Conquest of the Plane

Using The Economics Pack *Applications of Mathematica* for a didactic primer on Analytic Geometry and Calculus

Thomas Colignatus, March 2011

http://thomascool.eu

Applications of *Mathematica*

Thomas Colignatus is the name of Thomas Cool in science.

Published by Cool, T. (Consultancy & Econometrics)

Rotterdamsestraat 69, NL-2586 GH Scheveningen, The Netherlands

http://thomascool.eu

NUR 120, 846; JEL A00; MSC2010 97D01, 28A01

ISBN 978-90-804774-6-9

Mathematica is a registered trademark of Wolfram Research, Inc.

This book uses version 8.0.1

A primer with objectives

This book is a **primer,** and thus not a textbook, but it has a **textbook format**. It has different functions for different groups: students, teachers and the professions that apply mathematics.

Aims of this book when you are new to the subject

When you finish this book:

- You will better understand the major topics in analytic geometry and calculus of the Euclidean plane.
- You can better check whether inferences in those subjects are valid or not, and tell why.
- You can better understand arguments supported by geometry and analysis.
- You have had a fast track course and can oversee the landscape. It is not likely that you can construct proofs yet but you know better in what area to specialize.
- You can read this book *as it is*, thus also without *Mathematica*. Without ever running a program, you will still benefit from the discussion. However, if you have a computer and practice with the programs, then you end up being able to run the routines in this book and interprete their results.

Note that the software can be downloaded freely from the internet and be inspected; however, if you want to run it then you need a licence.

Aims of this book when you are an advanced reader

When you finish this book:

- You will even better understand what the key assumptions are and how it all fits together.
- You will refocus your research towards issues that matter more.
- You will be an advanced reader but you might lack in balance between either the math or the history and philosophy: you will have enhanced the balance.
- One of the aims of this book concerns new readers. You will be able to discuss and teach the subject in this manner for students as well.
- You will better appreciate *Elegance with Substance* by the same author (2009).

Abstract

- The book is *primer on analytic geometry and calculus.* A primer is not a textbook but this is a primer in textbook format. It works from the novice level up to where you could proceed with n-dimensionality. The first four parts have been written with students in mind. The fifth part explains the didactics to teachers and students in didactics of mathematics.
- New is the integral attention on didactics, the logical order, the didactic naming of lines, the notion of the dynamic quotient and the development of calculus from algebra, the definition of angles on the unit circumference circle, x_{ur} and y_{ur} on the unit radius circle, the integrated use of $\Theta = 2\ \pi$, the recovered exponent rex as a better term for the logarithm. (See Chapter 16 for The News.)
- The chapters have a direct hands-on approach so that you can directly learn from applying routines. This, and the sense of achievement, should stimulate you to continue, while it also provides a basis to reflect on what already has been learned.
- The chapters build up in logical order and provide theory on the way. When something is introduced it directly makes sense, there is no waiting for some unfulfilled promise.
- There are many formulas but there is no formal axiomatic development.
- The didactics are guided by the Van Hiele levels (Chapter 15) and we reject Freudenthal's "realistic math". You have sufficient experience with the plane since making drawings in kindergarten. When you think about a triangle it is as abstract as it can get because such thought is abstract by nature. What counts are the lingering notions in this abstract imagination that have to be activated. It can distract and confuse when mental clarification is mixed with the application to reality. In this book, geometry is treated at the Van Hiele base level and from there we proceed to analytic geometry.
- Application is relevant but should be dosed wisely. Examples are given from physics, economics and statistics.
- Discussed are co-ordinates, lines, circles, vectors, complex numbers, projection, systems of equations, trigonometry, parabola, the exponential number, Euler's form, calculus and a short section on non-Euclidean geometry. When required the routines in *Mathematica* are explained.

- This is a textbook without exercises. The idea is that *Mathematica* provides an interactive environment, that such exercises can be found in abundance on the internet and that *Mathematica* can help to solve those. What you learn here helps you to find those exercises elsewhere.
- Programs in the environment and language of *Mathematica* support the discussion. Download *The Economics Pack* from http://thomascool.eu, install and evaluate:

```
Needs["Economics`Pack`"]

ResetAll

Economics[Math`AnalyticGeometry, Math`Trigonometry, Math`Geometry, Math`Pythagoras,
Calculus, Physics, Taxes, Survival]
```

Note: On the palette for The Economics Pack there is a button for the User Guide. Click there and you will find the entire text of this book available there, also for evaluation in *Mathematica*.

Note: You can read this book *as it is,* thus also without *Mathematica*. Without ever running a program, you will still benefit from the discussion.

Note: See the internet for other programs in analytic geometry and calculus.

See D. L. Vossler "Exploring Analytic Geometry with *Mathematica*" at http://www.descarta2d.com/

- For the professions that apply mathematics like physics, engineering, biology, economics and evidence based medicine, this book provides documentation to judge on the proposal that national parliaments look into mathematics education, as explained in the book *Elegance with Substance* by the same author (2009). This very discussion is not mentioned in the body of this present book except in the Preface and the Conclusion at the end.

Keywords

Analytic geometry, analytical geometry, calculus, dynamic quotient, trigonometry, Θ, Xur, Yur, UMA, foundations of mathematics, mathematics education, didactics of mathematics, teaching of mathematics, epistemology, methodology of science, general philosophy, general economics

Preface

Aims and intentions

Mathematics can be liberating, fun, enlightening and empowering. A population well educated in mathematics will prosper and will have a bedrock foundation for democracy. Mathematics can also be taught badly, as strict, arcane and depending upon authority, and when you do not understand something then you do not belong to the class of the initiated and those who do understand. Egypt with its pyramids had a class of geometers who closely guarded their secrets and Sumer with its astronomers likewise. Since the Greeks mankind does much better. Euclid codifies geometry as a lawmaker but there is also the spirit of research in the laws of nature. The crucial idea that mathematics is respectful engagement in mutual discussion still has to sink in though. Current courses in mathematics are needlessly cumbersome and a barrier towards understanding, with rote training substituting for better didactics. With a better didactic approach more people will understand math and more people will see its fun. It will be greatly beneficial for society when more people - and even mathematicians - can take mathematics as it really is. This essentially means a need for re-engineering math and its education. My book *Elegance with Substance* of 2009 explains that the mathematical discipline is not up to the challenge so that society, parents, the applied sciences and teachers of physics, engineering, economics, biology, but also English and music, should take the lead and put the matter to institutes of government. EwS contains a shopping list of many points but a reader may wonder what it adds up to. The best way to show that improvement is possible is by actually doing it. *Conquest of the Plane* does so and fills in the blanks. It has the layout of a textbook for a math course. It has been written with students in mind so that they can directly benefit. Exercises are lacking though. The fifth part gives the didactic foundation for teachers. This book thus is a **primer**. A primer is not a textbook but this is a primer in textbook format. When students and teachers all over the world start using the course much of the mission will be accomplished. Yet, will this happen, with math teachers locked in tradition ? Not likely. Eventually this book can indeed be used as the example textbook that it is but for the first years it will primarily be a companion to *Elegance with Substance.* Readers interested in how mathematics could be re-engineered both as a subject and as a discipline are referred to *Elegance with Substance* - and see also the Conclusion at the very end. Then *Conquest of the Plane* is an existence proof that math can be improved indeed.

Intended readership

Given what is currently taught at the 3rd year of advanced highschool the course in this book might be started there ... However, the use of language is not adapted to that 3rd year yet, and exercises at that level are much lacking. It might succeed though if the teachers put effort into the experiment. Careful as I am myself, my intended readership has been as follows.

If you are a **student** then this book shows you what you should have been taught from 3rd year onwards. The book then provides a fast track (refresher) course. If you are a **parent** helping your kid with math then this book wil provide invaluable support. If you are a **teacher** then there are new insights both in didactics and mathematics itself. The book provides material to discuss with colleagues and ideas for in class and you might use the book as a textbook indeed for some classes like Summer school or a refresher course. If you are a student of the **didactics of mathematics** then you will see key ideas and see those presented and built up in textbook fashion so that there will be no confusion as to what they mean in practice.

Readership	*Student*	*Teacher*
Mathematics itself	(1) A course from 3 rd year of advanced highschool up to and including college freshmen (2) Support for parents	New insights like the definition of an angle, trigonometry and the derivation of calculus
Didactics of mathematics	(1) Key ideas on didactics (2) Built up in textbook manner	(1) Key ideas on didactics (2) Ideas for in class (3) A textbook for some.

This book is written in *Mathematica* which reduces the tedium of calculation. The routines for this book have been included in Cool (1999, 2001), *The Economics Pack*, that is, the update is in the software but not in the manual since that is given here. You can use this present book without running those programs. But if you have these programs available, then you can have a hands-on experience, verify the conclusions, try your own cases, and, write your higher-level programs.

My background

As a student in econometrics I participated in the mathematics courses for students of mathematics, physics and astronomy. Graduation in 1982 in Groningen gave only a teaching certificate for economics but this was resolved in 2008 with a MSc Teacher of mathematics in Leiden. The book on logic that I wrote as a student was eventually published as *A Logic of Exceptions* in 2007. The didactic study on mathematics resulted in *Elegance with Substance* in 2009. I am keenly interested in economics but for teaching I prefer mathematics. I have not used the new ideas here in class since they are not in the official program. Though I should be writing about the current economic crisis there is ample reason to compose a book on analytic geometry and calculus. A somewhat awkward point is that I am tempted to use the logic routines of *A Logic of Exceptions* as well. However, this presumes that you have consumed that other book and this will not do. Even more awkward is that *A Logic of Exceptions* presumes “some decent highschool mathematics” while the very purpose of this present book is to provide that. Therefor I have decided to reduce formal deduction to the minimum and to explain logic when it arises. In a future there would have to be a better tuning. The book on logic already explains that the subject better be taught already at elementary school so there is still some way to go. (It is a bit curious that The Economics Pack while embedded within *Mathematica* has its own subdirectories `Logic`` and `Math`` again.)

Logical constants

<table>
<tr><th>form</th><th>full name</th><th>aliases</th><th>form</th><th>full name</th><th>alias</th></tr>
<tr><td>∧</td><td>\[And]</td><td>⋮&&⋮, ⋮and⋮</td><td>⇒</td><td>\[Implies]</td><td>⋮=>⋮</td></tr>
<tr><td>∨</td><td>\[Or]</td><td>⋮||⋮, ⋮or⋮</td><td>⥰</td><td>\[RoundImplies]</td><td></td></tr>
<tr><td>¬</td><td>\[Not]</td><td>⋮!⋮, ⋮not⋮</td><td>∴</td><td>\[Therefore]</td><td>⋮tf⋮</td></tr>
<tr><td>∈</td><td>\[Element]</td><td>⋮el⋮</td><td>∵</td><td>\[Because]</td><td></td></tr>
<tr><td>∀</td><td>\[ForAll]</td><td>⋮fa⋮</td><td>⊢</td><td>\[RightTee]</td><td></td></tr>
<tr><td>∃</td><td>\[Exists]</td><td>⋮ex⋮</td><td>⊣</td><td>\[LeftTee]</td><td></td></tr>
<tr><td>∄</td><td>\[NotExists]</td><td>⋮!ex⋮</td><td>⊨</td><td>\[DoubleRightTee]</td><td></td></tr>
<tr><td>⊻</td><td>\[Xor]</td><td>⋮xor⋮</td><td>⫤</td><td>\[DoubleLeftTee]</td><td></td></tr>
<tr><td>⊼</td><td>\[Nand]</td><td>⋮nand⋮</td><td>∍</td><td>\[SuchThat]</td><td>⋮st⋮</td></tr>
<tr><td>⊽</td><td>\[Nor]</td><td>⋮nor⋮</td><td>|</td><td>\[VerticalSeparator]</td><td>⋮|⋮</td></tr>
<tr><td></td><td></td><td></td><td>∶</td><td>\[Colon]</td><td>⋮:⋮</td></tr>
</table>

In *Mathematica*, ⋮ is the Esc symbol.

These symbols are hardly used in this book. But you will recognize terms that abound in logical argumentation.

Consider two trains running on a round track. Each train has two options: run clockwise or counterclockwise. There are four combinations and only two work. This is equivalence or “if and only if”. It is expressed as (P ⇒ Q) ∧ (Q ⇒ P) or, one train runs clockwise if and only if the other does too.

For a link to Logicomix: http://thomascool.eu/Papers/ALOE/2010-02-14-Russell-Logicomix.pdf

Greek alphabet

form	*full name*	*aliases*	*form*	*full name*	*alias*
α	`\[Alpha]`	⋮a⋮	Α	`\[CapitalAlpha]`	⋮A⋮
β	`\[Beta]`	⋮b⋮	Β	`\[CapitalBeta]`	⋮B⋮
γ	`\[Gamma]`	⋮g⋮	Γ	`\[CapitalGamma]`	⋮G⋮
δ	`\[Delta]`	⋮d⋮	Δ	`\[CapitalDelta]`	⋮D⋮
ϵ	`\[Epsilon]`	⋮e⋮	Ε	`\[CapitalEpsilon]`	⋮E⋮
ζ	`\[Zeta]`	⋮z⋮	Ζ	`\[CapitalZeta]`	⋮Z⋮
η	`\[Eta]`	⋮h⋮	Η	`\[CapitalEta]`	⋮H⋮
θ	`\[Theta]`	⋮th⋮	Θ	`\[CapitalTheta]`	⋮Th⋮
ι	`\[Iota]`	⋮i⋮	Ι	`\[CapitalIota]`	⋮I⋮
κ	`\[Kappa]`	⋮k⋮	Κ	`\[CapitalKappa]`	⋮K⋮
λ	`\[Lambda]`	⋮l⋮	Λ	`\[CapitalLambda]`	⋮L⋮
μ	`\[Mu]`	⋮m⋮	Μ	`\[CapitalMu]`	⋮M⋮
ν	`\[Nu]`	⋮n⋮	Ν	`\[CapitalNu]`	⋮N⋮
κ	`\[Xi]`	⋮x⋮	Ξ	`\[CapitalXi]`	⋮X⋮
o	`\[Omicron]`	⋮om⋮	Ο	`\[CapitalOmicron]`	⋮Om⋮
π	`\[Pi]`	⋮p⋮	Π	`\[CapitalPi]`	⋮P⋮
ρ	`\[Rho]`	⋮r⋮	Ρ	`\[CapitalRho]`	⋮R⋮
σ	`\[Sigma]`	⋮s⋮	Σ	`\[CapitalSigma]`	⋮S⋮
τ	`\[Tau]`	⋮t⋮	Τ	`\[CapitalTau]`	⋮T⋮
υ	`\[Upsilon]`	⋮u⋮	ϒ	`\[CapitalUpsilon]`	⋮U⋮
ϕ	`\[Phi]`	⋮ph⋮	Φ	`\[CapitalPhi]`	⋮Ph⋮
χ	`\[Chi]`	⋮c⋮	Χ	`\[CapitalChi]`	⋮C⋮
ψ	`\[Psi]`	⋮ps⋮	Ψ	`\[CapitalPsi]`	⋮Ps⋮
ω	`\[Omega]`	⋮o⋮	Ω	`\[CapitalOmega]`	⋮O⋮

In *Mathematica*, ⋮ is the Esc symbol. See for curly variants ∂, φ, ϑ: tutorial/EnteringGreekLetters

Brief contents

Table of contents

Part I. Introduction

For some curious reasons that apparently cannot be avoided, this part has a Chapter 0 with a long introductory discussion. When you are a student and want to begin you better jump to the Chapter 1 on geometry. If you get stuck then you could return to this introduction to see whether it helps out.

