Laplace.

Despite the lack of a physical explanation of the inverse square law, Newton's theory gave an enormous boost to mathematical sciences and a corresponding kick to the egos of scientists: if the human mind was capable of (so easily and definitely) understanding the secrets of the Solar System, then there could be no limits to the possibilities of scientific progress...

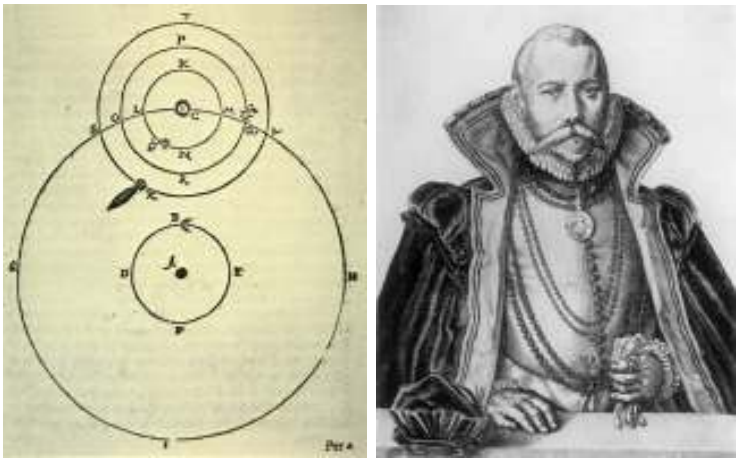


FIGURE 257.1. Tycho Brahe: “I believe that the Sun and the Moon orbit around the Earth but that the other planets orbit around the Sun”.

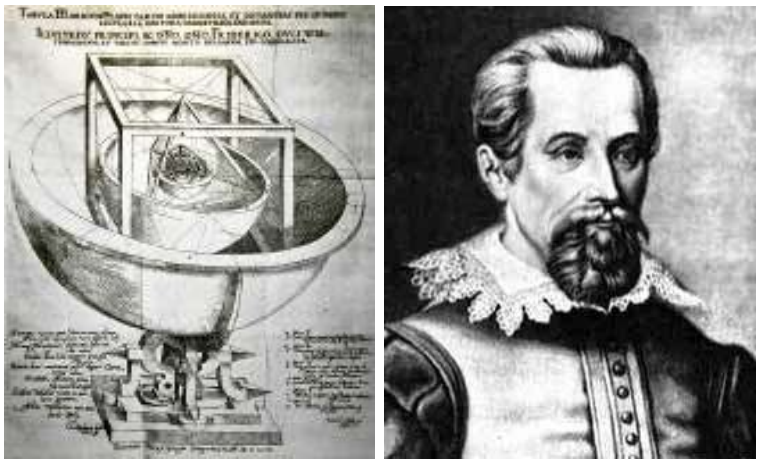


FIGURE 257.2. Johannes Kepler:“I believe that the planets are separated by invisible regular polyhedra: tetrahedron, cube, octahedron, dodekahedron and ikosahedron, and further that the planets including the Earth move in elliptical orbits around the Sun”.

257.2 Newton's Equation

The basis of celestial mechanics is Newton's second law,

$$F = m \cdot a, \tag{257.1}$$

expressing that a *force* F results in an acceleration of size a for a body of mass m, together with the expression for the gravitational force given by the inverse square law:

$$F = G\frac{mm_a}{r^2}, \tag{257.2}$$

where $G \approx 6.67 \cdot 10^{-11}\text{Nm}^2/\text{kg}^2$ is the *gravitational constant*, m_a is the mass of the attracting body and r is the distance to the attracting body.

Together (257.1) and (257.2) give a set of differential equations for the evolution of the Solar System. If we know the initial positions and velocities for all bodies in the Solar System, we can solve the system of differential equations, using the same techniques as presented above in Chapter Adaptive IVP-Solvers. We discuss this in detail below in Section 257.4. As a preparation, we rewrite (257.1) and (257.2) in dimensionless form, which will be convenient. The three fundamental units appearing in the equations are those of *space*, *time* and *mass*, which are represented by the variables x (or r), t and m. We now introduce new dimensionless variables, $x' = x/\text{AU}$, $t' = t/\text{year}$ and $m' = m/M$, where 1 AU is the mean distance from the Sun to Earth and M is the mass of the Sun. We can use the chain rule to obtain the dimensionless acceleration, $a' = \frac{d}{dt'}\frac{d}{dt'}x' = \frac{dt}{dt'}\frac{d}{dt}\frac{dt}{dt'}\frac{d}{dt}x/\text{AU} = \frac{\text{year}^2}{\text{AU}}a$. Combining (257.1) and (257.2) using our new dimensionless variables, we then obtain

$$m'M\frac{\text{AU}}{\text{year}^2}a' = G\frac{m'M \cdot m'_aM}{r'^2\text{AU}^2}, \tag{257.3}$$

or

$$a' = G'\frac{m'_a}{r'^2}, \tag{257.4}$$

where the new gravitational constant G' is given by

$$G' = \frac{G \cdot \text{year}^2 M}{\text{AU}^3}. \tag{257.5}$$

We leave it as an exercise to show that with suitable definitions of the units year and AU, the new dimensionless gravitational constant G' is given by

$$G' = 4\pi^2. \tag{257.6}$$

257.3 Einstein's Equation

In general relativity the basic concept is not *force*, as in Newtonian theory, but instead the *curvature* of space-time. Einstein explains the motion of

the planets in our Solar System in the following way: the planets move through space-time along straight lines, *geodesics*, which appear as circular (or elliptical) orbits only because space-time is curved by the large mass of the Sun. We shall now try to give an idea of how this works.

The curvature of space-time is given by its *metric*. A metric defines the distance between two nearby points in space-time. In Euclidean geometry that we have studied extensively in this book, the distance between to points $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3)$ is given by the square root of the scalar product $dx \cdot dx$, where dx is the difference $dx = x - y$. With the notation $ds = |x - y|$ we thus have

$$ds = \sqrt{dx \cdot dx} = \left(\sum_{i=1}^{3} dx_i^2 \right)^{1/2}, \tag{257.7}$$

or

$$ds^2 = \sum_{i=1}^{3} dx_i^2. \tag{257.8}$$

In the notation of general relativity, the Euclidean metric is then given by the matrix (tensor)

$$g = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{257.9}$$

as

$$ds^2 = dx^T g\ dx. \tag{257.10}$$

In space-time we include time t as a fourth coordinate and every event in space-time is given by a vector (t, x_1, x_2, x_3). In flat or *Minkowski* space-time in the absence of masses, the curvature is zero and the metric is given by

$$g = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{257.11}$$

which gives

$$ds^2 = -dt^2 + dx_1^2 + dx_2^2 + dx_3^2. \tag{257.12}$$

In the presence of masses, we obtain a different metric which does not even have to be diagonal.

From the metric g one can find the straight lines of space-time, which give the orbits of the planets. The metric itself is determined by the distribution of mass in space-time, and is given by the solution of Einstein's equation,

$$R_{ij} - \frac{1}{2} R g_{ij} = 8\pi T_{ij}, \tag{257.13}$$

where (R_{ij}) is the so-called *Ricci-tensor*, R is the so-called *scalar curvature* and (T_{ij}) is the so-called *stress-energy tensor*. Now (R_{ij}) and R depend on derivatives of the metric $g = (g_{ij})$ so (257.13) is a partial differential equation for the metric g.

The solution for the orbits of the planets obtained from Einstein's equation are a little different than the solution obtained from (257.4) given by Newton. Although the difference is small, it has been verified in observations of the orbit of the planet Mercury which is the planet closest to the Sun. We will not include these "relativistic effects" in the next section where we move on to the computation of the evolution of the Solar System.

257.4 The Solar System as a System of ODEs

To use the techniques developed in Chapter Adaptive IVP-Solvers to compute the evolution of the Solar System, we need to rewrite the second-order system of ODEs given by (257.4) in the standard form $\dot{u} = f$. We start by introducing coordinates $x^i(t) = (x_1^i(t), x_2^i(t), x_3^i(t))$ for all bodies in the Solar System, including the nine planets, then Sun and the Moon. This gives a total of $n = 9 + 2 = 11$ bodies and a total of $3n = 33$ coordinates. To rewrite the equations as the first-order system $\dot{u} = f$ we need to include also the velocities of all bodies, $\dot{x}^i(t) = (\dot{x}_1^i(t), \dot{x}_2^i(t), \dot{x}_3^i(t))$, giving a total of $N = 6n = 66$ coordinates. We collect all these coordinates in the vector $u(t)$ of length N in the following order:

$$\begin{aligned} u(t) = (&x_1^1(t), x_2^1(t), x_3^1(t), \ldots, x_1^n(t), x_2^n(t), x_3^n(t),\\ &\dot{x}_1^1(t), \dot{x}_2^1(t), \dot{x}_3^1(t), \ldots, \dot{x}_1^n(t), \dot{x}_2^n(t), \dot{x}_3^n(t)), \end{aligned} \tag{257.14}$$

so that the first half of the vector $u(t)$ contains the positions of all bodies and the second half contains the corresponding velocities.

To obtain the differential equation for $u(t)$, we take the time-derivative and notice that the derivative of the first half of $u(t)$ is equal to the second half of $u(t)$:

$$\dot{u}_i(t) = u_{3n+i}(t), \quad i = 1, \ldots, 3n, \tag{257.15}$$

i.e. for $n = 11$ we have $\dot{u}_1(t) = \dot{x}_1^1(t) = u_{34}(t)$ and so on.

The derivative of the second half of $u(t)$ will contain the second derivatives of the positions, i.e. the accelerations, and these are given by (257.4). Now (257.4) is written as a scalar equation and we have to rewrite it in vector form. For every body in the Solar System, we need to compute the contribution to the total force on the body by summing the contributions from all other bodies. Assuming that we work in dimensionless variables (but writing x instead of x', m_i instead of m_i' and so on for convenience)

we then need to compute the sum:

$$\ddot{x}_i(t) = \sum_{j \neq i} \frac{G' m_j}{|x^j - x^i|^2} \frac{x^j - x^i}{|x^j - x^i|}, \tag{257.16}$$

where the unit vector $\frac{x^j - x^i}{|x^j - x^i|}$ gives the direction of the force, see Figure 257.3.

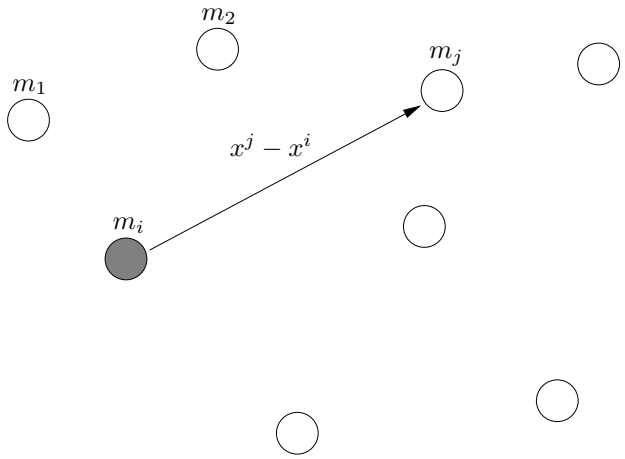


FIGURE 257.3. The total force on body i is the sum of the contributions from all other bodies.

Our final differential equation for the evolution of the Solar System in the form $\dot{u} = f$ is then given by

$$\dot{u}(t) = f(u(t)) = \begin{bmatrix} u_{3n+1}(t) \\ \vdots \\ u_{6n}(t) \\ \sum_{j \neq 1} \frac{G' m_j}{|x^j - x^1|^2} \frac{x_1^j - x_1^1}{|x^j - x^i|} \\ \vdots \\ \sum_{j \neq n} \frac{G' m_j}{|x^j - x^n|^2} \frac{x_3^j - x_3^n}{|x^j - x^n|} \end{bmatrix}, \tag{257.17}$$

where we have kept the notation $x^1 = (x_1^1, x_2^1, x_3^1)$ rather than (u_1, u_2, u_3) and so on in the right-hand side for simplicity. The evolution of our Solar System can now be computed by the standard techniques developed in Chapter Adaptive IVP-Solvers, using the initial data supplied in Table 257.1.

	Position	Velocity	Mass
Mercury	$x^1(0) = \begin{matrix}-0.147853935\\-0.400627944\\-0.198916163\end{matrix}$	$\dot{x}^1(0) = \begin{matrix}7.733816715\\-2.014137426\\-1.877564183\end{matrix}$	1.0/6023600
Venus	$x^2(0) = \begin{matrix}-0.725771746\\-0.039677000\\0.027897127\end{matrix}$	$\dot{x}^2(0) = \begin{matrix}0.189682646\\-6.762413869\\-3.054194695\end{matrix}$	1.0/408523.5
Earth	$x^3(0) = \begin{matrix}-0.175679599\\0.886201933\\0.384435698\end{matrix}$	$\dot{x}^3(0) = \begin{matrix}-6.292645274\\-1.010423954\\-0.438086386\end{matrix}$	1.0/328900.5
Mars	$x^4(0) = \begin{matrix}1.383219717\\-0.008134314\\-0.041033184\end{matrix}$	$\dot{x}^4(0) = \begin{matrix}0.275092348\\5.042903370\\2.305658434\end{matrix}$	1.0/3098710
Jupiter	$x^5(0) = \begin{matrix}3.996313003\\2.731004338\\1.073280866\end{matrix}$	$\dot{x}^5(0) = \begin{matrix}-1.664796930\\2.146870503\\0.960782651\end{matrix}$	1.0/1047.355
Saturn	$x^6(0) = \begin{matrix}6.401404019\\6.170259699\\2.273032684\end{matrix}$	$\dot{x}^6(0) = \begin{matrix}-1.565320566\\1.286649577\\0.598747577\end{matrix}$	1.0/3498.5
Uranus	$x^7(0) = \begin{matrix}14.423408013\\-12.510136707\\-5.683124574\end{matrix}$	$\dot{x}^7(0) = \begin{matrix}0.980209400\\0.896663122\\0.378850106\end{matrix}$	1.0/22869
Neptune	$x^8(0) = \begin{matrix}16.803677095\\-22.983473914\\-9.825609566\end{matrix}$	$\dot{x}^8(0) = \begin{matrix}0.944045755\\0.606863295\\0.224889959\end{matrix}$	1.0/19314
Pluto	$x^9(0) = \begin{matrix}-9.884656563\\-27.981265594\\-5.753969974\end{matrix}$	$\dot{x}^9(0) = \begin{matrix}1.108139341\\-0.414389073\\-0.463196118\end{matrix}$	1.0/150000000
Sun	$x^{10}(0) = \begin{matrix}-0.007141917\\-0.002638933\\-0.000919462\end{matrix}$	$\dot{x}^{10}(0) = \begin{matrix}0.001962209\\-0.002469700\\-0.001108260\end{matrix}$	1
Moon	$x^{11}(0) = \begin{matrix}-0.177802714\\0.884620944\\0.384016593\end{matrix}$	$\dot{x}^{11}(0) = \begin{matrix}-6.164023246\\-1.164502534\\-0.506131880\end{matrix}$	$1.0/2.674 \cdot 10^7$

TABLE 257.1. Initial data for the Solar System at 00.00 Universal Time (UT1, approximately GMT) January 1 2000 for dimensionless positions and velocities scaled with units 1 AU $= 1.49597870 \cdot 10^{11}$ m (one astronomical unit), 1 year $= 365.24$ days and $M = 1.989 \cdot 10^{30}$ kg (one solar mass).

257.5 Predictability and Computability

Two important questions that arise naturally when we study numerical solutions of the evolution of our Solar System, such as the one in Figure 257.4, are the questions of *predictability* and *computability*.

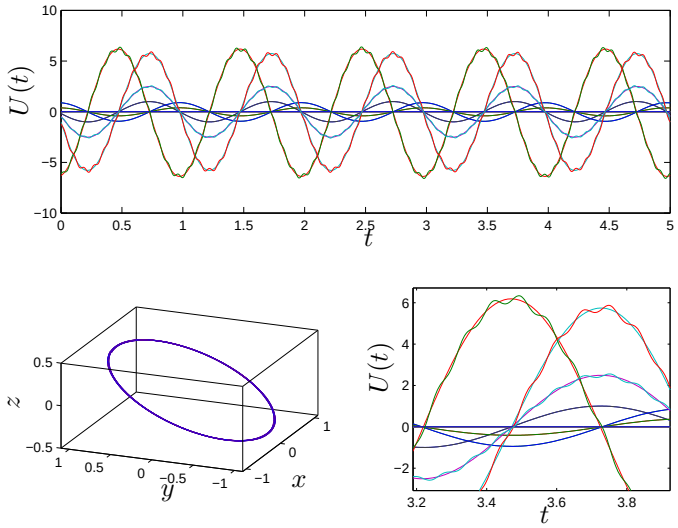


FIGURE 257.4. A numerical computation of the evolution of the Solar System, including Earth, the Sun and the Moon.

The predictability of the Solar System is the question of the accuracy of a computation given the accuracy in initial data. If initial data is known with an accuracy of say five digits, and the numerical computation is exact, how long does it take until the solution is no longer accurate even to one digit?

The computability of the Solar System is the question of the accuracy in a numerical solution given exact initial data, i.e. how far we can compute an accurate solution with available resources such as method, computational power and time.

Both the predictability and the computability are determined by the growth rate of errors. Luckily, the error does not grow exponentially as we saw for the Lorenz system. If we imagine that we displace Earth slightly from its orbit and start a computation, the orbit and velocity of Earth will be slightly different, resulting in an error that grows *linearly* with time. This means that the predictability of the Solar System is quite good, since every extra digit of accuracy in initial data means that the limit of predictability is increased by a factor ten. If now the solution is computed using a numerical method, such as the adaptive cG(1) method, this will

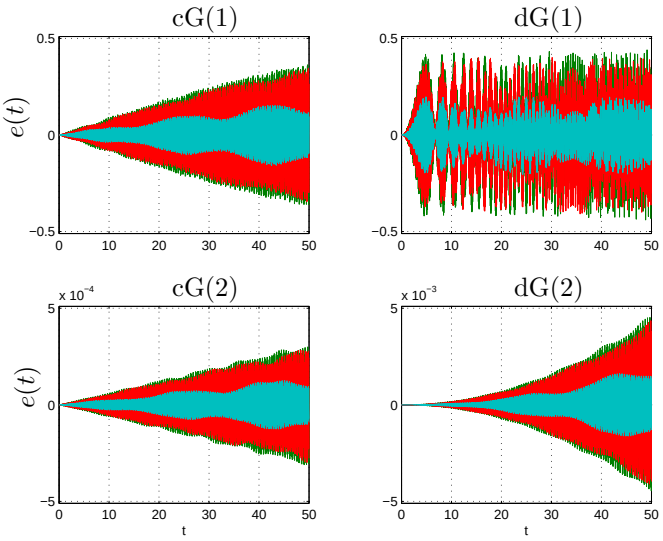


FIGURE 257.5. The growth of the numerical error in simulations of the Solar System using different numerical methods. The two methods on the left conserve energy, which results in linear rather than quadratic error growth.

result in additional errors. We can think of the error caused by a numerical method as a small perturbation introduced with every new time step. Adding the contributions from all time steps we find that the numerical error grows *quadratically*, see Problem 257.2.

As it turns out however, the error does not grow quadratically but only linearly for the cG(1) method as shown in Figure 257.5. This pleasant surprise is the result of an important property of the cG(1) method: it conserves energy. As a result, the cG(1) method performs better on a long time interval than the higher-order (more accurate) dG(2) method.