0.1 Conditions for using this book

The basic requirement for using this book is that you have at least a decent junior highschool level of understanding of mathematics or are willing to work up to that level along the way. We assume that this book could be used in advanced highschool or the first year of a college or university education.

You can read this book *as it is*, thus also if you do not have *Mathematica*. Even without ever running a program, you will still benefit from the discussion.

Readers new to the specific formats of *Mathematica* are advised to check the appropriate subsections on those, since those notations will be used.

Yet, if you have *Mathematica* and want to run the programs, then this book assumes that you have been introduced to *Mathematica*. You must be able to run *Mathematica*, understand its handling of input and output, and its other basic rules. Note that *Mathematica* closely follows standard mathematical notation. There are some differences with common notation though since the computer requires strict instructions. *Mathematica* comes with an excellent Help function that starts from the basics and works up to the most advanced levels. There are also many books that give an introduction.

When you want to run the `Math`AnalyticGeometry`` programs, you should also have a working copy of *The Economics Pack, Applications for Mathematica,* by the same author, as Cool (1999, 2001), with the software downloadable from the internet.

0.2 Structure of the book

This book is for both beginners and advanced readers. They are guided by the respective sections below.

The book basically has:

1. The basic ingredients: geometry, arithmetic and algebra, co-ordinates. Analytic geometry took off when Oresmus, Fermat and Descartes combined formulas with numbers and graphs on the plane.
2. The basic objects: line, circle and vector.
3. The consequences: trigonometry, the complex plane, linear algebra.
4. The news: "trig rerigged" and calculus developed fully from algebra.
5. Applications from physics, economics and statistics.
6. An extensive discussion of the didactics.

0.3 Using Mathematica

0.3.1 Axiomatics and other ways of proof

In their age-old civilisation, say 5000 years ago, the Egyptians developed a system for remarkably precise measurements. Complex constructions needed to be built, of which the pyramids were the largest ones. The stars needed to be traced. And when the Nile had flooded again and had destroyed some lands and created some new grounds, new lots had to be measured out for the displaced. When the Greeks came to visit, they noted this big body of geometric knowledge, and, perhaps not trusting all of it, they wondered: "Can you prove any of this ?" Eventually Euclid posed his axioms, and his book has been in use for a bit more than 2200 years now, see Struik (1977).

A long and wonderful story has been simplified here in perhaps too mundane terms. The discoveries of the notion of proof and of the axiomatic method are key events in human history. It is impossible to do them justice in just a few lines. Perhaps we should not look only to mathematics but look also for the source in codes of law, with that notion of proof. The subject of logic and proof has been developed in *A Logic of Exceptions* but will get attention below too, since Euclidean geometry has been a standard for strict reasoning and since this tradition is extended in analytic geometry. In logic, a main distinction is between (1) on one hand the axiomatic method that relies on substituting expressions into expressions, and (2) on the other hand the pure enumeration and investigation of all possible cases, which enumeration implies some notion of arithmetic and combinatorics.

In all cases a proof requires an understanding intellect that is willing to see, understand and accept "Yes, this convinces me". Perhaps harder is the "No, this does not convince me" in case that the proof fails. Often the voice of authority forces people to accept all kinds of statements even though the proof is weak or non-existent. See Aronson (1992ab) on how peer pressure can get a person to say that three lines are equal that aren't.

That being said, the second thing to say is that this book does not follow an axiomatic method, and we hardly prove anything. Our focus is on clarification and introduction in the subject, which at this level is complex enough. But there is a huge logical machinery behind it all.

0.3.2 *Mathematica* - also as a decision support environment

Mathematica is a language or system for doing mathematics on the computer. Note that mathematics itself is a language that generations of geniusses have been designing to state their theorems and proofs. This elegant and compact language is now being implemented on the computer, and this creates an incredible powerhouse that will likely grow into one of the revolutions of mankind - something that can be compared to the invention of the wheel or the alphabet; at least, it registers with me like that. Note that, actually, it is not the invention of precisely the wheel that mattered, since everybody can see roundness like in irisses, apples or in the Moon; it was the axle that was the real invention. In the same way next generations are likely to speak about the 'computer revolution', but the proper revolution would be this implementation of mathematics.

Mathematica already is a decision engine of a kind. If you run some algebraic solution routine then there is a lot of deduction before the answer pops up. However, that answer does not come as a neat English expression and does not read as a conclusion in the way that a good speaker would summarize his or her speech. The idea of this book is to learn how to interprete input and output in this language for analytic geometry and calculus.

Human mental processes are very sensitive to pictures. *Mathematica* is an enormous powerhouse for creating graphics. Thus, geometry could be developed using an abundance of pictures. Euclid's method partly relied on constructions that could be visualized, even though in proofs visualization was eliminated. The use of the power of *Mathematica* might cause a revival of *The Elements*. However, we arrive at a paradox now. While the present book uses pictures it also has a lot of text, formula's and numbers. The reason is that an abundance on pictures may also be an overabundance when the proper objective is exact determination and algebraic evaluation of solutions. Hence we use graphs but hopefully in a wise dose. We take geometry as a base area where proofs are not really required and then proceed to analytical geometry. We use a lot of concepts that you have actually some familiarity with. The objective is to create a deeper and more systematic understanding of those concepts.

The general format that people require for understanding contains four components and we try to provide these per topic - including active routines to move from the one to the other:

Text	Formula
Table	Picture

A good way of teaching is to present a paradox, have students think about it, present a punch line, and then all enjoy the moment when they get it in a flash. Alas, the presentation in this book is too new to follow that track. It is already quite a feat that we can tell the story as it develops below. (But it is an advice to you: to keep your mind in the active state of looking for questions that guide you through it all. And be aware of your learning style: active vs passive, abstract vs concrete.)

0.3.3 A guide

Since *Mathematica* is such an easy language to program in, it also represents something like a pitfall. It is rather easy to prototype the solution to a problem, or to write a notebook on a subject. But it still appears to be hard work to maintain conciseness, to enhance user friendliness and to document the whole.

Keep in mind the distinction between **(a)** issues in analytic geometry and calculus, **(b)** how a solution routine has been programmed, **(c)** the way how to use the routines.

This book focusses on (a). It however also provides a guide on (c) but neglects (b). Thus, the proper focus is on the *why*, i.e. the content of issues in analytic geometry and calculus, for which we want to apply these routines. But this also requires that we explain *how* to use them. If you want to know more about how the routines have been programmed, then you might use the routine `ShowPrivate[`*name*`]`.

0.4 Getting started

When you want to run the programs then you must do the following.

The Economics Pack becomes fully available by the single command `<<Economics`All``. It is good practice however to use a few separate command lines to better control the working environment. Three lines can be advised in particular.

0.4.1 The first line

You start by evaluating:

Needs["Economics`Pack`"]

This makes the `Economics[]` command available by which you can call specific packages and display their contents. Before you use this, read the following paragraphs first.

0.4.2 The second line

`CleanSlate`` is a package provided with *Mathematica* that allows you to reset the system. You thus can delete some or all of the packages that you have loaded and remove other declarations that you have made. The only condition is that `CleanSlate`` resets to the situation that it encounters when it is first loaded. You would normally load `CleanSlate`` after you have loaded some key packages that you would not want to delete. The ResetAll command is an easy way to call `CleanSlate``. Your advised second line is:

ResetAll

`ResetAll`	ResetAll calls CleanSlate, or if necessary loads it. This means that your notebook does not have to distinguish between calling CleanSlate` and evaluating CleanSlate[]

Note that if you first load CleanSlate` and then the Economics Pack, then the ResetAll will clear the Pack from your working environment, and thus also remove ResetAll. If you would happen to call ResetAll again after that, then the symbol will be regarded as a Global` symbol.

0.4.3 The third line

After the above, you could evaluate EconomicsPack to find the list of packages.

EconomicsPack

Select the package of your interest, load it, and investigate what it can do. For example:

Economics[Math`AnalyticGeometry, Math`Trigonometry, Math`Geometry, Math`Pythagoras, Calculus, Physics, Survival, Taxes]

You can suppress printing by an option `Print → False`. You can call more than one package in one call. If you want to work on another package and you want to clear the memory of earlier packages, simply call ResetAll first. This also resets the `In[]` and `Out[]` labels.

`Economics[xi, ...]`	shows the contents of xi` and if needed loads the package (s). Input xi can be Symbol or String with or without back–apostrophe. To prevent name conflicts, Symbols are first removed. Economics[ ] does not need the Cool`, Varianed` etc. prefixes
`Economics[All]`	assigns the Stub–attribute to all routines in the Pack (except some packages)
`EconomicsPack`	gives the list {directory → packages}

Note: Economics[x, Out → True] puts out the full name of the context loaded.

0.4.4 Using the palettes

The Pack comes with some palettes. These palettes have names and structures that correspond to the chapters in The Economics Pack itself.

- The master palette is "TheEconomicsPack.nb" and it provides the commands above and allows you to quickly call the other palettes or to go to the guide under the help function.
- The other palettes have "TEP_" as part of their name, so that they can easily be recognised as belonging to the Pack. These "TEP_" palettes contain coloured buttons for loading the relevant packages and text buttons for pasting commands.
- The exception here is "TEP_Arrowise.nb" that only deals with the package for making arrow diagrams.

An analytic geometry palette may at some time in the future be included.

0.4.5 All in one line

You can also load the Pack by the following single line.

```
<<Economics`All`
```

This evaluates Needs["Economics`Pack`"] and Economics[All], and opens the palettes. It does not call ResetAll, however.

0.5 Working environment

0.5.1 Language and evaluator

This book has been written in *Mathematica*, a system for doing mathematics. That program is both a text editor and an evaluator at the same time. Input in *Mathematica* are not just numbers but can be structured objects. That's why users of *Mathematica* rather don't speak about mere calculation but about *evaluation*. The text produced here is not only what the author has typed but also what the programs have generated. Those programs have been written to produce that output. You can change the input in *Mathematica* and generate different results.

0.5.2 Notation and help

One consequence of using *Mathematica* is that we will use its notation so that it can understand our formulas. The program has an extensive Help function. The full text of this book is available in *The Economics Pack*, and thus can be found as an Add-On Application in the Help function of *Mathematica.* See TheEconomicsPack palette for the Guide.

0.5.3 Input and evaluation

Next to plain text of the text editor there are also "input" and "output" cells for the evaluator. You enter commands in input cells (shown in bold type) and the result of the computer evaluation is printed below it (shown in "traditional form").

- This shows an input and output cell.

```
1 + 1
```

2

0.5.4 Full form and display

For pattern recognition, objects need to have a fixed format, called their FullForm. This form reads easier for computer programs but less easy for the human reader. Hence, the output cells can be displayed in a different form. When working with *Mathematica* you should always have the FullForm in mind.

- Operators can have various input formats. For example for division: (1) the FullForm, (2) the infix form, (3) the function call. When you type ESC div ESC then *Mathematica* creates a neat ÷ that also stands for division.

```
See[y / x,   Divide[y, x],   y ~Divide~ x,   y ÷ x,   FullForm[y/x]]
```

$\frac{y}{x}$ $\frac{y}{x}$ $\frac{y}{x}$ $\frac{y}{x}$ Times[Power[x, −1], y]

0.5.5 Getting used to *Mathematica*

Crucial notations to know are:

- $x == a$ means the logical statement that x and a are identical
- In a body of text $x = a$ will mean that $x == a$
- In a line of input $x = a$ means that variable x gets the value a
- Result or Out[ ] refers to the result of the former evaluation.
- A function call can be entered as $f[x]$ or as x // f
- x /. r means that substitution rules r are applied to x
- Lists are put in curly brackets like {*a*, *b*, *c*, ...}. The order within the list may or may not be relevant, depending upon context. The default situation is that the order does matter. Then {*a*, *b*, *c*, ...} can be a program with executable statements.

The above only gives the basic necessities that you require for understanding the notation, texts and programs below. If you encounter problems below on how issues are implemented in *Mathematica* then it is advisable to dwell a bit longer on them, since discovering more about *Mathematica* is an investment that can pay off in various subjects. One key advantage is that you can write your own programs once you become comfortable with the language. A good way to look at *Mathematica* is to regard it as a language indeed (and not just a computer program).

0.6 For teachers

When you are a teacher (or a student of the didactics of mathematics) then this section is merely to draw your attention to Part V in the back of the book with the didactic deliberations. As a teacher you have these options: (1) first read the book as a student, to see how it would come to them, without the deliberations in Part V, or (2) first read these didactics so that you are better informed but less equipped to trace the fresh student, or (3) jump to places as you are guided by curiosity. The advice is to do (1), make notes, and later compare your notes with Part V. Before you decide be sure to read the Preface first.