257.6 Adaptive Time-Stepping

If we compute the evolution of the Solar System using the adaptive cG(1) method, we find that the time steps need to be small enough to follow the orbit of the Moon (or Mercury if we do not include the Moon). This is inefficient since the time scales for the other bodies are much larger: the period of the Moon is one month and the period of Pluto is 250 years, and so the time steps for Pluto should be roughly a factor 3,000 larger that the time steps for the Moon. It has been shown recently that the standard methods cG(q), including cG(1), and dG(q) can be extended to individual, *multi-adaptive*, time-stepping for different components. In Figure 257.6 we show a computation made with individual time steps for the different planets.

Notice how the error grows quadratically, indicating that the method does not conserve energy. (It is possible to construct also multi-adaptive methods which conserve energy.)

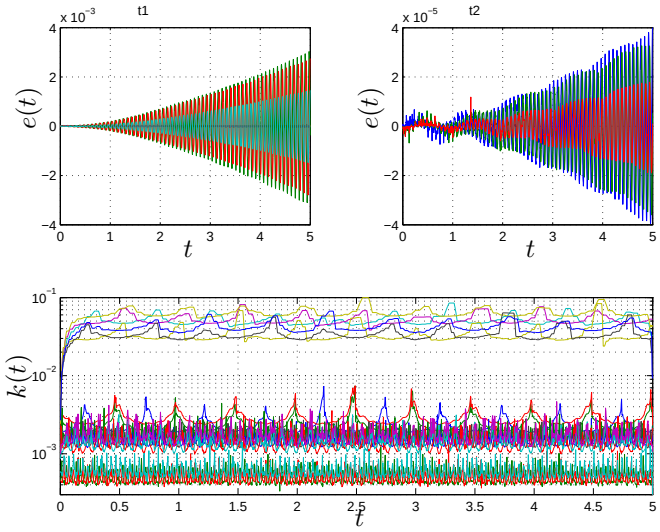


FIGURE 257.6. A computation of the evolution of the Solar System with individual, multi-adaptive, time steps for the different planets.

257.7 Limits of Computability and Predictability

Using the multiadaptive cG(2) it appears that the limit of computability of the Solar System (with the Moon and the nine planets) using double precision, is of the order 10^6 years. Concerning the predictability of the same system it appears that for every digit beyond 5 in the precision of data we gain a factor of ten in time, so that for example predicting the position of the Moon 1000 years ahead would require about 8 correct digits in e.g. the initial positions and velocities, masses and gravitational constant. We conclude that it appears that normally the precision in data would set the limit for accurate simulations of the evolution of the Solar System, if we use a high order multiadaptive solver.

Chapter 257 Problems

257.1. Prove that with suitable definitions of the units year and AU the gravitational constant is $G' = 4\pi^2$. *Hint:* assume that Earth is in a circular orbit where the centripetal force mv^2/r is balanced by the gravitational force GmM/r^2.

257.2. Motivate the quadratic growth of the numerical error for the Solar System. *Hint:* Assume that an error of size ϵ is added to the velocity of a planet in every time step.

257.3. (*Hard!*) Prove that in general if an error in initial data grows as

$$|e(T)| \leq S(T)|e(0)|,$$

for a specific initial value problem, then the error in a numerical solution of the initial value problem grows as

$$|e(T)| \lessapprox \epsilon \int_0^T S(t)\ dt,$$

assuming that the additional error in every time step is kept below $k_n\epsilon$.

257.4. Prove that the cG(1) method conserves energy for a Hamiltonian system, i.e. prove that for a system given by $\ddot{x} = F = -\nabla_x P(x)$, the total energy

$$E(t) = K(\dot{x}(t)) + P(x(t)),$$

is conserved. Here $P(x)$ is a given potential field, and $K(\dot{x}) = \frac{\dot{x}^2}{2}$ is the kinetic energy. *Hint:* Write as a first-order system for the vector $[u, v] = [x, \dot{x}]$, take $[\dot{v}, \dot{u}]$ as the test function and use the chain rule.

257.5. Investigate numerically the predictability and computability of the Solar System. Can you verify the linear error growth for the cG(1) method?

258
Newton's Nightmare*

> God does not care about mathematical difficulties. He integrates empirically. (Einstein)

Newton's theory of gravitation states that the gravitational force field $F(x)$ generated by a point mass m at the origin is the potential field

$$F(x) = -m\frac{x}{\|x\|^3} = \nabla(\frac{m}{\|x\|}), \tag{258.1}$$

corresponding to the potential $\varphi(x) = m/\|x\|$, in units where the gravitational constant is one. This means the gravitational force from the mass m at the origin on a unit point mass at position x is equal to $F(x)$. Taking norms gives

$$\|F(x)\| = \frac{m}{\|x\|^2},$$

which is known as *Newton's Inverse Square Law* . More generally, the gravitational force field of a mass m a position y is given by

$$F(x) = -m\frac{x-y}{\|x-y\|^3}, \tag{258.2}$$

with $F(x)$ being the force on a unit point mass a position x and the corresponding potential $\varphi(x) = m/\|x-y\|$.

Over a long period, Newton tried to show one consequence of his new theory of gravitation: the gravitational force between two solid balls is the same as if the total mass of each ball was concentrated at the center of

FIGURE 258.1. Isaac Newton 1689: "I have not been able to discover the cause of those properties of gravity from phenomena, and I frame no hypotheses; for whatever is not deduced from the phenomena is to be called a hypothesis, and hypotheses, whether metaphysical or physical, whether of occult qualities or mechanical, have no place in experimental philosophy".

mass of each ball. This result has important practical implications. For example, it would allow the modeling of the solar system as 9 small point masses representing the planets orbiting around one fixed big point mass representing the Sun, that is as a 9-body system. Without the simplifying basic result, we would have to take into account the gravitational attraction between the parts of each of the bodies and we would end up with a very complicated model. The practicality of Newton's gravitational theory could easily be questioned by anyone having some interest in that direction, like the Church. Lacking this basic result, Newton delayed the publication of his monumental *Principia Mathematica* many years. Newton states that he purposely made Principia difficult to read "to avoid being bated by little smatterers of mathematics". Newton did not like critics.

It fact, even a 9-body system of point masses may be far beyond comprehension or mathematical analysis. Luckily, the solar system is a very special 9-body system in which the motion of each planet can be viewed to good approximation as a 1-body system, i.e. as each planet orbiting undisturbed around one heavy Sun. Such 1-body systems have a full analytical solution available, as we saw in the Chapter *Lagrange and the Principle of Least Action*.

The basic result that Newton finally succeeded in proving can be phrased as follows: Consider a thin spherical shell S of radius r and uniform thickness centered at the origin and assume the total mass of the shell is m. Let $F(x)$ be the gravitational force field generated by the spherical shell so that $F(x)$ is the gravitational force of the shell on a unit point mass at

position x outside the sphere. Newton proved that

$$F(x) = -m\frac{x}{\|x\|^3} \quad \text{for } \|x\| > r,$$

which says that the gravitational field generated by the sphere on a point outside the sphere is the same as the field generated by a point mass m at the center of the sphere.

The gravitational field $F(x)$ of the shell/sphere is the sum of the gravitational fields of all the little pieces $ds(y)$ of the surface of mass $dm(y)$ at position y making up the sphere S, that is

$$F(x) = \int_S f(y)ds(y),$$

where

$$f(y)ds(y) = -dm(y)\frac{x-y}{\|x-y\|^3}$$

is the gravitational field of the piece of surface $ds(y)$ of mass $dm(y)$ at position y. We note that

$$dm(y) = \frac{mds(y)}{4\pi r^2},$$

since the area of the sphere is $4\pi r^2$ and the total mass is m, and thus

$$f(y) = -\frac{m}{4\pi r^2}\frac{x-y}{\|x-y\|^3}. \tag{258.3}$$

Newton thus wanted to verify that

$$\int_S f(y)\, ds(y) = -m\frac{x}{\|x\|^3} \quad \text{for } \|x\| > r, \tag{258.4}$$

where $f(y)$ is given by (258.3). Once this basic result for a sphere is established, the corresponding result for a solid ball follows by simply viewing the ball as the union of a collection of thin spheres of varying radii. The desired final result for two solid balls follows similarly.

We now prove (258.4) giving the gravitational field of a thin spherical shell S of radius r and total mass m centered at the origin. We assume that $x = (R, 0, 0)$ with $R > r$. By symmetry, this covers the general situation. We note that the components $F_2(x)$ and $F_3(x)$ of the gravitational force vanish because the gravitational force is directed from $(R, 0, 0)$ towards the origin, and we have simply to verify that

$$F_1(x) = -\frac{m}{4\pi r^2}\int_S \frac{R-y_1}{\|x-y\|^3}ds(y) = -\frac{m}{R^2}.$$

To compute the surface integral, we use spherical coordinates

$$y = (r\cos(\varphi), r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta))$$

with $0 \le \varphi \le \pi$ and $0 \le \theta \le 2\pi$, see Fig. 258.2, and recall from the Chapter Surface integrals that

$$ds(y) = r^2 \sin(\varphi) d\varphi\, d\theta.$$

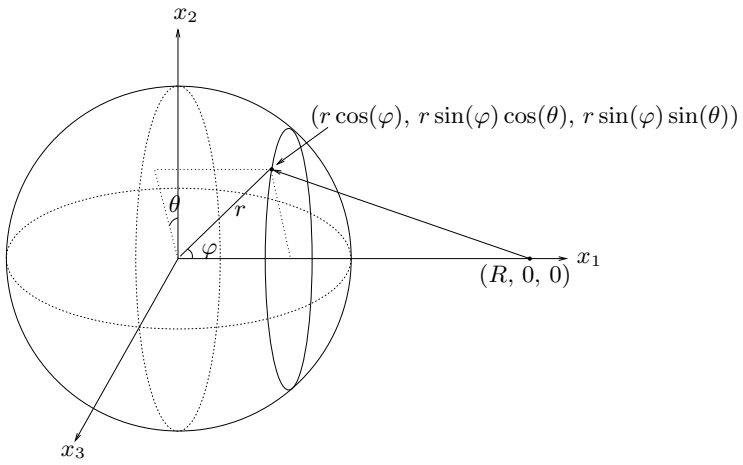


FIGURE 258.2. Newtons nightmare:

We have according to Fig. 258.2,

$$\begin{aligned} F_1(x) &= -\frac{m}{4\pi r^2} \int_S \frac{R - y_1}{\|x - y\|^3} ds(y) \\ &= -\frac{m}{4\pi} \int_0^{\pi} \int_0^{2\pi} \frac{(R - r\cos(\varphi))\sin(\varphi)}{((R - r\cos(\varphi))^2 + (r\sin(\varphi))^2)^{3/2}}\, d\theta d\varphi \\ &= -\frac{m}{2} \int_0^{\pi} \frac{(R - r\cos(\varphi))\sin(\varphi)}{((R - r\cos(\varphi))^2 + (r\sin(\varphi))^2)^{3/2}}\, d\varphi \end{aligned}$$

where we performed the integration with respect to θ using the fact that the integrand is independent of θ. We thus need to verify that

$$I = \int_0^{\pi} \frac{(R - r\cos(\varphi))\sin(\varphi)\, d\varphi}{((R - r\cos(\varphi))^2 + (r\sin(\varphi))^2)^{3/2}} = \frac{2}{R^2}. \tag{258.5}$$

To this end, we change variables to set $t = \cos(\varphi)$ and we use $dt = -\sin(\varphi) d\varphi$ to get

$$I = \int_{-1}^{1} \frac{(R - rt)\, dt}{(R^2 + r^2 - 2Rrt)^{3/2}} = \frac{1}{R^2} \int_{-1}^{1} \frac{(1 - at)\, dt}{(1 + a^2 - 2at)^{3/2}},$$

where $a = \frac{r}{R} < 1$. By a routine computation, we find that $a < 1$ implies

$$\begin{aligned}
&\int_{-1}^{1} \frac{(1-at)\,dt}{(1+a^2-2at)^{3/2}} \\
&= \int_{-1}^{1} \frac{(\frac{1+a^2}{2}-at)\,dt}{(1+a^2-2at)^{3/2}} - \int_{-1}^{1} \frac{(\frac{1+a^2}{2}-1)\,dt}{(1+a^2-2at)^{3/2}} \\
&= \frac{1}{2a}\big[-(1+a^2-2at)^{1/2}\big]_{-1}^{1} - \frac{a^2-1}{2a}\big[(1+a^2-2at)^{-1/2}\big]_{-1}^{1} \\
&= \frac{1}{2a}\big(1+a-(1-a)\big) - \frac{a^2-1}{2a}\big(\frac{1}{1-a}-\frac{1}{1+a}\big) = 1+1 = 2,
\end{aligned}$$

and the desired result follows:

$$F_1(x) = -\frac{m}{R^2} \quad \text{if } x = (0,0,R), \quad R > r.$$

Below, we give a much shorter proof of this result using some tools of Calculus to be developed in the next chapters.

Chapter 258 Problems

258.1. Prove that the gravitational field from a thin sphere is equal to zero *inside* the sphere.

258.2. Compute the gravitational field $F(x)$ for $x \in \mathbb{R}^3$ of a solid ball of total mass m and radius r centered at the origin

258.3. Compute the gravitational field of a "black hole" with mass density $\frac{exp(-r)}{r}$, $r = \|x\|$.

258.4. Determine the gravitational field generated by a thin straight uniform rod.

258.5. Determine the gravitational field generated by a thin circular flat (a) ring (b) disc.

258.6. (a) Consider a particle cloud of uniform density in the form of a ball. Assume the particles attract each other according to Newton's Law of gravitation. Compute the evolution of the cloud for $t > 0$ assuming the particles are at rest at $t = 0$. (b) Do the same with a cloud in the form of the volume between two concentric spheres. (c) Extend to clouds of variable density.

259
Chemical Reactions*

We already know the laws that govern the behavior of matter under all but the most extreme situations. In particular, we know the basic laws that underlie all of chemistry and biology. Yet we have certainly not reduced these objects to the status of solved problems; we have, as yet, had little success in predicting human behavior from mathematical equations. So even if we do find a complete set of basic laws, there will still be in the years ahead the intellectual challenging task of developing better approximation methods, so that we can make useful predictions of the probable outcomes in complicated and realistic situations. (S. Hawking in A Brief History of Time)

It is especially difficult to find exact solutions of the equations, as the equations (Einstein's equations) are non-linear. (Einstein)

Inasmuch as a propagating flame may be considered as a wave of chemical reactions sweeping across a flowing gas, it offers an excellent proving ground for the analytical skills of a fluid dynamicist, a heat and mass transfer specialist and a physical chemist, all put together into a well-rounded applied mathematician. (M. Kanury)

259.1 Constant Temperature

We consider N different chemical species A_1,..., A_N, which participate in J reactions with stoichiometric (positive) integer coefficients $\nu_{n,j}$ for species n appearing as reactant in reaction j and $\lambda_{n,j}$ for species n appearing as product in reaction j (with the coefficients being zero if the species is not

a reactant or product). This is commonly expressed as

$$\sum_{n=1}^{N} \nu_{n,j} A_n \to \sum_{n=1}^{N} \lambda_{n,j} A_n \quad \text{for } j = 1, ..., J. \tag{259.1}$$

We say that the *order* of reaction j is equal to $\sum_{n=1}^{N} \nu_{n,j}$. We denote the *molar concentration* (expressed in moles per unit volume) of species A_n by c_n. The *reaction rate* r_j of reaction j is supposed to be given by

$$r_j = k_j(T) \prod_{m=1}^{N} c_m^{\nu_{m,j}},$$

where the *reaction coefficient* or *Arrhenius factor* $k_j(T)$ is given by

$$k_j(T) = B_j T^{\alpha_j} \exp(-\frac{E_j}{RT}),$$

with $E_j > 0$ representing the *activation energy*, $B_j T^{\alpha_j}$ representing the *frequency factor*, B_j and α_j are positive constants, the absolute temperature T is assumed to be the same for all species, and R is the gas constant. The basic idea behind the product formula $\prod_{m=1}^{N} c_m^{\nu_{m,j}}$ is that the reaction rate is proportional to the molar concentrations of the reactants with each reactant A_m counted $\nu_{m,j}$ times. The Arrhenius factor is small if T is below some threshold value corresponding to the quotient $\frac{E_j}{RT}$ being moderately large.

The net production rate (moles per volume per unit time) of species A_n in reaction j is given by $\alpha_{n,j} r_j$, where

$$\alpha_{n,j} = \lambda_{n,j} - \nu_{n,j},$$

and the total net production rate s_n of species n is given by

$$s_n = \sum_{j=1}^{J} \alpha_{n,j} r_j.$$

We now assume that the temperature T is constant and is given, and we seek the vector of concentration $c(t) = (c_1(t), ..., c_N(t))$ as a function of time t describing the dynamics of the set of reactions for $t > 0$, assuming that $c(0) = c^0$, where $c^0 = (c_1^0, ..., c_N^0)$ is a given vector of initial concentrations. Using the balance equation $\dot{c}_n = s_n$ for each species $n = 1, 2, ..., N$, we obtain the following initial value problem for a system of ordinary differential equations: Find $c(t) = (c_1(t), ..., c_N(t))$ such that

$$\begin{cases} \dot{c}_n(t) = \sum_{j=1}^{J} \alpha_{n,j} k_j(T) \prod_{m=1}^{N} c_m(t)^{\nu_{m,j}} & \text{for } t > 0,\ n = 1, ..., N, \\ c(0) = c^0. \end{cases} \tag{259.2}$$

This is an initial value problem of the form $\dot{u}(t) = f(u(t))$ for $t > 0$, $u(0) = u^0$, where $u(t) = c(t)$ and $f : \mathbb{R}^N \to \mathbb{R}^N$ is a given function.

An *Equilibrium* for a given temperature T corresponding to $\dot{c}_n(t) = 0$ for $t > 0$, $n = 1, ..., N$, is characterized by the algebraic system of equations

$$\sum_{j=1}^{J} \alpha_{n,j} k_j(T) \prod_{m=1}^{N} c_m^{\nu_{m,j}} = 0 \quad n = 1, ..., N, \tag{259.3}$$

corresponding to the equation $f(u) = 0$.

EXAMPLE 259.1. The reaction

$$2\mathrm{NO} + \mathrm{Cl}_2 \to 2\mathrm{NOCl},$$

can be put in the form (259.1) with $A_1 = \mathrm{NO}$, $A_2 = \mathrm{Cl}_2$, $A_3 = \mathrm{NOCl}$, $N = 3$, $J = 1$, $\nu_{1,1} = 2$, $\nu_{2,1} = 1$, $\nu_{3,1} = 0$, $\lambda_{1,1} = 0$, $\lambda_{2,1} = 0$, $\lambda_{3,1} = 2$, $\alpha_{1,1} = -2$, $\alpha_{2,1} = -1$, and $\alpha_{3,1} = 2$.