Two small examples of the approach in this book may be given so that you can make a better informed choice. (1) Traditionally we have 2½ for two-and-a-half and 2 a for two-times-a, such as $2\sqrt{2}$ for twice the square root of two. A student may write 2½ as 2 ½ because fixed positions are hard in handwriting; and later conclude 2 ½ = 1. Where do we go wrong and what is the solution ? We should write 2 + ½ and accept that addition does not always simplify. In the same way 2 × 2 = 4 simplifies but $2\sqrt{2}$ does not. The notation 2½ may not be a problem for textbooks with fixed positions but for the handwriting by students it is a minefield, and the whole problem is essentially caused by a mathematical unawareness of the active or passive meaning of the plus-sign. (2) The traditional definition of the cosine is Cos = a / h as the ratio of the adjacent to the hypotenuse. Subsequently Cos is treated as a function of the angle and transformed into a formal definition Cos[φ] = a / h. The standard format for defining a function like x-squared however is $f[x] = x^2$. In the traditional definition of Cos[φ] there is no φ on the right hand side. It is an inverse definition without the explanation that things are taken inversely. It actually is an equation $a = h$ Cos[φ] and the issue is equation solving and not defining a function. If you have an angle then you can calculate the a / h ratio from the cosine. Also, textbooks fear Greek letters and write Cos[x] = a / h with x for the angle or arc but in the standard setup on the unit circle a is on the horizontal axis and is a value of x too, so that Cos[x] = x ? The better way is to present φ = ArcX[x / h] so that when you have the ratio then you use that ArcX function to find the arc - and the X in the name helps the identification. These are just two examples of a longer list and some of the more accessible ones.

1. Geometry

1.1 Learning by doing

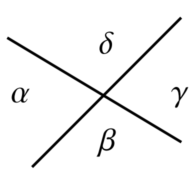
You are now out on a conquest of the plane. You have been familiar with the plane ever since someone gave you a pencil and a piece of paper to draw on. It can be exciting to look at it in a new way. The systematic way. This will give you a jolt that can startle you. The idea of this part is that you first grasp the basics. The best introduction to geometry is to do it.

The name geometry derives from Greek *gaia* or *ge* for *land* and *metrein* for *measurement.* With *orgein* for *work* we see that George is a farmer and Georgia a region with farmers.

1.2 Lines

1.2.1 Intersecting lines and their angles

Regard the following two intersecting lines. They cut out four angles that we have labeled with the Greek lowercase letters alpha, beta, gamma and delta (see the beginning of the book). An angle is the section of the plane between the two half lines from the point of intersection. The angles add up to 1, standing for the full plane as our unit of measurement. Thus we have the equation $\alpha + \beta + \gamma + \delta = 1$.



When we fold the paper through the point of intersection we can let the lines overlap, so that we conclude that $\alpha = \gamma$ and $\beta = \delta$. A bit more involved is this calculation that also generates the insight of the value 1/2. Since the lines are straight and since straightness is not affected by the intersection, we have:

1. $\alpha + \beta = \gamma + \delta = 1/2$ since these are halves on the side of one line.
2. $\beta + \gamma = \alpha + \delta = 1/2$ since these are halves on the side of the other line.

3. Subtraction gives $\alpha - \gamma = \gamma - \alpha = 0$, or $\alpha = \gamma$.
4. Thus also $\beta = \delta$.

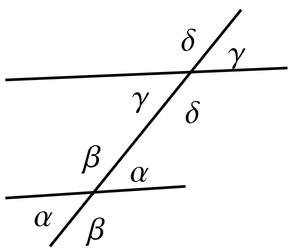
Conclusion: when two lines intersect then the opposing angles are equal.

NB. Where lines intersect is a *point*. Overlapping lines are the same one. A point is location without size. A line is length without width. Two points define a line.

NB. When all angles are equal then $\alpha = \beta = 1/4$ and these we call right angles. The lines then are said to be perpendicular. A fun way to create perpendicular lines is this: Put a piece of paper on a flat surface. Place two dots on it. Fold the paper such that the fold is exactly over the two dots. Then open the paper again and fold the paper such that the lines of the first fold are neatly overlapping. There is your perpendicular cross. (This uses 3D.)

1.2.2 Parallel lines and their angles

If two lines do not intersect then they are parallel. Take two parallel lines intersected by a third line. We label the angles, using the knowledge of opposing angles.



Since the lines are straight we have:

1. $\alpha + \beta = \gamma + \delta = 1/2$ since these are halves on the side of one line.
2. We see that $\alpha = \gamma$ and $\beta = \delta$. For example if we cut the paper at halfway the intersecting line-section and move the two partial sheets over each other, then the lines overlap. (Seeing subsitutes for axioms or complex reasoning.)
3. Since $\alpha + \beta = 1/2$ we also get $\alpha + \delta = \beta + \gamma = 1/2$. If this were not the case then the lines would eventually cross, and we have assumed that they are parallel. (Euclids fifth postulate.)

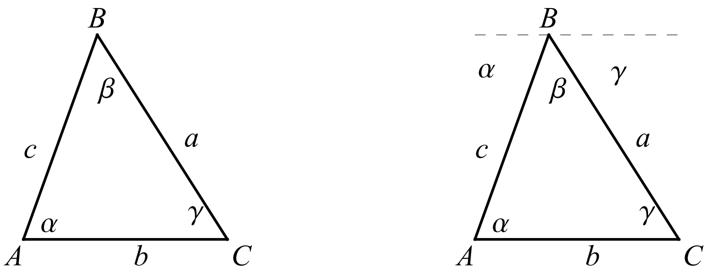
Conclusion: when parallel lines intersect with a third one then there are only two angles, that add up to 1/2.

NB. In the intersection we see Z and F shapes that help identify equal angles. Parallel lines can be denoted as *k* // *m*.

1.3 Angular shapes

1.3.1 The sum of angles in a triangle

A triangle arises when there are three points not on a single line. The connecting line sections are called sides. The area of the triangle is enclosed by those sides. We label the corners by upper case *A, B* and *C,* the opposing sides by lower case *a, b* and *c,* and the angles with α, β and γ. The triangle itself is identified by ABC. The layout is shown on the left hand side.



When we draw a line through *B* that is parallel to *b* then we see (inverted) *Z* shapes that allow us to identify angles of sizes α and γ around β. We conclude that $\alpha + \beta + \gamma = 1/2$. The angles within a triangle add up to half a plane.

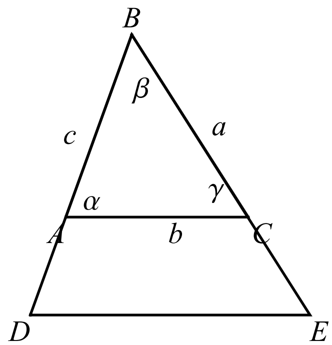
A corollary is that when all angles are equal then each is 1/2 * 1/3 = 1/6. The sides then must be the same since when you flip it around the results neatly overlap. This is an equilateral triangle.

An isosceles triangle has at least two equal sides, and then has at least two equal angles and conversely. For example, $a = c$ iff $\alpha = \gamma$. *Iff* means *if and only if* and expresses an equivalence. Both necessary and sufficient conditions are satisfied. A proof is to flip an isosceles triangle around and fit it onto itself: if the sides are the same then the angles are the same, and if the angles are the same then the sides are the same.

NB. An angle can also be written ∠A. Or, when A is in the middle of more intersecting lines as ∠BAC where the middle letter identifies the corner point. A side can also be named by its corner points, for example $AB = c$.

1.3.2 Proportionality

If line sections *AB* and *BC* are extended to *BD* and *BE,* and *DE // AC* then the following sections are in proportion: *BA : BD = BC : BE = AC : DE.*

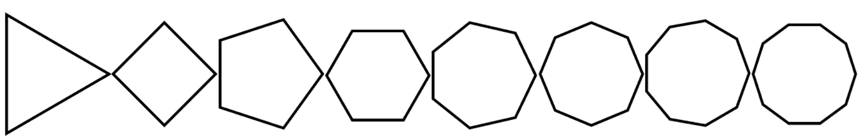


This is rather difficult to prove. A shortcut is to simply define proportional sides as such that all sides are multiplied with the same factor. But then we have to prove that those triangles indeed exist (for the definition might be nonsense). For now we rely on our intuition that if the proportionality would not hold then the lines would cross somewhere. (If you think that is sloppy then you are right.)

`TrianglePlot[α:1.2]`	gives the basic layout of the triangle, with names for corners, angles (α) and sides. Option Projection → 1 shows 1 projection, All shows all. Option Add → True shows the proof that the angles add up to 1/2
`ProportionalTriangle-Plot[]`	shows proportionality

1.3.3 Polygons

Angular shapes are called polygons - from the Greek *poly* for *many* and *gonon* for *angle, corner, vertex*. Regular polygons are such that the vertices lie on a circle and that the angles are equal. The triangle is actually a trigonon.



Taken from *Mathematica*'s documentation.

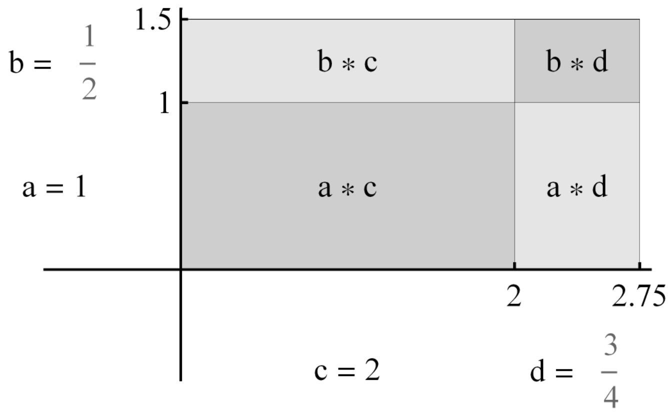
1.3.4 Calculation of circumference and area

Geometric questions can be about circumference *Cir* and area or surface *Sur*. Circumference has to do with addition. Area has to do with multiplication.

- The surfaces of the separate rectangles add up to the grand total.

(a + b) (c + d) // Expand

$a c + a d + b c + b d$



For a rectangle with sides h and w the circumference is $Cir = h + w + h + w = 2\ (h + w)$ and the area is $Sur = h\ w$. When we multiply height and width by a proportionality factor p then the circumference rises with the same factor to $2\ (h + w)\ p$, but the area rises by its square to $h\ w\ p^2$.

An area is generally measured by rectangular shapes, unless we find other tricks.

For example, to find the area of a triangle we try to fit it into a rectangle. You may try to do so yourself. As a start, take the diagonal in the rectangle above and see the two triangles. Then generalize. OK, the area of a triangle is $Sur = h\ b\ /\ 2$, where h is the height of the triangle and b its base.

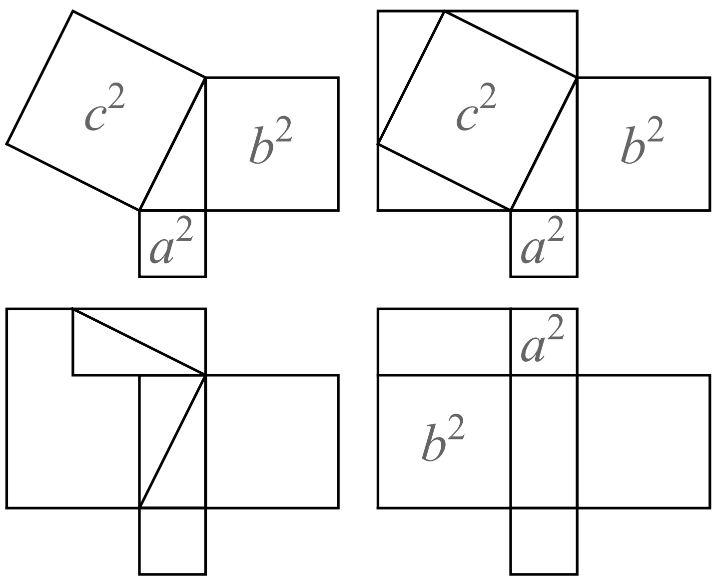
`APlusBTimesCPlusDGO[`*a*`,` *b*`,` *c*, *d*, *opts*`]`
is a graphics object that displays the multiplication of (a + b) (c + d) as the surface of a rectangle with those sides

1.4 The Pythagorean Theorem

1.4.1 The theorem

For a triangle with a right angle at C (a right triangle) we find $c^2 = a^2 + b^2$. This theorem is named after Pythagoras.

- Prove the theorem. Use the triangle to complement c^2 to a larger square. Shift the complementing triangles till the equality with the sum of squares is clear.



`Pythagoras[`*a*`,` *b*`,` *opts*`]`

gives a graphical proof of the Pythagorean Theorem for a right angled triangle with short sides a and b, i.e. the theorem that the square of the hypotenusa equals the sum of squares of the sides $\left(c^2 = a^2 + b^2\right)$.

1.4.2 Pythagoras and ratios

The historical link between Pythagoras and the theorem is a bit vague. But there is a good link to Pythagoras and his harmonies and theory of numbers. This concerns the idea that numbers could always be expressed as a ratio of integers, or that they thus are “rational”. With the theorem we can prove that this is not so. Hence next to the rational numbers there also are ir-rational numbers.

Note what is involved. We take a particular length or line section as the unit of measurement. For example a meter. We presume that we can measure all other lengths as multiples or ratios of this unit of measurement. A ratio is $n : d$, with the numerator and denominator taken as integer numbers. However, such ratios appear to fail. Lengths can be expressed in lengths, but not in rational numbers.

Consider a square with sides 1. With the theorem we see that the diagonal is $\sqrt{2}$. The diagonal exists so this length must also exist. Suppose that it can be expressed as a ratio of integers in the numerator and denominator, say $\sqrt{2} = n / d$. When both n and d would be even then we divide both by 2 to simplify and we work with the result. Hence we can assume that n and d cannot both be even. Squaring gives $2 = n^2/ d^2$ or $n^2 = 2\, d^2$. Note that if p is an odd number then p^2 is odd too. We see that n^2 is even (it is 2 times something) so that n cannot be odd. Thus it is even. If n is even then there is some number z so that $n = 2\, z$. Then $n^2 = 4\, z^2$. We already had $n^2 = 2\, d^2$. From this it follows that $d^2 = 2\, z^2$. We conclude that d is even. We started out saying that n and d cannot both be even but now we have deduced that they are. Contradiction. The only exit route is that $\sqrt{2}$ cannot be expressed as a ratio of two integer numbers.