EXAMPLE 259.2. The two reactions

$$2\mathrm{NO} + \mathrm{Cl}_2 \to^{k_1} 2\mathrm{NOCl},$$
$$2\mathrm{NOCl} \to^{k_2} 2\mathrm{NO} + \mathrm{Cl}_2,$$

can be put in the form (259.1) with $A_1 = \mathrm{NO}$, $A_2 = \mathrm{Cl}_2$, $A_3 = \mathrm{NOCl}$, $N = 3$, $J = 2$, $\nu_{1,1} = 2$, $\nu_{2,1} = 1$, $\nu_{3,1} = 0$, $\lambda_{1,1} = 0$, $\lambda_{2,1} = 0$, $\lambda_{3,1} = 2$, $\alpha_{1,1} = -2$, $\alpha_{2,1} = -1$, $\alpha_{3,1} = 2$, $\nu_{1,2} = 0$, $\nu_{2,2} = 0$, $\nu_{3,2} = 2$, $\lambda_{1,2} = 2$, $\lambda_{2,2} = 1$, $\lambda_{3,2} = 0$, $\alpha_{1,2} = 2$, $\alpha_{2,2} = 1$, and $\alpha_{3,2} = -2$. Equilibrium is characterized by

$$k_1 c_1^2 c_2 = k_2 c_3^2, \quad \text{or} \quad \frac{c_1^2 c_2}{c_3^2} = \frac{k_2}{k_1}.$$

EXAMPLE 259.3. An *ideal first order tank reactor* is modeled by the equation

$$qc^0 - Vkc = qc,$$

where c^0 is the reactant concentration at inflow, c is the concentration in the reactor, q is the inflow (= outflow) rate, V is the volume of the reactor and k is a reaction coefficient. The equation expresses that the (rate of) reactant inflow minus the reactant consumed in the reaction is equal to the reactant outflow. Introducing $\tau = \frac{V}{q}$, which is the time the reactant stays in the reactor, we get

$$c = \frac{c^0}{1 + \tau k}.$$

The *efficiency* of the reactor is given by

$$\eta = \frac{c^0 - c}{c^0} = \frac{\tau k}{1 + \tau k} = \frac{1}{1 + \frac{1}{\tau k}}.$$

We see in particular that the efficiency decreases as τ decreases.

EXAMPLE 259.4. An *ideal first order tube reactor* occupying the interval $(0,1)$, which may be viewed as a set of ideal first order tank reactors coupled in series, is modeled by

$$qc(x) - A\Delta x k c(x) = qc(x+\Delta x) \quad \text{for } 0 < x < 1,$$

where q is the (constant) flow rate, A the cross section of the tube, and Δx is a small increment in x. Dividing by Δx and letting Δx tend to zero leads to the initial value problem of finding the concentration $c(x)$ for $0 \le x \le 1$ such that

$$\frac{dc}{dx} = -\tau k c \quad \text{for } 0 < x \le 1, \quad c(0) = c^0,$$

where $\tau = \frac{A}{q}$. The solution is given by $c(x) = c^0 e^{-\tau k x}$, and the efficiency $\eta = \frac{c^0 - c(1)}{c^0} = 1 - e^{-\tau k}$. Using the fact that $\frac{x}{1+x} < 1 - e^{-x}$ for $x > 0$, it follows that the ideal tube reactor is more efficient than the ideal tank reactor.

259.2 Variable Temperature

Suppose now that the temperature $T(t)$ is variable with time t, and is unknown along with the concentrations $c_1(t), ..., c_N(t)$. The *heat of reaction* of reaction j is given by

$$(-\sum_{m=1}^{N} \alpha_{m,j} h_m) r_j,$$

where h_m is the *molar enthalpy* of species A_m. The heat of reaction is positive for an *exothermic reaction* and negative for an *endothermic reaction*.

The problem is now to find $c(t) = (c_1(t), ..., c_N(t))$ and $T(t)$ for $t > 0$ such that

$$\begin{cases} \dot{c}_n = \sum_{j=1}^{J} \alpha_{n,j} k_j(T) \prod_{m=1}^{N} c_m^{\nu_{m,j}}, & t > 0,\ n = 1, ..., N, \\ C_p \dot{T} = \sum_{j=1}^{J} (-\sum_{m=1}^{N} \alpha_{m,j} h_m) k_j(T) \prod_{m=1}^{N} c_j^{\nu_{m,j}}, & \\ c(0) = c^0,\ T(0) = T^0, & \end{cases} \tag{259.4}$$

where $c^0 = (c_1^0, ..., c_N^0)$ and T^0 are given initial concentrations and temperature, and C_p is the specific heat of the mixture of species.

259.3 Space Dependence

Adding spacial dependence in a domain Ω in $\mathbb{R}^3$, we are led to the following model: Find $c(x,t) = (c_1(x,t), ..., c_N(x,t))$ and $T(x,t)$ for $x \in \Omega$, $t > 0$,

such that

$$\begin{cases} \dot{c}_n + \nabla\cdot(c_n\beta) - \nabla\cdot(\epsilon_n\nabla c_n) \\ \qquad = \sum_{j=1}^{J}\alpha_{n,j}k_j(T)\prod_{m=1}^{N}c_m^{\nu_{m,j}} \quad \text{for } x\in\Omega,\ t>0,\ n=1,...,N, \\ C_p\dot{T} + \nabla\cdot(C_pT\beta) - \nabla\cdot(\epsilon_0\nabla T) \\ \qquad = \sum_{j=1}^{J}(-\sum_{m=1}^{N}\alpha_{m,j}h_m)k_j(T)\prod_{m=1}^{N}c_m^{\nu_{m,j}} \quad \text{for } x\in\Omega,\ t>0, \\ c(x,0)=c^0, \quad T(x,0)=T^0 \quad \text{for } x\in\Omega, \end{cases} \tag{259.5}$$

where $\beta(x,t)$ is a given convection velocity, and the ϵ_n are given diffusion coefficients. The system is complemented by boundary conditions of Dirichlet, Neumann or Robin type for each equation.

EXAMPLE 259.5. A stationary one species constant temperature first order reaction with constant diffusion and zero convection is modeled in dimensionless form by the equation

$$\Delta u = \varphi^2 u \quad \text{in } \Omega,$$

together with Dirichlet, Neumann or Robin boundary conditions, where φ is the *Thiele modulus*, and Ω is a domain in $\mathbb{R}^d$, $d=1,2,3$. A quantity of interest as a function of Ω, the reaction coefficient φ^2 and the boundary conditions, is the *total production* $\int_\Omega u(x)\,dx$.

EXAMPLE 259.6. A simple model for *flame propagation* in a channel takes the form

$$\begin{cases} \dot{u}_1 - \Delta u_1 + \beta_1\frac{\partial u_1}{\partial x_1} = u_2 f(u_1) & x\in\Omega,\ t>0, \\ \dot{u}_2 - \Delta u_2 + \beta_1\frac{\partial u_2}{\partial x_1} = -u_2 f(u_1) & x\in\Omega,\ t>0,\ x\in\Omega, \end{cases} \tag{259.6}$$

together with appropriate boundary conditions, where $\Omega=\mathbb{R}\times(0,1)$, u_1 represents temperature, u_2 represents a reactant concentration, β_1 is the velocity of the reactant in the x_1 direction, and $u_2f(u_1)$ represents a reaction rate with $f:\mathbb{R}^+\to\mathbb{R}^+$ given. With a proper choice of β_1 we may seek a stationary solution with $\dot{u}=0$ corresponding to a propagating flame front.

EXAMPLE 259.7. A basic model for *combustion* in a domain Ω in $\mathbb{R}^3$ takes the form: Find the concentration c and temperature T such that:

$$\begin{cases} \dot{c} - \epsilon_1\Delta c = -B_1 e^{-\frac{E}{RT}}c, & x\in\Omega,\ t>0, \\ \dot{T} - \epsilon_0\Delta T = B_0 e^{-\frac{E}{RT}}c & x\in\Omega,\ t>0, \end{cases} \tag{259.7}$$

together with, say, homogeneous Neumann boundary conditions, and with B_0 and B_1 positive constants. Depending on the activation energy E and initial conditions, the process may be fast or slow locally in space and time.

Axiom 1: All bodies are either in motion or at rest.
Axiom 2: Each single body can move at varying speeds.
Lemma 1: Bodies are distinguished from one another in respect of motion and rest, quickness and slowness, and not in respect of substance.
Lemma 2: All bodies agree in certain respects.
Lemma 3: A body in motion or at rest must have been determined to motion or rest by another body, which likewise has been determined to motion or rest by another body, and that body by another, and so ad infinitum.
...
Lemma 6: If certain bodies composing an individual thing are made to change the existing direction of their motion, but in such a way that they can continue their motion and keep the same mutual relation as before, the individual thing will likewise preserve the same mutual relation as before, the individual thing will likewise preserve its own nature without change of form. (Spinoza 1632-1677, Ethica II)

260
Meteorology and Coriolis Forces*

> Any teacher who stands up in front of a class and says that Coriolis force determines which way the water flows from a sink or bathtub, should not only read Fraser's Bad Coriolis Web page (www.ems.psu.edu/ fraser/Bad/BadCoriolis.html), but be required to copy it on the blackboard 100 times. (Jack Williams, USA TODAY)

260.1 Introduction

A common weather map shows the level curves of the air pressure p, the so-called *isobars*. Intuition might suggest that the wind will blow from high pressure to low pressure, i.e. in the opposite direction to the pressure gradient ∇p and orthogonal to the isobars. However, this turns out to be completely false. In fact, the wind circles around a center of low pressure in a counter-clockwise direction on the North hemisphere and in a clockwise direction on the Southern hemisphere, and in the opposite directions around centers of high pressure. Thus the wind blows along the isobars, instead of orthogonal to the isobars. This fact is well-known to sailors, making it possible to easily and accurately predict the wind direction if the centers of the low and high pressures are known. The reason is that the Earth is rotating, which creates a force of acceleration called the *Coriolis force*. This causes the wind to deviate to the right on the Northern hemisphere and to the left on the Southern hemisphere (away from the equator). The effect is that the wind circles around a center of low pressure in a counter-clockwise

direction on the Northern hemisphere, as any weather map in a newspaper indicates. The Coriolis force is felt on a turn-around when seeking to change position in the radial direction, which causes an (unexpected) force in the tangential direction.

260.2 A Basic Meteorological Model

We shall now derive a simple model for the motion of the atmosphere, which predicts that the wind should revolve around centers of low and high pressure. The model takes the form

$$\nabla p = \rho 2\omega \times v, \tag{260.1}$$

where p is the pressure, v is the wind velocity, $\omega \in \mathbb{R}^3$ is the angular velocity of the Earth, and ρ is the density of the atmosphere. The quantity $2\omega \times v$ is an approximation of the Coriolis acceleration and the equation $\nabla p = \rho 2\omega \times v$ gives a balance of the pressure force ∇p and the Coriolis force $\rho 2\omega \times v$. Here ∇p represents the gradient in the plane of the surface of the Earth and the model applies to "caps" on the Northern or Southern away from the Equator, say above or below the 60 degree latitude, where we can approximate the surface of the Earth by a flat disc, see Fig. 260.1, that is, the "world" of the sailor and the wind is a big flat turn-around.

We see that (260.1) states that ∇p is orthogonal to the direction of the wind. If we know p, we can determine the wind direction and speed from (260.1).

260.3 Rotating Coordinate Systems and Coriolis Acceleration

To derive the expression $2\omega \times v$ of the Coriolis force, we need to study coordinate transformations from one fixed coordinate system to a rotating coordinate system. We thus let $\{e_1, e_2, e_3\}$ be a fixed orthonormal reference coordinate system for $\mathbb{R}^3$, and we let $\{\bar{e}_1, \bar{e}_2, \bar{e}_3\}$ be another orthonormal coordinate system with the same origin, which rotates around the fixed vector $\omega \in \mathbb{R}^3$ with the angular speed $\|\omega\|$. More precisely, if $x(t) = x_1(t)e_1 + x_2(t)e_2 + x_3(t)e_3$ are the reference coordinates of a fixed point in the rotating coordinate system, then according to Fig. 260.2 we have

$$\frac{dx}{dt} = \omega \times x, \tag{260.2}$$

since $\frac{dx}{dt}$ is perpendicular to both ω and x, and $\|\frac{dx}{dt}\| = \|\omega\|\|x\|\sin(\theta)$, where $\theta \in [0, \pi]$ is the angle between ω and x. In particular we have for the basis

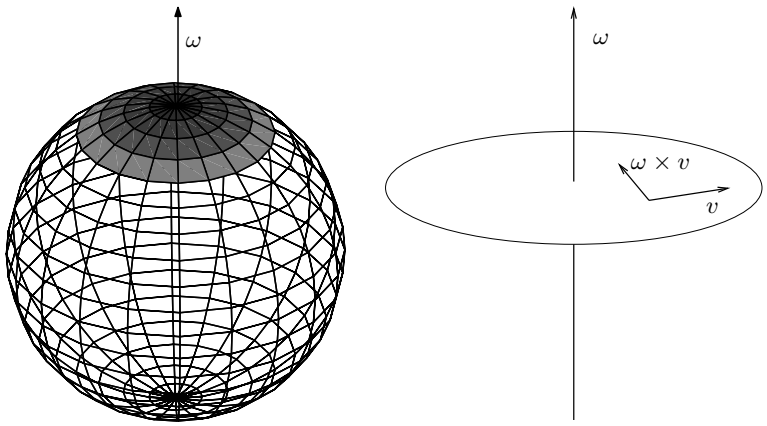


FIGURE 260.1. Northern Hemisphere **Change w to** ω

vectors of the moving coordinate system

$$\frac{d\bar{e}_i}{dt} = \omega \times \bar{e}_i, \quad i = 1, 2, 3. \tag{260.3}$$

Consider now a moving point with coordinates $x(t)$ in the fixed reference system and coordinates $\bar{x}(t)$ in the rotating system, so that

$$x(t) = x_1(t)e_1 + x_2(t)e_2 + x_3(t)e_3,$$
$$\bar{x}(t) = \bar{x}_1(t)\bar{e}_1(t) + \bar{x}_2(t)\bar{e}_2(t) + \bar{x}_3(t)\bar{e}_3(t),$$

and of course $x(t) = \bar{x}(t)$. In particular, we may this way seek the coordinates of the basis vectors $\bar{e}_i(t)$ in the fixed system $\{e_1, e_2, e_3\}$. We now compute the velocity $\frac{dx}{dt}$ by differentiating $x(t) = \bar{x}(t)$ with respect to t to get

$$\begin{aligned}\frac{dx}{dt} &= \frac{d}{dt}\bar{x}(t) = \frac{d\bar{x}_1}{dt}\bar{e}_1 + \frac{d\bar{x}_2}{dt}\bar{e}_2 + \frac{d\bar{x}_3}{dt}\bar{e}_3 + \bar{x}_1\frac{d\bar{e}_1}{dt} + \bar{x}_2\frac{d\bar{e}_2}{dt} + \bar{x}_3\frac{d\bar{e}_3}{dt} \\ &= \frac{d\bar{x}_1}{dt}\bar{e}_1 + \frac{d\bar{x}_2}{dt}\bar{e}_2 + \frac{d\bar{x}_3}{dt}\bar{e}_3 + \bar{x}_1(\omega \times \bar{e}_1) + \bar{x}_2(\omega \times \bar{e}_2) + \bar{x}_3(\omega \times \bar{e}_3),\end{aligned}$$

where we used (260.3). We can write this expression as

$$\frac{dx}{dt} = \frac{\bar{d}\bar{x}}{dt} + \omega \times x, \tag{260.4}$$

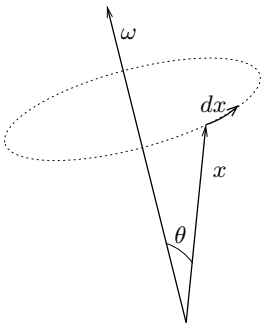


FIGURE 260.2. A vector x rotating with angular velocity ω.**Change w to ω and q to θ**

if we agree to write

$$\frac{\bar{d}\bar{x}}{dt} = \frac{d\bar{x}_1}{dt}\bar{e}_1 + \frac{d\bar{x}_2}{dt}\bar{e}_2 + \frac{d\bar{x}_3}{dt}\bar{e}_3.$$

The velocity of $x(t) = \bar{x}(t)$ in the fixed reference system is $\frac{dx}{dt}$, while $\frac{\bar{d}\bar{x}}{dt}$ is the velocity vs the rotating system involving the derivatives $\frac{\bar{d}}{dt}\bar{x}_i(t)$. In particular, if the point is fixed in the rotating system so that $\frac{\bar{d}\bar{x}}{dt} = 0$, then we retrieve (260.2) and (260.3).

We now seek a corresponding formula for the accelerations. We differentiate with respect to t once more, and using (260.4) with x replaced by $\frac{\bar{d}\bar{x}}{dt}$, we get

$$\frac{d^2x}{dt^2} = \frac{d}{dt}\big(\frac{\bar{d}\bar{x}}{dt} + \omega \times x\big) = \frac{d}{dt}\big(\frac{\bar{d}\bar{x}}{dt}\big) + \omega \times \frac{dx}{dt}$$
$$= \frac{\bar{d}}{dt}\big(\frac{\bar{d}\bar{x}}{dt}\big) + \omega \times \frac{\bar{d}\bar{x}}{dt} + \omega \times \big(\frac{\bar{d}\bar{x}}{dt} + \omega \times x\big).$$

We can write this as

$$\frac{d^2x}{dt^2} = \frac{\bar{d}^2\bar{x}}{dt^2} + 2\omega \times \frac{\bar{d}\bar{x}}{dt} + \omega \times \big(\omega \times x\big). \tag{260.5}$$

Here, $\omega \times \big(\omega \times x\big)$ represents the *centripetal acceleration* and $2\omega \times \frac{\bar{d}\bar{x}}{dt}$ the *Coriolis acceleration*, and $\frac{d^2x}{dt^2}$ is the acceleration vs the reference system and $\frac{\bar{d}^2\bar{x}}{dt^2}$ the acceleration vs the rotating system.

By Newton's Law $F = ma$, *acceleration* is directly coupled to *force*, and thus both the centripetal and the Coriolis acceleration show up as forces in the fixed reference system. Both these forces in fact have a somewhat mysterious character; we have through massive daily experience become quite familiar with the centripetal acceleration, while the Coriolis force still presents surprises to most of us.

If the rotation speed $\|\omega\|$ is relatively small, then we can neglect the centripetal acceleration and we get

$$\frac{d^2x}{dt^2} \approx \frac{\bar{d}^2\bar{x}}{dt^2} + 2\omega \times \frac{\bar{d}\bar{x}}{dt}, \tag{260.6}$$

which leads to the model (260.1). Note that we use the rotating coordinate system in our "world", and thus $\frac{\bar{d}\bar{x}}{dt}$ is the relevant velocity.

Chapter 260 Problems

260.1. Motivate (260.1) using (260.6).

260.2. Inspect the isobars of a weather map and compute wind direction from (260.1) and compare with the wind direction of the map.

260.3. Study the effect of the Coriolis acceleration at the Equator.

260.4. Show that the centripetal acceleration of a body moving in a circle with radius r with speed v is equal to $\frac{v^2}{r}$.

260.5. The Gulf Stream is the reason Scandinavia is not deep frozen like Alaska. Explain why the Gulf Stream bends over from North America to North Europe.

260.6. Consider a car driving East-West along a certain latitude. At what speed is the Coriolis force on the car of the same size as the centripetal force? Determine this speed as a function of the latitude and find out at which latitudes the minimum and maximum is attained.

260.7. A bucket of water is spinning around its center with angular velocity ω. What is the shape of the water surface?

260.8. A pendulum of length l swings back and forth once every period of length $t = \sqrt{l/g}$, where g is the acceleration of gravity. Compute the Coriolis force on the pendulum at latitude θ (i.e. at an angle θ from the equator). This Coriolis force makes the plane in which the pendulum swings rotate, i.e. if the pendulum swings north–south at one instant, it will later swing west–east. Find the time T after which the pendulum swings in the initial direction once again as function of the latitude. What is the period on you latitude?