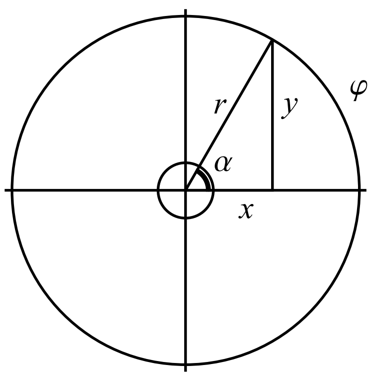
The ancient Greeks did not develop a theory of arithmetic that allowed them to deal with this. They used the letters of their alphabet to denote their numbers and this is rather a drawback. Instead, they focussed on geometry where it is no problem to work with lengths like $\sqrt{2}$. The core of Euclid's *Elements* consists of the theory of proportions that allows him to measure what is needed. Euclids "number" then is a ratio of line sections in geometry. For us, $\sqrt{2}$ still presents a challenge when you consider the decimal expansion. The mathematical description of the continuum still assumes a basic notion of what a continuum is, as distinct from what we call "number sense".

1.5 The circle

1.5.1 The circle

A circle is a curved line such that all points are at the same distance to a common center. The common distance is called the radius, here r. Practical examples come from a compass or from swinging something around an axle.

A diameter is a line section from one point on the circle passing through the center onwards to the other side of the circle. This graph has two diameters that are perpendicular, i.e. they create four equal angles each of 1/4. We know that those exist, creating them is another issue. The surface of the circle is contained in the encompassing square with sides $2r$, thus 4 times the r^2 squares. A lower boundary, by eye, is 3 of those. Thus $3\ r^2 \leq$ surface $\leq 4\ r^2$.



This book is essentially a rewriting of this short paragraph into a better understanding of what we say here.

1.5.2 Two independent factors in a two dimensional plane

A first step is to put symbols at the elements in the graph. Consider an angle α in the circle between the horizontal diameter and the drawn radius. Associated with that angle is an arc φ on the circle. Where the drawn radius meets the arc we take a line parallel to the vertical diameter, and then see a triangle with sides r, y and x. Properties are:

1. Angle α and arc φ are directly dependent and related proportionally.
2. Given the right angle we have $x^2 + y^2 = r^2$. PM. We use the word "projection" for the procedure of mapping some point perpendicular onto a line.
3. We call $s = y / x = s[\alpha]$ the *slope* caused by α. The slope is the *rise* divided by the *run*. For example a 50% slope means that you have to climb 50 centimeter for every meter that you progress sideways. A mountain can have the same slope

for a long while. If we divide x and y by r then the ratio $(y / r) / (x / r) = y / x$ again. Due to proportionality r drops out.

4. When α and r change then φ, x and y change.

5. When x and y change then φ, α and r change.

It appears that the circle has only two independent factors. Since there are two independent factors we say that the plane is two-dimensional. We will not do much with this observation until a later chapter. This is however the place to become aware of it and to record it for later reference.

We can express the dependencies in functions and their inverses. These are called the trigonometric functions (simply because there is a triangle involved: tri, gonon, measurement). We will discuss these functions later. We now proceed with some geometric insights that use the circle.

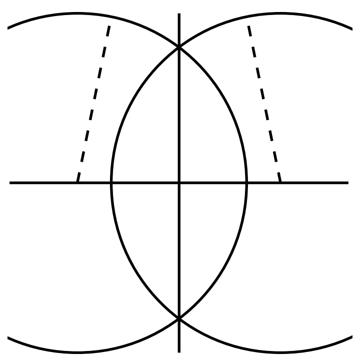
PM. Elements of a circle are at http://demonstrations.wolfram.com/GeometricElementsOfACircle/

```
UnitCirclePlot[]
```

displays a unit circle, with radius r = 1, co−ordinates {x, y},
angle φ in radian as the arc from {1, 0} to {x, y},
an inner circle with circumference 1,
and an angle α as the inner arc measured in Unit Turn or Unit Measure Around (UMA). Here x = Xur[α] = Cos[φ] and y = Yur[α] = Sin[φ]. Opts affect display

1.5.3 Using ruler and compass

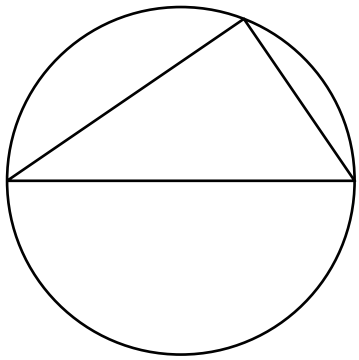
Some handicraft is to use a compass to draw a circle. You can find a perpendicular line by drawing two circles with centers on a base line. The line through the points of intersection of the circles is perpendicular to that base line. Normally you would not draw full circles but just some small arcs around the points of intersection. (Using a compass is 3D too but more in line with geometric tradition.)

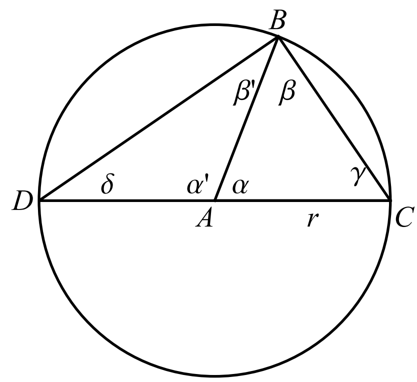


Check http://demonstrations.wolfram.com/IndestructiblePerpendicularLines/

1.5.4 The theorem by Thales

The theorem by Thales is on record as officially the first theorem in geometry. On a circle we take two points on a diameter and one not so. The latter has a right angle.





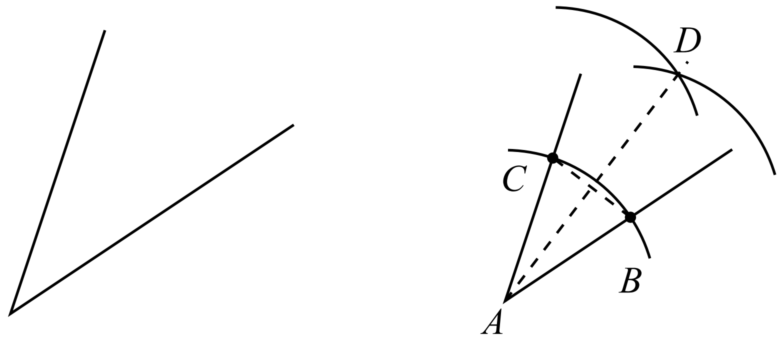
The recipe for finding a proof is: write down all that you know and with some thought the answer pops up. The proposition is depicted in the left graph. The right graph puts in all what we know, including the crucial radius. Then we find:

1. $\alpha + \beta + \gamma = 1/2$, $\delta + \alpha' + \beta' = 1/2$, $\delta + \gamma + \beta + \beta' = 1/2$ for all triangles
2. $\alpha + \alpha' = 1/2$
3. ABC is isosceles because of the radius and thus $\beta = \gamma$
4. ABD similary and thus $\beta' = \delta$
5. $\delta + \gamma + \beta + \beta' = \beta + \beta' + \beta + \beta' = 1/2$ gives $\beta + \beta' = 1/4$

`ThalesTheoremPlot[(None, All) φ:1.2]`	with or without labels for angle φ (default 1.2)

1.5.5 Halving an angle

When two lines are given, how would you draw the line that bisects their angle ? The graph to the left is pretty empty and daunting while the proof on the right may require some thought.



Take the intersection of the lines as the center *A* of a circle with arbitrary radius. We find points *B* and *C*. Thus *ABC* is an isosceles triangle. Take *B* as the center of a circle with arbitrary radius (say the same). Take *C* as the center of a circle with the same radius. Their intersection is at *D*. The line *AD* bisects the angle at *A*.

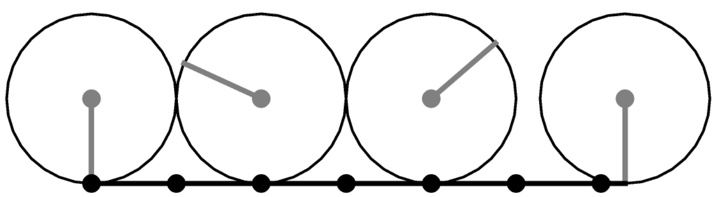
`BisectionPlot[None, a, b]`	shows just two lines from the origin to a and b, that form an angle that needs to be bisected
`BisectionPlot[Label, a, b, r]`	shows the labelled solution by drawing arcs with radius
`BisectionPlot[All, a, b, r]`	shows both in a GraphicsRow

1.5.6 Circumference and area of the circle

Calculation of circumference and area of the circle is a bit of a conundrum since how can we find a rectangle to multiply the sides ?

Regarding circles with the same center but different radii we see that they are proportional. If you do not believe that then take the regular polygons that fit within, and use the proportions on the triangles within those (§1.3.2). Thus we can use a proportionality constant Θ such that the *Cir* = *r* Θ. We use the Greek capital theta to reflect the shape of a circle. A circle with radius $r = 1$ thus has a circumference that is precisely that proportionality constant Θ.

If we want to measure Θ empirically then we take a circle with radius 1 and cycle it along a ruler. We find a value of approximately 6.28. Thus $\Theta \approx 6.28$.



Adapted from Stephen Wolfram http://demonstrations.wolfram.com/CircumferenceOfACircle/

For the area we find $Sur = ½\, r^2\, \Theta$. See Chapter 9 for the proof. Again: circumference is proportional, and area depends upon the square of the proportionality factor, in this case the radius.

In geometry we frequently do not specify a unit of measurement. This only becomes relevant for practical application when the engineers take over from the mathematicians. Circumference and radius can then be measured in any unit (meters, feet). When we take the ratio then that unit or measurement drops out again (Meter / Meter = 1). Thus Θ can be taken as a dimensionless number. Just like everything in geometry tends to be taken without a unit of measurement. In another respect it is not dimensionless since 'going around' is a phenomenon of its own. The issue is similar to the rate of interest. You put $100 in a bank and a year later you get $105. The dimension seems Dollar / Dollar = 1 but actually there is a time difference so that the unit of measurement of the 5% rate of interest is Per Year = (1 / Year). The unit of Θ can be Per Turn.

1.6 Measurement of angles and arcs

1.6.1 Angle and arc

Geometry is not just about lines but a lot of analysis concerns angles and arcs. A key contribution of analytic geometry is that it has found a precise measurement of the latter.

In above diagram the circle turned only once. The depicted radius cycled over the whole plane. We have been measuring angles as ray sections of the plane, taking the plane itself as the unit of measurement 1. This is equivalent to taking a turn of the circle. It is conceptually more agreeable to do the latter.

Thus our angles α, β, ... can be expressed in the number of turns of a circle. For example a half turn, a quarter turn, etcetera.

Measuring turns is not so simple. Halving angles and halving again and again like we have shown to be possible will get us far but we should be able to do better. A neat approach is to regard a circle with radius $\varrho = 1 / \Theta$. Then its circumference is $Cir = \varrho\, \Theta = (1 / \Theta)\, \Theta = 1$. That means that the fractions that we have been using for the angles can also be located on this special circle. We now redefine *angle* as the arc on this inner circle. Our newly defined angle is an arc too, but at a special location and with the neat property that a full turn has length 1.

Regard the circle diagram in subsection 1.5.1. For the outer circle we take $r = 1$. We call this the Unit Radius Circle, or Unit Circle for short. Within this circle there is drawn another much smaller circle, with a radius of 1 / Θ. This is the Unit Circumference Circle, or the Angular Circle for short. Note why this circle seems so small: a circle with a circumference of about 1 meter must have a radius of about 16 cm.

Thus these definitions apply:

Definitions	*Angular Circle*	*Unit Circle*
Radius	$1 / \Theta$	1
Circumference	1	Θ
Arc	α	$\varphi = \alpha \Theta$
Angle	α	α

$\Theta \approx 6.2831853071795 \approx 6.28$ and $\rho = 1 / \Theta \approx 0.1591549430918953 \approx 0.16 \approx 1/6$.

Unit Circle = Unit Radius Circle. Angular Circle = Unit Circumference Circle.

1.6.2 Numerical values

Let us consider the numerical values of the angle α on the angular circle and the arc φ on the unit circle a bit closer.

- Take the decimal fractions of the angle and compute the distance rolled.

$$\begin{pmatrix} \alpha & 0 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1. \\ \text{Distance} & 0 & 0.63 & 1.26 & 1.88 & 2.51 & 3.14 & 3.77 & 4.4 & 5.03 & 5.65 & 6.28 \end{pmatrix}$$

- Take the distance rolled and compute the α. The distance steps are multiples of $r = 1$. Included are the quarters of Θ too.

$$\begin{pmatrix} \text{Distance} & 0 & 1 & 1.57 & 2 & 3 & 3.14 & 4 & 4.71 & 5 & 6 & 6.28 \\ \alpha & 0 & 0.16 & 0.25 & 0.32 & 0.48 & 0.5 & 0.64 & 0.75 & 0.8 & 0.95 & 1. \end{pmatrix}$$

When we are interested in the distance rolled when $r = 1$ then the data presented in the second table are relevant and then the distance φ gives the relevant information. But if we are interested in distance cycled and there is another radius, say $r = 25$, then α and φ remain the same but the distance now is $25\ \varphi$. These tables thus are only relevant for the angular and unit circles.

`AngleDistanceTable[ b:0, f:1, n:2]`	uses the angles ranging from b to factor f (e.g. 1/2) times 10 and rounds to n
`DistanceAngleTable[ b:0, f:1, n:2]`	uses the distances ranging from b to factor f (e.g. 1/2) times 10 and quarters of $\Theta = 2\pi$, and rounds to n

1.6.3 Traditional transforms of Θ: degrees and radians

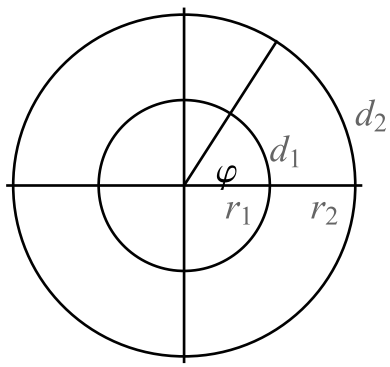
Relevant are the following traditional transforms of Θ.

1. Sumerian astronomers divided the sky, a year and the circle in 360 parts, in days or degrees denoted as 360°. Turning around is 180° and a right turn is

90°. The advantage is that common angles like the 24 hours of a day or the 12 months per year or the 8 corners of the wind have integer values. Many typical values come from regular polygons, with regular steps in the cycle. A historical estimate of $\pi = \Theta / 2 \approx 3 \approx 22/7$ is used.

2. Modernity adopted the definition of radians or Rad, and calculates π to ever greater precision (it is not entirely clear when they have the courage to stop).

The following plot shows how radians are defined. One radian is the arc of length 1 on the unit circle. Just length but on an arc, and specifically on the unit circle so that there are $\Theta = 2\pi$ radians to go around. Since circles are proportional, the number of radians is given by the ratio of arc to radius. Regard a circle with r_1 and arc d_1 and another circle with r_2 and d_2. When the angles are the same then the radian measure is $\varphi = d_1 / r_1 = d_2 / r_2$. When $r_1 = 1$ then $\varphi = d_1$ is the arc on the unit circle. (Special is $\varphi = d_1 = r_1 = 1$. The graph uses $d_1 = r_1$ so that $d_2 = r_2$, and we can choose either as our unit of measurement.)



Below we will give a deeper comparison between the various standards of measurement. Before we draw a conclusion it is useful to better know what is involved.

A quick comment now is that φ in the circle above should rather be drawn on the arc at $r = 1$ where it really is. For $\alpha = \varphi / \Theta$ the ratio $d[r] / r$ is also constant but the angle α is then defined as a standard fraction $1 / \Theta$ of that ratio. Users of radians and π commonly express results not only in radians but also as fractions of a full turn around the circle and then divide by 2π too; but curiously this is not done for the unit of measurement. Instead of dividing each time it seems better to do it only once at the definition of the unit of measurement.

Let us compare the measures numerically. Above tables were for angle α on the angular circle and arc φ on the unit circle, i.e. the distance around or rolling when $r = 1$. Now we can say that those distance values are also radians. Commonly those radian values are expressed as multiples or fractions of π rad. Those coefficients still have only meaning in terms of how many turns are made. The Sumerian degrees are handy since they use integer values, where we would use percentages.

StandardAnglesTable[True]

$$\begin{pmatrix} \alpha & 0 & \frac{1}{10} & \frac{1}{5} & \frac{3}{10} & \frac{2}{5} & \frac{1}{2} & \frac{3}{5} & \frac{7}{10} & \frac{4}{5} & \frac{9}{10} & 1 \\ \mathrm{N}[\alpha] & 0 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 & 1. \\ \mathrm{N}[\alpha]\ \% & 0 & 10. & 20. & 30. & 40. & 50. & 60. & 70. & 80. & 90. & 100. \\ \mathrm{Deg} & 0 & 36 & 72 & 108 & 144 & 180 & 216 & 252 & 288 & 324 & 360 \\ \mathrm{Rad}\ \varphi & 0 & \frac{\pi}{5} & \frac{2\pi}{5} & \frac{3\pi}{5} & \frac{4\pi}{5} & \pi & \frac{6\pi}{5} & \frac{7\pi}{5} & \frac{8\pi}{5} & \frac{9\pi}{5} & 2\pi \\ \mathrm{N}[\mathrm{Rad}] & 0 & 0.63 & 1.26 & 1.88 & 2.51 & 3.14 & 3.77 & 4.4 & 5.03 & 5.65 & 6.28 \end{pmatrix}$$

Let us use the angles for months and corners of the wind as those are commonly used in textbooks. Not all fit in this table so we only consider a half plane. The choice for $\pi = \Theta / 2$ relates to the property that we may use half of the values: an angle above the horizontal is positive while an angle below the horizontal is negative.

$$\begin{pmatrix} \alpha & 0 & \frac{1}{12} & \frac{1}{8} & \frac{1}{6} & \frac{1}{4} & \frac{1}{3} & \frac{3}{8} & \frac{5}{12} & \frac{1}{2} \\ \mathrm{N}[\alpha] & 0 & 0.08 & 0.12 & 0.17 & 0.25 & 0.33 & 0.38 & 0.42 & 0.5 \\ \mathrm{N}[\alpha]\ \% & 0 & 8.33 & 12.5 & 16.67 & 25. & 33.33 & 37.5 & 41.67 & 50. \\ \mathrm{Deg} & 0 & 30 & 45 & 60 & 90 & 120 & 135 & 150 & 180 \\ \mathrm{Rad}\ \varphi & 0 & \frac{\pi}{6} & \frac{\pi}{4} & \frac{\pi}{3} & \frac{\pi}{2} & \frac{2\pi}{3} & \frac{3\pi}{4} & \frac{5\pi}{6} & \pi \\ \mathrm{N}[\mathrm{Rad}] & 0 & 0.52 & 0.79 & 1.05 & 1.57 & 2.09 & 2.36 & 2.62 & 3.14 \end{pmatrix}$$

The different measures can be fully translated into each other. There is no issue about what we know about measuring angles (seen as sections of the plane between intersecting lines). There is an issue though on what is agreeable to the human mind and effective in education. The approach adopted here is to use α on the angular circle as *the* angle.

```
StandardAnglesTable[decimalq:False, b:0, f:1, n:2]
```

gives the angles ranging from b to factor f (e.g. 1/2)
and rounds to n. If decimalq is True then decimal steps

```
DefinitionRadians[phi, r]
```

shows how radians are defined: phi is the arc, the radius of the first circle is 1,
and r > 1 is the radius of the second circle. phi =
d1/r1 = d2/r2 (in this case r1 = 1 and r2 = r), or the ratio of the section
of the circle with the radius. Since this ratio is constant for phi,
for whatever radius, we define phi as that ratio. With r1 = 1,
d1 are the radians. The whole circle has Θ = 2 Pi radians

1.7 Review and looking ahead

Above gives only a tiny fraction of geometry in standard course books. But presently we are concerned with just the basics in order to continue with analytic geometry. The key aspects of an introduction to geometry have been covered:

1. There are point and line but also length (line section) and area.
2. Proportionality.
3. Angles arise in angular shapes but there are also curved shapes.
4. The circle combines notions of both distance (points at equal distance) and angle and arc.
5. There is the notion of proof: definitions, theorems and proofs. Understanding the reason why something is as it is removes the tedium of mere memorizing.
6. There is arithmetic with calculation. We may prove that something like an intersection exists but calculating where it actually is is a different issue. It is unclear how we got the value of $\Theta \approx 6.28$.

The next step is to introduce more involved calculation and algebra.

2. Arithmetic and algebra

2.1 Numbers and symbols

Arithmetic and algebra are linked since (1) algebra is an abstraction of arithmetic calculation, and (2) calculation can be seen as elementary algebra. The basic steps are all the same, with numbers or with unknowns or variables. A good educational foundation in arithmetic helps doing algebra later.

- Examples of addition and multiplication.

(a + b) (c + d) == e // Expand

$$a c + a d + b c + b d = e$$

(2 x + 1) (x + 4) == 25 // Expand

$$2 x^2 + 9 x + 4 = 25$$

2.2 Approximation and rounding off

2.2.1 Different formats cause choice

The number 1/3 is a pure number that can be approximated in a decimal expansion as 0.33333... Putting the dots there indicates that the 3's continue forever. Since we cannot easily work with those dots we round off to some accuracy. The accuracy is the number of digits behind the decimal dot. For example 0.3333 is an approximation of 1/3 with an accuracy of 4. A number is called pure when it has infinite accuracy.

When the true number is 0.3333 then we can also say that 1/3 is an approximation of the latter. What is an approximation depends upon the target, and it is not correct to identify decimal numbers with approximation itself. It is immaterial whether we write 1/4 = 25/100 = 0.25 since these are all equivalent expressions of a pure number (a quarter). For an approximation we use other indicators than the decimal dot.

This differs from conventions in computer science where decimal numbers are considered to be approximate by definition.

- *Mathematica* distinguishes between equality in approximate numerical form and identity in terms of form.

```
See[ 0.4 == 2/5,   0.4 === 2/5]
```

True   False

- *Mathematica* assumes infinite accuracy for quotients and state dependent accuracy for decimals (e.g. computer precision).

```
See[ Accuracy[2/5],   Accuracy[0.4] ]
```

∞   16.3525

In education it is better however to accept 0.4 as an expression of infinite accuracy too and just 4/10 written in decimal form. We need a separate notation for an approximate number. To understand the issue let us first look at computers. There is a whole science on how to implement numbers on a computer. Students use calculators and must understand essentials, and our overal use of numbers must be consistent.

2.2.2 On the computer

In computer calculations errors in accuracy can multiply faster than the numbers themselves. For example, 0.2 * 11 - 2 clearly must be 0.2 again. When we write this as a function and use the result upon itself again in another step on the computer, then this may seem to go well for a decent number of steps but at some point the numerical inaccuracy takes over.

- Nestlist uses a startvalue and then applies the function to its former results.

```
f[x_] := x * 11 - 2; NestList[f, 0.2, 20]
```

{0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.200005, 0.200051, 0.200557, 0.206132, 0.267457, 0.942028, 8.36231, 89.9854, 987.84, 10864.2}

Mathematica allows to include the accuracy in a number, for example 0.2`20 means that there are 19 zeros included. This works OK for a while but eventually we still land at 0. Working with pure number 1/5 remains best.

```
f[x_] := x * 11 - 2; NestList[f, 0.2`20, 20]
```

{0.20000000000000000000, 0.2000000000000000000, 0.200000000000000000, 0.20000000000000000, 0.2000000000000000, 0.200000000000000, 0.20000000000000, 0.2000000000000, 0.200000000000, 0.20000000000, 0.2000000000, 0.200000000, 0.20000000, 0.2000000, 0.200000, 0.20000, 0.2000, 0.20, 0.2, 0×10^{-1}, 0}

```
f[x_] := x * 11 - 2; NestList[f, 1/5, 20]
```

$$\left\{\frac{1}{5}, \frac{1}{5}\right\}$$

Thus importantly: computer languages use decimal numbers as indicators of approximate numbers, and even *Mathematica* follows that convention. When for us 0.2 is a pure number and a convenient manner of writing 2/10 then for accurate computing we must make sure that the computer input is not 0.2 but 2/10. In *Mathematica* we can use Rationalize, in Excel we better use 2 * 11 - 20 and divide those results by 10 to 0.2 if needed.

The example about 0.2 * 11 - 2 was presented by Jon McLoone of WRI.
See the computer arithmetic package tutorial in *Mathematica*'s help function.

2.2.3 Denoting an approximation

What is generally missing from mathematical textbooks and computers is a standard way to express how a pure number like 0.25 differs from a decimal approximate number 0.25... where we have rounded off to 2 digits. The lingering dots are not deemed acceptable.

A solution suggestion here is to use a tilde (~) on the last digit to indicate that the number concerns an approximation. Thus we write $x = 0.2\underset{\sim}{5}$ when 0.25 is the pure number and x only an approximation. We can also use the position under or above to indicate the kind of approximation.

- Put a tilde (~) over (under) the last digit if the true value lies above (below) the approximation. NB. These are Strings and not Numeric.

```
{NTilde[0.254591, 2], NTilde[0.2494591, 2]}
```

$\{0.2\tilde{5}, 0.2\underset{\sim}{5}\}$

In some European nations, exams are graded on a scale from 0 to 10. You pass when the grade is at least 5.5 and this is rounded to 6. Some students like to round 5.48 to 5.5 and then to 6. The tilde helps to remind us that one 5.5 is not quite another 5.5.

$$\begin{pmatrix} 5.48 & 5.\underset{\sim}{5} & 5 \\ 5.49 & 5.\underset{\sim}{5} & 5 \\ 5.5 & 5.5 & 6 \\ 5.51 & 5.\tilde{5} & 6 \\ 5.52 & 5.\tilde{5} & 6 \end{pmatrix}$$

A drawback is that many tildes can be written since a lot of numbers are approximations. This must be balanced with the use of fractions when those are awkward. In general we write the easiest form as long as we are certain that people will get what we mean.

PM. This is not just a mathematical issue: when we measure the length of 7 cars and arrive at an average of 3.15 meters then it is clear that rounding off to 3.2 m also involves measurement error. In physics there are rules for handling the interval but we neglect those here.

PM. A common way to *try* to express approximation is to say that $x = 0.25$ is an identity and that $x \approx 0.25$ is an approximation. This is confusing however. Is 0.25 now a pure number 1/4 and only x the approximation, or is it intended that *both* are approximations (as in the computer) ? This traditional notation puts emphasis on x and the equality sign while the issue actually lies with the number that has been rounded off. However, when we thus establish and adopt the rule that a decimal number always is pure, at least in text and not for computers, then writing $x \approx 0.25$ is no longer confusing since we then know that the approximation lies with x and not in 0.25. The tilde then comes in handy only in case of doubt or possibly when translating to and from a computer. When $x \approx 0.25$ then $x = 0.25 = 1/4$ is not excluded, while in $x = 0.2\underset{\sim}{5}$ it no longer can be the case that $x = 0.25 = 1/4$.

`NRoundAt[`*x*, *n*:0`]`	rounds number x to n decimal places
`NRoundAt[`*expr*, *n*:0`]`	rounds numbers in *expr* to n decimal places
`NTilde[`*expr*, *n*:0`]`	is like NRoundAt and puts a tilde (~) over (under) the last digit if the true value lies above (below) the approximation. Output is a String and not Numeric.

Repeated application NRoundAtTilde[NRoundAtTilde[x, 2]] does not work since input must be Numeric and not a String. If *Mathematica* does not show the number of digits, use N[*x*, *n*].