261
The Crash Model*

On October 24, 1929, people began selling their stocks as fast as they could. Sell orders flooded market exchanges. On a normal day, only 750-800 members of the New York Stock Exchange started the Exchange. However, there were 1100 members on the floor for the morning opening. Furthermore, the Exchange directed all employees to be on the floor since there were numerous margin calls and sell orders placed overnight and extra telephone staff was arranged at the members' boxes around the floor. The Dow Jones Industrial Index closed at 299 that day. October 29 was the beginning of the Crash. Within the first few hours the stock market was open, prices fell so far as to wipe out all the gains that had been made in the previous year. The Dow Jones Industrial Index closed at 230. Since the stock market was viewed as the chief indicator of the American economy, public confidence was shattered. Between October 29 and November 13 (when stock prices hit their lowest point) over $30 billion disappeared from the American economy. It took nearly twenty-five years for many stocks to recover. (www.arts.unimelb.edu.au/amu/ucr/student/1997/Yee/1929.htm)

261.1 Introduction

Why did the Wall fall on November 9 1989? Why did the Soviet Union dissolve in January 1992? Why did the Stock market collapse in October 1929 and 1987? Why did Peter and Mary break up last Fall after 35 years of marriage? What caused the September 11 attack? Why does the flow in

the river go from orderly laminar to chaotic turbulent at a certain specific point? All the situations behind these questions share a common feature: Nothing particularly dramatic preceded the sudden transition from stable to unstable, and in each case the rapid and dramatic change away from normality came as big surprise to almost everyone.

We now describe a simple mathematical model that shows the same behavior: the solution stays almost constant for a long time and then quite suddenly the solution explodes.

We consider the following initial value problem for a system of two ordinary differential equations: find $u(t) = (u_1(t), u_2(t))$ such that

$$\begin{cases} \dot{u}_1 + \epsilon u_1 - \lambda u_2 u_1 = \epsilon & t > 0, \\ \dot{u}_2 + 2\epsilon u_2 - \epsilon u_1 u_2 = 0 & t > 0, \\ u_1(0) = 1, u_2(0) = \kappa\epsilon, \end{cases} \tag{261.1}$$

where ϵ is a small positive constant of size say 10^{-2} or smaller and λ and κ are positive parameters of moderate size ≈ 1. If $\kappa = 0$, then the solution $u(t) = (1, 0)$ is constant in time, which we view as the *base solution*. In general, for $\kappa > 0$, we think of $u_1(t)$ as a primary part of solution with initial value $u_1(0) = 1$, and $u_2(t)$ as a small secondary part with an initial value $u_2(0) = \kappa\epsilon$ that is small because ϵ is small. Both components $u_1(t)$ and $u_2(t)$ will correspond to physical quantities that are non-negative and $u_1(0) = 1$ and $u_2(0) = \kappa\epsilon \geq 0$.

261.2 The Simplified Growth Model

The system (261.1) models an interaction between a primary quantity $u_1(t)$ and a secondary quantity $u_2(t)$ through the terms $-\lambda u_1 u_2$ and $-\epsilon u_2 u_1$. If we keep just these terms, we get a simplified system of the form

$$\begin{cases} \dot{w}_1(t) = \lambda w_1(t) w_2(t) & t > 0, \\ \dot{w}_2(t) = \epsilon w_2(t) w_1(t) & t > 0, \\ w_1(0) = 1, \quad w_2(0) = \kappa\epsilon. \end{cases} \tag{261.2}$$

We see that the coupling terms are *growth terms* in the sense that both the equation $\dot{w}_1(t) = \lambda w_1(t) w_2(t)$ and $\dot{w}_2(t) = \epsilon w_2(t) w_1(t)$ say that $\dot{w}_1(t)$ and $\dot{w}_2(t)$ are positive if $w_1(t) w_2(t) > 0$. In fact, the system (261.1) always blow up for $\kappa > 0$ because the two components propel each other to infinity as t increases in the sense that the right hand sides get bigger with $w_1(t) w_2(t)$ and this increases the growth rates $\dot{w}_1(t)$ and $\dot{w}_2(t)$, which in turn makes $w_1 w_2(t)$ even bigger, and so on towards blow up, see Fig. 261.1.

We can study the blow up in (261.2) analytically assuming for simplicity that $\lambda = \kappa = 1$. In this case, it turns out that the two components $w_1(t)$ and

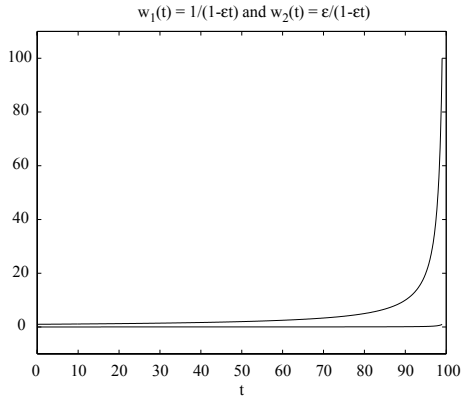


FIGURE 261.1. Solution of simplified growth model

$w_2(t)$ for all t are coupled by the relation $w_2(t) = \epsilon w_1(t)$, that is $w_2(t)$ is always the same multiple of $w_1(t)$. We check this statement by first verifying that $w_2(0) = \epsilon w_1(0)$ and then by dividing the two equations to see that $\dot{w}_2(t)/\dot{w}_1(t) = \epsilon$. So, $\dot{w}_2(t) = \epsilon \dot{w}_1(t)$, that is $w_2(t) - w_2(0) = \epsilon w_1(t) - \epsilon w_2(0)$, and we get the desired conclusion $w_2(t) = \epsilon w_1(t)$ for $t > 0$. Inserting this relation into the first equation of (261.2), we get

$$\dot{w}_1(t) = \epsilon w_1^2(t) \qquad \text{for } t > 0,$$

which can be written as

$$-\frac{d}{dt}\frac{1}{w_1(t)} = \epsilon \qquad \text{for } t > 0.$$

Recalling the initial condition $w_1(0) = 1$, we get

$$-\frac{1}{w_1(t)} = \epsilon t - 1 \qquad \text{for } t \geq 0,$$

which gives the following solution formula in the case $\lambda = \kappa = 1$:

$$w_1(t) = \frac{1}{1-\epsilon t}, \quad w_2(t) = \frac{\epsilon}{1-\epsilon t} \qquad \text{for } t \geq 0. \tag{261.3}$$

This formula shows that the solution tends to infinity as t increases towards $1/\epsilon$, that is, the solution explodes at $t = 1/\epsilon$. We notice that the time of blow up is $1/\epsilon$, and that the *time scale* before the solution starts to increase noticeably, is of size $\frac{1}{2\epsilon}$, which is a long time since ϵ is small. Thus, the solution changes very slowly for a long time and then eventually blows up quite a bit more rapidly, see Fig. 261.1.

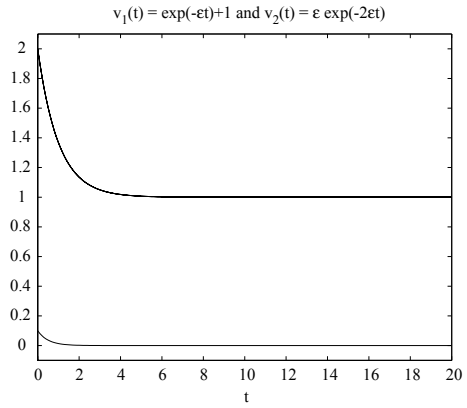


FIGURE 261.2. Solution of simplified decay model

261.3 The Simplified Decay Model

On the other hand, if we forget about the growth terms, we get another simplified system:

$$\begin{cases} \dot{v}_1 + \epsilon v_1 = \epsilon & t > 0, \qquad t > 0, \\ \dot{v}_2 + 2\epsilon v_2 = 0 & \qquad\quad t > 0, \\ v_1(0) = 1 + \delta, & v_2(0) = \kappa\epsilon, \end{cases} \tag{261.4}$$

where we have also introduced a small perturbation δ in $v_1(0)$. Here the two terms ϵv_1 and $2\epsilon v_2$ are so called *dissipative* terms that cause the solution $v(t)$ to return to the base solution $(1, 0)$ regardless of the perturbation, see Fig. 261.2. This is clear in the equation $\dot{v}_2 + 2\epsilon v_2 = 0$ with solution $v_2(t) = v_2(0)\exp(-2\epsilon t)$, which decays to zero as t increases. Rewriting the equation $\dot{v}_1 + \epsilon v_1 = \epsilon$ as $\dot{V}_1 + \epsilon V_1 = 0$, setting $V_1 = v_1 - 1 = \exp(-\epsilon t)$, we find that $v_1(t) = \delta\exp(-\epsilon t) + 1$, and thus $v_1(t)$ approaches 1 as t increases. We summarize: the solution $(v_1(t), v_2(t))$ of (261.4) satisfies

$$v_1(t) = \delta\exp(-\epsilon t) + 1 \to 1, \quad v_2(t) = \kappa\epsilon\exp(-2\epsilon t) \to 0, \quad \text{as } t \to \infty.$$

We say that (261.4) is a *stable* system because the solution always returns from $(1+\delta, \kappa\epsilon)$ to the base solution $(1, 0)$ independently of the perturbation $(\delta, \kappa\epsilon)$ of $(v_1(0), v_2(0)$.

We note that the time scale is again of size $1/\epsilon$, because of the presence of the factors $\exp(-\epsilon t)$ and $\exp(-2\epsilon t)$.

261.4 The Full Model

We can now sum up: The real system (261.1) is a combination of the unstable system (261.2) that includes only the growth terms only and whose the solution always blows up, and the stable system (261.4) that excludes the growth terms. We shall see that depending on the size of $\lambda\kappa$ the unstable or stable feature will take over. In Fig. 261.3 and Fig. 261.4, we show different solutions for different values of the parameters λ and κ with different initial values $u(0) = (u_1(0), u_2(0)) = (1, \kappa\epsilon)$. We see that if $\lambda\kappa$ is sufficiently large, then the solution $u(t)$ eventually blows up after a time of size $1/\epsilon$, while if $\lambda\kappa$ is sufficiently small, then the solution $u(t)$ returns to the base solution $(1, 0)$ as t tends to infinity.