2.3 Verb and noun

2.3.1 Active versus passive

English distinguishes between verbs and nouns, like riding and a ride. Mathematical discourse does so too but mathematical language tends to use the same word, leaving the interpretation to context. An operator can be used both as an *instruction* (a verb, an activity) to do something, or as a *result* (a noun, something passive).

2.3.2 The square root and approximation

In this book we tend to use integers with few digits for mental calculations and use the computer for more involved numbers. It still is an issue how we will use the different operators.

- Evaluating $\sqrt{2.}$ gives a decimal number. The square root here is a verb, an activity of calculation and the result is an approximation up to some accuracy. This result is again a noun.
- Evaluating $\sqrt{2}$ again gives $\sqrt{2}$. It is a pure number, with infinite accuracy. It is a noun since no activity has been performed (except checking that none will be performed).

See[$\sqrt{2.}$, $\sqrt{2}$, N[$\sqrt{2}$]]

1.41421 $\sqrt{2}$ 1.41421

Students new to the issue tend to regard 1.41421 as more accurate than $\sqrt{2}$. When they see $\sqrt{2}$ then they become restless and want to compute it on the calculator, and only when it has been calculated and replaced with a decimal number then they have a sense of accomplishment. Engineers have the same habit. They have to build things and thus have to know where the number is located on the axis. Mathematicians tend to be more comfortable with $\sqrt{2}$ since it is infinitely accurate. When squared it precisely gives 2, while 1.41421 squared gives an error.

1.41421 ^ 2

1.99999

This book has a mathematical inclination but uses decimals when they support understanding.

2.3.3 Proportion, ratio, division, quotient, fraction and decimal number

2.3.3.1 Principle

The distinction between verb and noun also arises with respect to proportion, ratio, division, quotient, fraction and number. (1) Number is one-dimensional, ratio is two-dimensional. (2) Number is a result and division is a process. The number line is written as $\mathbb{R}$, the set of real numbers. The objective of division is to associate a ratio with a location on the number line.

This section gives much attention to this subject because of its importance for calculus later on. How to deal with seeming divisions by zero ? This section can best be read as a review of what you learned in elementary school. The idea and technique will have become automatic but now we rekindle awareness. Division is a process, with ratio as input, number as output, and a whole technique inbetween, with all kinds of terms for the intermediate steps. The decimal number 2.5 or the mixed number 2½ have an integer part and a fractional part. A proper fraction is less than one, which is the fraction used in the mixed number. But once you create that concept of fraction then there arises the notion of an improper fraction like 5/2, which causes some duplicity in terms with ratio.

The ratio 1 : 7 associates with the number and fraction 1/7. Apparently in some Asian systems of arithmetic the fraction does not appear altogether since there is no value added in going from "one-to-seven" to "one-seventh". That is, the value added is in pure number theory as mathematicians discovered but not in daily life for mental arithmetic. That Asian system is Euclid's two-dimensional ratio, but without his complex Theory of Proportions. It may be more efficient to use only ratio and decimal number (and drop fractions) (but this conflicts with how computers treat decimals).

Van Hiele (1973:196-204) suggests that fractions are overrated in elementary school (and vectors underrated). Exercises like adding 11/56 and 31/41 would have little value both for mathematical insight and daily life so that it does not come as a surprise that what is learned is quickly forgotten. Instead of $a \,/\, b$ it might be better to consider $a * b^{-1}$. The latter would not be a mere step for example in the transformation to a decimal notation but would amount to an entire elimination of the fractional sign. Interestingly, within *Mathematica*, the FullForm of $a \,/\, b$ already is $a * b^{-1}$. My suggestion here is to employ $a \,/\, b$ as an easy notation and indeed with that kind of interpretation, and thus without the need of immediate calculation (like long division).

PM. Henry Gurr: "The semester long effort to move the students toward a generalized and expanded understanding of proportion, exposes/reveals those students who for a variety of reasons can not or will not move in that direction. There appears to be a "Barrier". There is, however, a successful way to "penetrate" the barrier and move students toward "deeper" understanding of proportion, a topic to which we now turn. Increasing Student Understanding of Proportion and Proportionality Constants Through the Table of Proportional Quantities."

http://www.usca.edu/math/~mathdept/hsg/ProportionPaperV03.html
http://www.usca.edu/math/~mathdept/hsg/FallSemesterTableProportionQuantityFig3.html

There however is a linguistic soup to be digested before facing up that barrier.

2.3.3.2 Watch the dimensions, the order and the whole

If the dimensions (e.g. gram, meter) in a ratio are the same then the result of division is a dimensionless number; if they differ then the result is a rate.

When we take 5 minutes to fill 2 buckets with water then the process has 2/5 buckets per minute and if the cause is time T and the effect is the number of buckets B, then the relation is $B = 2/5\ T$.

Simply saying that 2 : 5 can be written as 2/5 is uninformative as to what is at hand and what the relation between cause and effect is. Here 2 units of effect require 5 units of cause, but we discussed it in a 5 : 2 ratio.

When we make bread and mix 5 parts flour, the main ingredient, with 2 parts water, the parts or shares are 2/7 and 5/7. Mixing is a different kind of process, with ambiguous cause and effect.

An odds ratio gives probability of winning versus probability of losing. It is like mixing.

Let us exchange 5 apples for 2 bananas on the market. The quantities are $q_a = 5\ a$ and $q_b = 2\ b$ where the a and b give the dimensions of the numbers. The prices are $p_a = x$ \$/a and $p_b = y$ \$/b in dollars per unit. The transaction equation in money values is $p_a\ q_a = p_b\ q_b$. The latter gives $5\ x$ \$ = $2\ y$ \$ or $x / y = 2 / 5$. If we set the price of apples at \$1 then $x = 1$ and $y = 5/2$ or p_b= 2½ \$/b. The exchange ratio is expressed in output (receiving) versus input (sending) or output/input = $2\ b / (5\ a) = 2/5\ (b/a)$. The input/output price ratio is $p_a / p_b = x\ \$/a / (y\ \$/b) = (x / y) * (b / a) = (2\ b) / (5\ a)$, or the exchange ratio again. It is a bit ambiguous what cause and effect are: the desire to sell or the desire to buy.

2.3.3.3 Input and output of the process of division

Ratio is the input of division. Number is the result of division, if it succeeds. When variables like length and width of a rectangle are in proportion then their numerical values are divided, not the concepts themselves.

- A number is a one-dimensional concept. 2/5 is *identically* equal to 4/10; and depending upon definition also to 0.4. Rational numbers can be found from a ratio of integers. Irrational numbers cannot be found from such ratio.
- A ratio is a two-dimensional concept, with numerator and denominator as separate and independent phenomena. A ratio of 2 to 5 is equivalent to a ratio of 4 to 10: *equivalence* rather than equality by identity. The elements in the ratio are still separately relevant and accessible for discussion.

Consider a van and a bus with passengers going to Amsterdam or Rotterdam. A ratio of 2 to 5 in the van is seen as different from the ratio of 4 to 10 in the larger bus, see the tickets that have to be checked and note that the numbers involved are only discovered along the way while checking. A ratio may however be simplified to find the equivalence. Two equivalent ratios are said to be proportional. If it takes 10 minutes to fill 4 buckets then this is proportional to the 5 minutes that it took for 2 buckets. At first it may seem a bit curious to make such a fuss about 2/5 versus 4/10 but when it is accepted that these are really different points of data then it starts making some sense.

The term ratio is the Latin translation of Euclid's *logos*, meaning reason but also reckoning or computation. A computer used to be a person doing calculations but that job and name have been transferrd to machines nowadays. Calculation can also be seen as algebra and reasoning indeed. Ratios are Euclid's numbers, as he did not have our methods available. Expressed a bit crude and unrespectful: for a right triangle with short sides 1 and long side $\sqrt{2}$ and another triangle scaled up with factor 2, Euclid rather states that they are in proportion, instead of that its long side is $2\sqrt{2}$ which might be found on the calculator. When you are a geometer and only know rational numbers then those irrational numbers are difficult to handle. Education is still struggling with this legacy of always looking for proportions. Mathematicians will define radians as the ratio of arc to radius, while an engineer will normalize to one radian as one unit length on the unit circle.

2.3.3.4 Technical terms for division itself

The operational terms describe the process of going from a ratio to an equivalence and then normalize to a number. These operational terms show up in teaching; and teaching tends to create names for all the small steps.

- Division is an active process. For example: long division.

```
4 /   500   \125
      4
      10
       8
       20
       20
        0
```

See Beck http://library.wolfram.com/infocenter/Courseware/140/

- A quotient in elementary school: the result of dividing dividend by divisor, especially when that result is an integer. In 30 ÷ 3 the quotient is 10. A quotient is a noun and a one-dimensional number. Quotient can be different from fraction if it is rounded down to the nearest integer. *Mathematica's* Quotient is equivalent to $\mathrm{Floor}[m/n]$ and thus always takes the integer part, rounding down. Fitting people into a bus, you round down. Buying cans of paint for your wall you round up however.
- Fraction = numerator / denominator. A fraction is a noun and a one-dimensional number.
- There is duplicity of terms for quotient and fraction except for (a) the distinction between proper and improper fractions, (b) proper fractions occur in the mixed number 2½, (c) simplification in division: If division may also be written as n / d then a fraction is a quotient not rounded down and left standing

when it can no longer be simplified.

- A *proper* fraction is intended for values below 1. As it says: part of a whole. A number is split in the integer part and the fractional part. An *improper* fraction has denominator larger than numerator, like 5 / 2. The form 2½ is a *mixed number* and consists of integer and fraction, and equals the improper fraction 5/2.
- A division of 4 by 10 can be simplified into 2 by 5, still two-dimensional, but then into one-dimensional noun and fraction 2/5. Here 1/2 can also be expressed in small letters as ½ or as a decimal number 0.5 (see above on approximation).

NB. The step of rounding down the quotient is not adopted universally so there is duplicity with fraction. Given this duplicity there is a shift in meaning. Speaking about the quotient 4/25 we tend to mean the *form* and speaking about the fraction 4/25 we tend to mean the *number.* This only happens when the form *a* / *b* is present so that both aspects indeed can be identified. This book hence regards quotient as the *form* of the formula.

NB. The exercise to teach kids at elementary school to write mixed numbers breaks down a bit when they at a later age learn about irrational numbers like Θ. The accurate number 2 + Θ/4 is hardly expressible in a mixed number with proper fraction.

2.3.3.5 Notation of a mixed number

The situation is complicated by the notation. Convention has that "two and a half" is denoted as 2½ and not as 2 + ½ (what the linguistic expression says). When we compare with $2\sqrt{2}$ then we expect to multiply, but $2½ \stackrel{?}{=} 2 * ½ = 1$. A computer has fixed positions and in typescript the expression 2½ seems recognizable enough, so that we can learn the two different codings. Handwriting can be sloppy however and we might leave a small space between the 2 and the ½. Hence this book adopts this notation:

- 5 / 2 = 2 + ½ so that addition can be a noun and a verb (like the square root). It may be called a *fractional number form* (and it still is a rational number).
- 2 + Θ/4 is adequate as well even though Θ/4 is not a proper fraction.
- The *Mathematica* standard notation of 5/2 that does not simplify.

```
a (b + 5/2)
```

$$a\left(b+\frac{5}{2}\right)$$

- Integer part and fractional part.

{num = 5/2, IntegerPart[num], FractionalPart[num]}

$\left\{\frac{5}{2}, 2, \frac{1}{2}\right\}$

Mathematica has the symbol \[ImplicitPlus] that allows the formatting of mixed fractions in input; however, output is a ratio again.

$\mathbf{2\frac{1}{2}}$

$\frac{5}{2}$

New included routines are RationalHold and Fraction. RationalHold puts a mixed number addition into Hold, and Fraction puts it into a String.

`RationalHold[`*expr*`]`	puts all Rational[x, y] in expr into HoldForm[IntegerPart[expr] + FractionalPart[expr]]
`Fraction[`*expr*`]`	Fraction[x] = ToFraction[FromFraction[x]]
`ToFraction[number]`	solves into fractional notation, e.g. 9/2 = 4 1/2, and puts this into a String
`ToFraction[`*expr*`]`	turns all rational expressions into fractional notation. Use Rationalize first if you want to turn Reals into Rationals
`FromFraction[string]`	assumes a string with only fractional notation, e.g. 4 1/2 instead of 9/2, and solves into *Mathematica* standard notation. The fractions must be denoted using the \! notation
`FromFraction[`*expr*`]`	applies FromFraction to any String in expr

RationalHold and ReleaseHold are the better formats. Division[x, y] can be used when Indeterminate stands for missing data.

- Routine RationalHold keeps the fractional part intact so that addition is not only a verb but also a noun.

a (b + 5/2) // RationalHold

$a\left(b+\left(2+\frac{1}{2}\right)\right)$

Result // FullForm

Times[*a*, Plus[*b*, HoldForm[Plus[2, Rational[1, 2]]]]]

- This is the standard confusing notation where addition seems like multiplication. The mixed number is put in a string so that it does not actually multiply (whence we read that it should be a mixed number).

Fraction[a (b + 5/2)]

$$a\left(2\frac{1}{2}+b\right)$$

Result // FullForm

Times[*a*, Plus[" 2\!\(1\/2\)", *b*]]

2.3.3.6 Proportion

Merriam-Webster on *proportion* has:

1. harmonious relation of parts to each other or to the whole (see balance, symmetry)
2. proper or equal share <each did her proportion of the work> (see also quota, percentage)
3. the relation of one part to another or to the whole with respect to magnitude, quantity or degree: Ratio
4. see size, dimension
5. a statement of equality between two ratios in which the first of four terms divided by the second equals the third divided by the fourth (as in 4/2 = 10/5) - (...)

Some hold that in mathematics only the last holds but this is not quite true since proportion can be used as share and ratio.

Thus, in case 3: If P and Q are in proportion, expressed as P to Q or $P \propto Q$ or $P :: Q$, then $P : Q$ is their ratio. The ratio then has numerical value P / Q which is the proportionality constant.

Thus, in case 5: (a) it is not equality of ratios but equivalence, (b) two rectangles are in proportion when L1 : W1 :: L2 : W2. This can be reduced to a single ratio again.

Proportions are important for dealing with reality. They can be handled easily but it requires some training. There are two reasons why proportions feature strongly in science and mathematics education:

- Students appear to have difficulty in manipulating the terms:

$$\left(\frac{a}{b} == \frac{c}{d}\right) \Leftrightarrow \left(a == \frac{b\,c}{d}\right) \Leftrightarrow a\,d == b\,c \Leftrightarrow \left(\frac{a}{c} == \frac{b}{d}\right) \Leftrightarrow \left(\frac{d}{b} == \frac{c}{a}\right)$$

- There is Euclid's Theory of Proportions as a more primitive theory and precursor to modern arithmetic. It is geometry before analytic geometry. Our culture and language is infused with this traditional geometrical approach of always looking for proportions. We are not adapted yet to our advance in numerical capabilities.

Fortunately, there is much equivalence between proportion, ratio, division, quotient, and fraction, and thus we can often exploit these terms for linguistic variation in text editing just to create more appealing texts. Henceforth this book uses ratio as a more beautiful word than fraction or quotient, and we write a / b. If we intend the Greek Theory of Proportions then we support this by writing $a : b$.

NB. This chapter gives much attention to this subject because of its importance for calculus later on. How to deal with seeming divisions by zero ?

2.3.4 Active division and the dynamic quotient

2.3.4.1 Going from arithmetic to algebra

There is a difference between numeric 2 / 2 = 1 and algebraic $a / a = 1$ when possibly $a = 0$: but curiously the same slash is used. Division is a verb and an active process while a fraction is a noun and passive result, but with algebra there is a small but important difference in situation, since algebra might leave the process incomplete since no number is specified. Let us extend division with the handling of algebra:

- a // b will be the *dynamic quotient*. The process of division forks out into the two suboperations, either for numbers or for variables.

Thus 5/2 is a result and improper fraction, and may still be simplified into the fractional number form, but when we want to express that the focus should be on activity then we use 5 // 2, and then we do not know yet of any result and we still have to find out whether it indeed is an improper fraction or not. Since division is a dynamic or active process by itself this new procedure cannot be called dynamic division. It may be ambiguous for the term quotient whether the result must be rounded down but in this book we don't do so. Hence we can use the term. Note that we might use the term ratio again but given the historical burden we do not do so. A ratio is not necessarily expressed as a fraction while with the dynamic quotient we try to do so.

PM. In *Mathematica* $f[x]$ can also be written as $x \,//\, f$ but in this case we presume that it is clear that f is a function so that the confusion does not arise or is easily resolved.

PM. In *Mathematica* we do not need a special symbol for the dynamic quotient since it is merely simplification:

a / a // Simplify

1

2.3.4.2 Division by zero ?

What about the danger of dividing by zero ? When $3\,x = 5\,x$ then the solution is $x = 0$. If we divide both sides by x then we get $3\,x\,/\,x = 5\,x\,/\,x$ or $3 = 5$. Such problems arise easily with more complicated formulas when the horrors of simplification are not easily seen.

To make it strict, let $y\,/\,x$ be as commonly used and the dynamic quotient $y\,//\,x$ be the following process or program:

$y\,//\,x \equiv \{\,y\,/\,x$, unless x is a variable and then: assume $x \neq 0$, simplify the expression $y\,/\,x$, declare the result valid also for the domain extension $x = 0\,\}$.

As always: it depends upon the application whether this operation is valid. We cannot just simply assume that we can always do this but have to check whether it applies.

For example: From $3\,x = 5\,x$ we still cannot go to $3\,x\,//\,x = 5\,x\,//\,x$ since again $3 = 5$. The reason is that $3\,x = 5\,x$ implies that $x = 0$ or that x is a constant and not a variable. Hence we would have $3\,x\,//\,0 = 5\,x\,//\,0$ which still is the standard division, and division by zero is undefined. The acute distinction is between true variables and seeming variables.

For example: The expression $\frac{x\,y}{x^2 + y^2}$ is undefined for $x = 0$ and $y = 0$. When we consider $y = a\,x$ for the case of proportional dependence then however a nonzero expression arises.

$$\frac{x\,y}{x^2 + y^2} \to \frac{a\,x^2}{a^2\,x^2 + x^2} \to \frac{a}{a^2 + 1}$$

This simplification however assumes that there really is such dependence. For a value at $x = 0$ and $y = 0$ we also would need a particular value of a for, as it stands now, any a would do.

2.3.4.3 Handling the domain

Regard the quotient $Q = \left(1 - x^2\right) \big/ (1 - x)$. Since $1 - x^2 = (1 + x)(1 - x)$ we find that Q simplifies to $1 + x$. We namely eliminate $1 - x$ from both numerator and denominator. This does not necessarily mean a division, since there can be simplification in the two separate terms, and actually a revision of what we really divide. It may also mean that we indeed adjust the domain: then simplification becomes an alternative to limit theory in finding a value where division by zero is undefined.

■ Standard simplification.

```
(1 - x^2) / (1 - x) // Simplify
```

$x+1$

Of course, dividing by 1 - x is tricky when x = 1 since then we divide by zero. It thus depends what we want with this Q. After the simplification, x = 1 gives Q = 2. The traditional approach requires us to specify the two separate cases. The dynamic quotient is faster.

Traditional definition overload	*With the dynamic quotient*
$(1 - x^2) / (1 - x) = \begin{cases} 1 + x & \text{if } x \neq 1 \\ 2 & \text{if } x = 1 \end{cases}$	$(1 - x^2) // (1 - x) = 1 + x$

In the traditional approach we also would use limit theory (not explained here) to determine that $Q \to 2$ in the limit for $x \to 1$, while with the dynamic quotient we simply extend the domain. The crux is that we rely on the algebraic form to find a value where the domain used to be undefined. We will use this property in calculus later on.

2.4 Multiplication again

2.4.1 Multiplication of terms

Say that you want to multiply (1 + 1/2) with (2 + 3/4). Our strategy is to let *Mathematica* take the tedium of calculation.

■ This is how *Mathematica* does it.

```
{term[1] = (1 + 1/2),   term[2] = (2 + 3/4),   term[1] * term[2]}
```

$\left\{\frac{3}{2}, \frac{11}{4}, \frac{33}{8}\right\}$

Mathematica's strategy on 1 + 1/2 is to use the same denominator, so that 2/2 + 1/2 becomes 3/2. In the same way 2 + 3/4 = 8/4 + 3/4 = 11/4. For multiplication it is a fast method indeed.

■ The conventional way of writing fractions is dangerous for handwriting and does not fit the development of algebra (a + b)(c+ d) later on.

```
Fraction[term[1]] Fraction[term[2]]
```

$1\frac{1}{2}\ 2\frac{3}{4}$

- This seems clearer, see also the graph we already presented in §1.3.4.

RationalHold[term[1]] RationalHold[term[2]]

$$\left(1+\frac{1}{2}\right)\left(2+\frac{3}{4}\right)$$

- PM. This should not be a surprise. Complex numbers (above) use it a lot.

(a + b) (a − b) // Expand

$a^2 - b^2$

2.4.2 The solution of squares

Suppose that we have a square with a surface of 2 square meters. What are the sides of this square ? This is an inverse question. A square has equal sides, let us call one side x, thus $x^2 = 2$. Finding x is inverse to surface calculation.

- The routine Solve gives the answer. NSolve would give numbers.

Solve[x^2 == 2, x]

$\{\{x \to -\sqrt{2}\}, \{x \to \sqrt{2}\}\}$

There is again the distinction between verb versus noun. $\sqrt{2}$ is a result, a noun. Solving an equation is an active process, a verb, and now generates two solutions.

New students tend to think in this way: Seeing $x^2 = 2$ they conclude that x must be the square root of 2, and then they proceed accordingly (entering it in the calculator). They use $\sqrt{\blacksquare}$ as a verb. They get positive feedback since for a square and its sides the negative part of the solution drops out. However, this is mathematically incorrect. Later when we consider co-ordinates the two solutions appear to be relevant. Given the confusion amongst new students it is useful to distinguish between Sqrt[ ] as the passive result and DoSqrt[ ] as the active solver. Using DoSqrt as a stepping stone they soon will write Solve.

- DoSqrt[y] generates the solution of $x^2 = y$.

DoSqrt[2]

$\{-\sqrt{2}, \sqrt{2}\}$

2.5 Recovered exponents

When 10^3 is calculated to be 1000 then the exponent 3 disappears into the resulting number. When we have 1000 and we use base 10 and if we want to recover the exponent then it appears to be 3. The name of this operation can be "recovered exponent", or "rex" in short, with 3 = rex[10, 1000]. This name of the operation expresses that we are only recovering the exponent. Currently mathematics uses Napier's word "logarithm" but this is rather opaque for what is just a recovered exponent. It may not matter whether your wife is called Mary or Nicole but try it out. Hence, rex[b, x] = p if and only if $x = b^p$.

When the input is numerical.

Rex[10, 1000]

3

Rex is purely another name. When the input is not numerical then Log still appears in output. The default base is e = 2.71828...

Rex[x^a y^b] // PowerExpand

$a \log(x) + b \log(y)$

$Rex can be used as a name that does not further evaluate. In TraditionalForm it displays as lower case rex.

$Rex[x^a y^b] == (Rex[x^a y^b] // PowerExpand // To$Rex)

$\text{rex}[x^a\, y^b] = a\,\text{rex}[x] + b\,\text{rex}[y]$

An advantage is that pure numerical input does not evaluate either.

Result /. {a → 2, x → e, b → 10, y → e^2}

$\text{rex}[e^{22}] = 2\,\text{rex}[e] + 10\,\text{rex}[e^2]$

$Rex also displays the base as a suffix.

$Rex[10, 1000]

$\text{rex}_{10}(1000)$

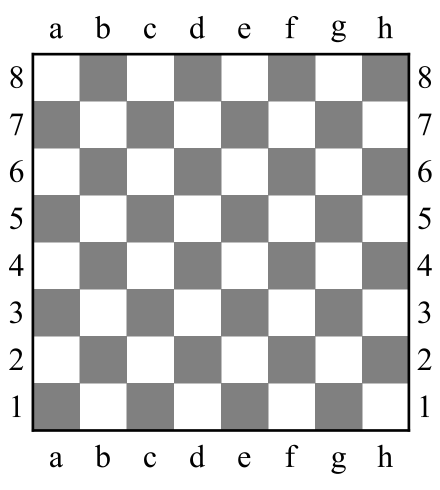
3. Co-ordinates

3.1 Two axes

3.1.1 What co-ordinates are

Co-ordinates give information to locate something. For a person it might be a telephone number or an address. When you meet people and want to contact them later then you can ask for their co-ordinates and they will give you their business card. In the same way for the plane: we use a system of co-ordinates so that every point on the plane can be identified.

A chess board is a familiar system of co-ordinates. The columns are labelled with the first eight letters of the alphabet (lower case makes for better reading) and the rows are just counted. White starts at the bottom and black at the top. The square at the bottom right hand at h8 will be white. The queen of white will start at d1 and the queen of black will be opposite at d8.



3.1.2 X and Y

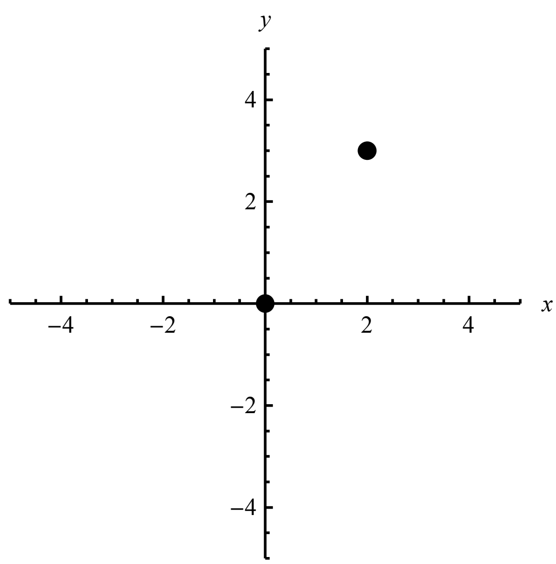
With a ruler on a piece of paper we draw a *horizontal* line and we call it the x-axis. Perpendicular to it we draw a *vertical* line and call it the y-axis. To identify what axis is what, we label the axes x and y.

Where the lines cross will be called the *point of origin*. From there we can step right, left, up or down.

We can put numbers on the axes. We copy numbers from the ruler to the axes. The origin will get the number 0. On the horizontal axis we count positive numbers to the right and negative numbers to the left. On the vertical axis we count positive numbers up and negative numbers down. When we go along an axis from 1 to 2, or from 2 to 3, etcetera, then we will call this a full *step*.

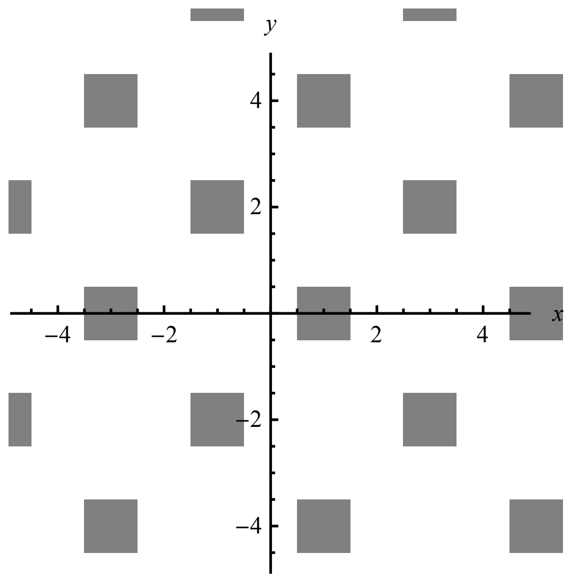
We can use curly brackets around two numbers to identify a point on the plane. To start with, {0, 0} will denote the *point of origin*. Then, for example, {2, 3} will mean the point that we can find by moving from the origin, first stepping to number 2 on the horizontal axis and then making 3 steps up.