Thus, there seems to be a *threshold value* for $\lambda\kappa$ above which the initially disturbed solution eventually blows up and below which the initially disturbed solution returns to the base solution. We can view κ as a measure of the size if the initial disturbance, because $u_2(0) = \kappa\epsilon$. Further, we can view the factor λ as a quantitative measure of the *coupling* between the growth components $u_2(t)$ and $u_1(t)$ through the growth term $\lambda u_1 u_2$ in the evolution equation for u_1.

Our main conclusion is that if the initial disturbance times the coupling is sufficiently large, then the system will blow up. Blow up thus requires both the initial disturbance and the coupling to be sufficiently large. A large initial disturbance will not cause blow up unless there some coupling. A strong coupling will not cause blow up unless there is an initial disturbance.

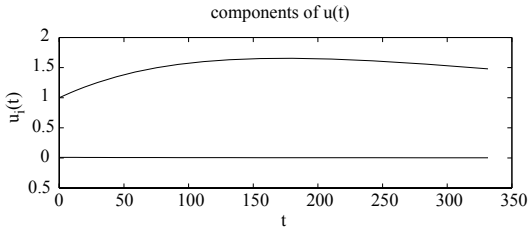


FIGURE 261.3. Return to the base solution if $\lambda\kappa$ is small enough

We now investigate the qualitative behavior of (261.1) in a little more detail. We see that $\dot{u_1}(0)/u_1(0) = \lambda\kappa\epsilon$, while $\dot{u_2}(0)/u_2(0) = -\epsilon$, which shows that initially $u_1(t)$ grows and $u_2(t)$ decays at relative rates of size ϵ. Now, $u_1(t)$ will continue to grow as long as $\lambda u_2(t) > \epsilon$, and further $u_2(t)$ will start to grow as soon as $u_1(t) > 2$. Thus, if $u_1(t)$ manages to become larger than 2, before $u_2(t)$ has decayed below ϵ/λ, then both components will propel each other to cause a blow up to infinity. This happens if $\lambda\kappa$ is above a certain threshold.

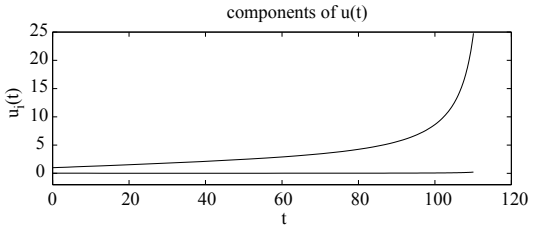


FIGURE 261.4. Blow up if $\lambda\kappa$ is large enough

We notice that the time scale for significant changes in both u_1 and u_2 is of size ϵ^{-1}, because the growth rates are of size ϵ. This conforms with the experience from the simplified models. The scenario is thus that the primary part $u_1(t)$ grows slowly starting from 1 at a rate of size ϵ and the secondary part $u_2(t)$ decays slowly at a rate of size ϵ^2, over a time of size $1/\epsilon$. If $\lambda\kappa$ is above a certain threshold, then $u_1(t)$ reaches the value 2, at which point $u_2(t)$ starts to grow and eventually blow up follows on a somewhat shorter time scale. If $u_1(t)$ does not reach the value 2 in time, then $(u_1(t), u_2(t))$ returns to the base solution $(1, 0)$ as t increases.

We hope the presented scenario is quite easy to grasp intuitively, and conforms with every-day experiences of quit sudden blow-up, as a result of an accumulation of small events over a long period.

We can give the Crash model very many interpretations in real life, such as

- stock market (u_1 stock prize of big company, u_2 stock prize of small innovative company),
- chemical reaction (u_1 main reactant, u_2 catalyst),
- marriage crisis (u_1 main discontent, u_2 small irritation factor),
- spread of infection (u_1 infected people, u_2 amount of germs),
- symbiosis (u_1 main organism, u_2 small parasite),
- population model (u_1 rabbits, u_2 vitalizing carrots),

and many others.

In particular, the model describes an essential aspect of the process of transition from laminar to turbulent flow in for example a pipe. In this case u_1 represents a flow component in the direction of the pipe and u_2 represents a small perturbation of the flow in the transversal direction. The time to explosion corresponds to the time it takes for the flow to go turbulent starting as laminar flow at the inlet. In the famous experiment of Reynolds from 1888, ink is injected at the inlet of a transparent pipe and the observer can follow the streamline traced by the ink, which forms

a straight line in the laminar part and then successively becomes more and more wavy until it breaks down to completely turbulent flow at some distance from the inlet. The distance to breakdown varies with the flow speed and viscosity and perturbations resulting from e.g. roughness of the surface of the pipe or a heavy-weight truck passing by at some distance from the experimental set-up.

Chapter 261 Problems

261.1. Develop the indicated applications of the Crash model.

261.2. Solve the full system (261.1) numerically for various values of λ and κ and try to pin down the threshold value of $\lambda\kappa$.

261.3. Develop a *Theory of Capitalism* based on (261.1) as a simple model of the economy in a society, with u_1 representing the value of a basic resource like land, and u_2 some venture capital related to the exploitation of new technology, with $(1,0)$ a base solution without the new technology, and with the coefficient λ of the u_1u_2 term in the first equation representing the positive interplay between base and new technology, and the terms ϵu_i representing stabilizing effects of taxes for example. Show that the possible pay-off $u_1(t) - u_1(0)$ of a small investment $u_2(0) = \kappa\epsilon$ may be large, and that an exploding economy may result if $\lambda\kappa$ is large enough. Show that no growth is possible if $\lambda = 0$. Draw some conclusions from the model coupled to for example the role of the interest rate for controlling the economy.

261.4. Interpret (261.1) as a simple model of a stock market with two stocks, and discuss scenarios of overheating. Extend to a model for the world stock market, and predict the next crash.

261.5. Consider the linear model

$$\begin{aligned} \dot\varphi_1 + \epsilon\varphi_1 - \lambda\varphi_2 &= 0 \quad t > 0, \\ \dot\varphi_2 + \epsilon\varphi_2 &= 0 \quad t > 0, \\ \varphi_1(0) = 0, \quad \varphi_2(0) &= \kappa\epsilon, \end{aligned} \tag{261.5}$$

which is obtained from (261.1) by setting $\varphi_1 = u_1 - 1$ and $\varphi_2 = u_2$ and replacing $u_1\varphi_2$ by φ_2 assuming u_1 is close to 1. Show that the solution of (261.5) is given by

$$\varphi_2(t) = \kappa\epsilon\exp(-\epsilon t), \quad \varphi_1(t) = \lambda\kappa\epsilon t\exp(-\epsilon t).$$

Conclude that

$$\frac{\varphi_1(\frac{1}{\epsilon})}{\varphi_2(0)} = \lambda\frac{\exp(-1)}{\epsilon},$$

and make an interpretation of this result.

261.6. Expand the Crash model (261.1) to

$$\begin{aligned} \dot{u}_1 + \epsilon u_1 - \lambda u_1 u_2 + \mu_1 u_2^2 &= \epsilon \qquad t > 0, \\ \dot{u}_2 + 2\epsilon u_2 - \epsilon u_2 u_1 + \mu_2 u_1^2 &= 0 \qquad t > 0, \\ u_1(0) = 1, \quad u_2(0) &= \kappa\epsilon, \end{aligned}$$

with decay terms $\mu_1 u_2^2$ and $\mu_2 u_1^2$, where μ_1 and μ_2 are positive coefficients. (a) Study the stabilizing effect of such terms numerically. (b) Seek to find values of μ_1 and μ_2, so that the corresponding solution starting close to $(1, 0)$ shows an intermittent behavior with repeated periods of blow up followed by a decay back to a neighborhood of $(1, 0)$. (c) Try to find values of μ_1 and μ_2 so that multiplication of the first equation with a positive multiple of u_1 and the second by u_2, leads to bounds on $|\epsilon u_1(t)|^2$ and $|u_2(t)|^2$ in terms of initial data. Hint: Try for example $\mu_1 \approx 1/\epsilon$, and $\mu_2 \approx \epsilon^2$.

261.7. Study the initial value problem $\dot{u} = f(u)$ for $t > 0$, $u(0) = 0$, where $f(u) = \lambda u - u^3$, with different values of $\lambda \in \mathbb{R}$. Relate the time-behavior of $u(t)$ to the set of solutions $\bar{u}$ of $f(u) = 0$, that is, $\bar{u} = 0$ if $\lambda \leq 0$, and $\bar{u} = 0$ or $\bar{u} = \pm\sqrt{\lambda}$ if $\lambda > 0$. Study the linearized models $\dot{\varphi} - \lambda\varphi + 3\bar{u}^2\varphi = 0$ for the different $\bar{u}$. Study the behavior of the solution assuming $\lambda(t) = t - 1$.

261.8. Study the model

$$\begin{aligned} \dot{w}_1 + w_1 w_2 + \epsilon w_1 &= 0, \qquad t > 0, \\ \dot{w}_2 - \epsilon w_1^2 + \epsilon w_2 &= -\gamma\epsilon, \qquad t > 0, \end{aligned} \tag{261.6}$$

with given initial data $w(0)$, where γ is a parameter and $\epsilon > 0$. This problem admits the stationary "trivial branch" solution $\bar{w} = (0, -\gamma)$ for all γ. If $\gamma > \epsilon$, then also $\bar{w} = (\pm\sqrt{\gamma - \epsilon}, -\epsilon)$ is a stationary solution. Study the evolution of the solution for different values of γ. Study the corresponding linearized problem, linearized at $\bar{w}$.