When you have copied this then you would get a graph like the one below. In this present graph we have put thick dots at {0, 0} and {2, 3}.



3.1.3 Practice makes perfect

It can be good practice to step through this maze using integer points only and without hitting a square. Start at {1, 2} and try to get to {-4, -3}.



A path is {1, 2} to {2, 2} to {2, -3} to {-4, -3}.

Another exercise is to assign letters to points and translate a word into a list of numbers, so that we get a coded message. Try to code FINE using F = {0, 0}, I = {-3, 4}, N = {4, -2} and E = {-4, -3}.

3.2 A major discovery

Now that we have assigned numbers to the plane we can directly link arithmetic and algebra to points on the plane. This is a major discovery. This is what analytic geometry is about.

The core idea of analytic geometry is to combine spatial issues with formulas (algebra) including numbers (arithmetic), and indeed copy those onto a piece of paper or our mental image of such a plane. The pheonomena are interlinked, as we saw that the surface of a rectangle can be seen as deriving from a calculation that again relates to algebra. It took mankind some millenia to grow aware of these linkages but once the idea was developed by Nicole d'Oresme (1323 - 1382), Pierre de Fermat (1601/8 - 1665) and René Descartes (1596 - 1650), progress has been great.

Part II. Line, circle and vector

4. Line

4.1 A map of the plane

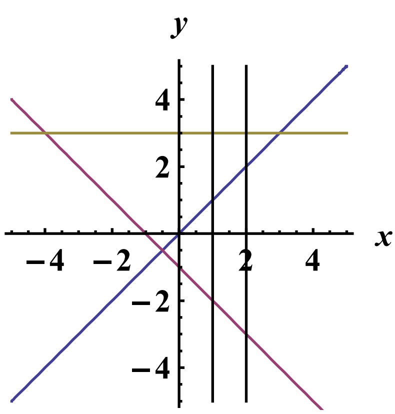
4.1.1 Horizontal and vertical lines, and diagonals

Horizontal lines are the (**zero,** main) horizontal line, the **first** horizontal line, the **second** horizontal line, ... There are also the **negative first** horizontal, ... etcetera. Vertical lines can be named as the (**zero**, main) vertical line, the **first** vertical line, the **second** vertical line, ...

There are also diagonals that are either rising or declining. Diagonals through the origin are the (**zero**, main) diagonals. The **first** rising or declining diagonals pass through {0, 1}. Negative **first** rising or declining diagonals pass through {0, -1}.

In the graph below we see the following lines:

- The (zero, main) horizontal line or the x-axis and the third horizontal
- The (zero, main) veritical line or the y-axis and the first and second vertical
- The (zero, main) (rising) diagonal line
- The negative first declining diagonal line



See if you can identify the point of intersection of the negative fourth vertical and the second horizontal (not shown). Indeed, this is {-4, 2}.

See if you can find the intersection of the negative first rising diagonal and the third vertical. E.g. first take the intersection of the (zero) (rising) diagonal and the third vertical, and then make a parallel shift downwards. The answer is {3, 2}.

What is the intersection of the tenth declining diagonal with the sixth vertical ? Here it is {6, 4}.

PM. Diagonal lines cross both axes so there is a choice how to name them. It is best to do so as we just have done. PM. In a classroom situation we can ask students to put their name at their chosen line so that the numerical names get some flair.

4.1.2 Properties of these lines

4.1.2.1 Properties without numbers

Properties of these lines are:

- Lines have a slope. The main diagonal is upward sloping. The main declining diagonal is downward sloping. A horizontal line is flat: we can say that *no slope = zero slope*.
- Lines may rise or fall. A vertical line does both. A horizontal line does neither. There are only four options.
- Lines might be parallel. This happens if and only if their slopes are equal.
- Lines might intersect. That point is unique. Overlapping lines are the same.
- If we have a point and a slope then we have a line.
- If we have two points then we have a line (and thus a slope).

These properties cause some questions. Can we determine the point of intersection? Can we calculate the slope ?

4.1.2.2 A slope can have a number assigned to it

Let us consider the slope of the diagonals. For example the points {0, 0}, {3, 3}, {5, 5} and {10, 10} are all on the same line with the same slope.

- The slope s is defined as: when x takes one step to the right then y changes by s. Alternatively said: slope s is the *rise* divided by the *run*. We can calculate (and check):

- For a horizontal line the value of y does not change and thus the slope is 0. For a vertical line the slope is indeterminate (∞) (since we cannot take a step sideways).
- For the main rising diagonal the step from {1, 1} to {2, 2} gives slope 1.
- For the main declining diagonal the step from {1, -1} to {2, -2} gives slope -1.
- For the main diagonals (only) we find the slope as the ratio $s = y / x$.
- The other diagonals are parallel and thus have the same slopes.

4.1.2.3 Similarly for a starting height

Next to a slope there is also something that we can call the starting height. Regard the third rising diagonal. We find two properties:

- Its slope is 1 since it is parallel to the main diagonal. Every time that x rises by 1 then also y rises by 1.
- The vertical difference between points on the third rising diagonal and points on the main diagonal always is {0, 3} which is precisely the value on the vertical axis. We call the latter point the *starting point*.

$$\begin{pmatrix} \text{Third rising diagonal} & \{0, 3\} & \{1, 4\} & \{2, 5\} & \{3, 6\} & \{4, 7\} & \{5, 8\} & \{6, 9\} \\ \text{Main rising diagonal} & \{0, 0\} & \{1, 1\} & \{2, 2\} & \{3, 3\} & \{4, 4\} & \{5, 5\} & \{6, 6\} \\ \text{Difference} & \{0, 3\} & \{0, 3\} & \{0, 3\} & \{0, 3\} & \{0, 3\} & \{0, 3\} & \{0, 3\} \end{pmatrix}$$

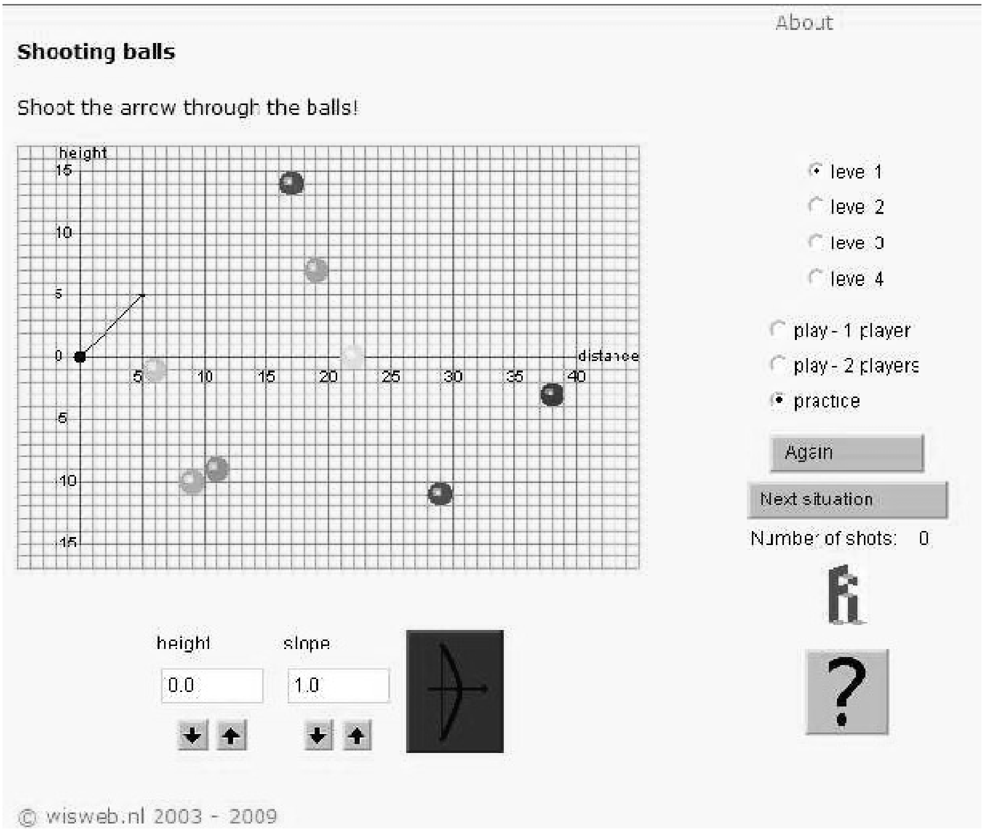
4.1.2.4 Summary conclusion

We can generalize this result for all diagonals and horizontal lines. All those lines have a starting point and a slope. Vertical lines are the exception however.

4.1.3 Exercise

Now that we have built up experience with lines and slopes and have assigned names and numbers, we are well-equipped to wonder about arbitrary lines. Can we use the notions of a starting height and a slope to say something about lines in general ? An excellent applet for exploration is given by the following link:

http://www.fi.uu.nl/toepassingen/00065/toepassing_wisweb.en.html.



Shooting Balls at Wisweb.nl, authored by Michiel Doorman and programmed by Petra Oldengram.

Note:

- Two points always give a line. Thus it is possible to hit two balls with one shot (except when the two balls are on a vertical line).
- You can direct the shot by guessing visually but this can give errors and the result is more certain by entering numerical values. (When you play at a higher level then the visual method of control also disappears.)
- The co-ordinates of the balls can be found from reading the axes or by clicking on the balls. How can that information be transformed into a shooting instruction ?

Homework: Play the applet, check that you can hit two balls by one shot indeed. Make a table: recording the co-ordinates of two pairs of balls and the shooting instruction (starting value and slope) (shown by the applet), and your success in shooting them both. When you have a success then try to design a formula that describes how these numbers are related, so that you can calculate the shooting instruction from what you know about the balls. Try the formula on a new pair of balls. Hint: Try first to hit a single ball from {0, 0}.

4.2 Algebraic formulas

4.2.1 Sorting and cataloguing and finding the formulas

When we compare two points, say {1, 2} and {1, 5}, then we compare the separate entries in the lists. We find $1 = 1$ and $2 \neq 5$. In general we say that $\{x, y\} = \{3, 7\}$ when $x = 3$ and $y = 7$.

Let us sort out what we have found and catalogue the different cases:

- For the horizontal line with the points {1, 5}, {20, 5} and {whatever, 5} we find that y is always 5. A good formula for this line is $y = 5$.
- For the vertical line with the points {3, 1}, {3, 5} and {3, whatever} we find that x is always 3. A good formula for this line is $x = 3$.
- For the main diagonal with the points {0, 0}, {1, 1} and {whatever, same} we find that y is always x. A good formula for this line is $y = x$.
- For the main declining diagonal we find $y = -x$.
- For the two main diagonals we thus find $y = s\,x$ for $s = 1$ or $s = -1$.
- For the third rising diagonal we found that y is always 3 higher than x. Thus a good formula is $y = x + 3$.
- When we put $x = 0$ in the formula $y = x + 3$ then this gives $y = 3$. This thus gives the point {0, 3}.
- The starting point $\{0, c\}$ can be given by a value c on the y-axis.

The expressions $x = 3$, $y = 5$, $y = s\,x$ and $y = x + 3$ are equations. A crucial finding thus is that these lines can also be identified by their equations.

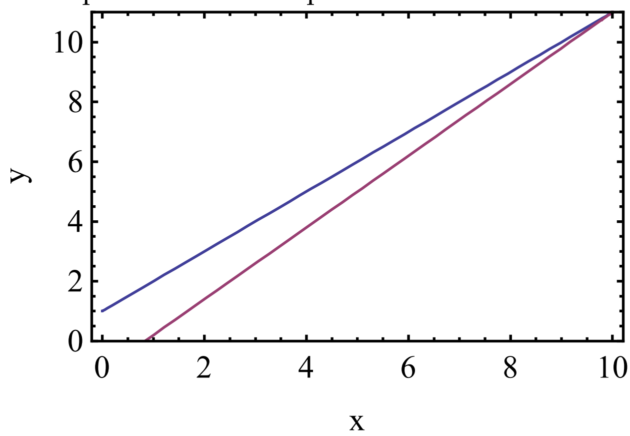
Can you identify the line with the formula $y = -x - 5$? Indeed, it is the negative fifth declining diagonal.

4.2.2 Notation

An umbrella formula for our lines is $y = c + s\,x$. Clearly we now allow all kinds of values of s other than 1 and -1. The starting values c may take any value other than integers too.

An example is a bucket that can contain 10 liters of water. It already contains one liter. The bucket is filled from a faucet with water running one liter per minute. How long does it take to fill the bucket ? Explain that this process is only approximately linear. These two questions were easy but try to identify the graphs now.

- This plot has x and y and two lines. What is time and what is the bucket, which line represents the example ?



Homework: Try the applet again. Clarify how your understanding has changed compared with the former time. PM. Balls are allocated at random so scores are not quite comparable.

4.2.3 Definition of a functional relation

We say that y (the effect) is a *function* of x (the cause) when each x has precisely one value of y. We write $y = f[x]$.

Above definition $y = c + s\, x$ is the functional definition of the line.

The drawback of the use of functions is that vertical lines are not covered. We say that there is a *correspondence* when x (the cause) can be associated with more values for y. Thus $x = 7$ is a correspondence since it applies for all y.

Presumed cause and presumed effect may actually not be related. Namely in these cases: (1) for any cause the effect remains the same: a horizontal line, (2) the cause has one value and consequences are over the whole range.

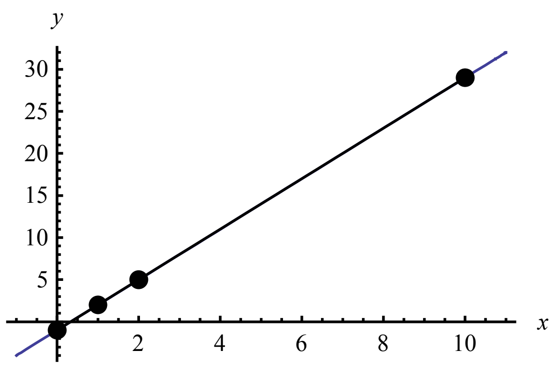
For $y = f[x]$ it is often useful to solve into $x = g[y]$ for some g. When such a function exists then it is called the *inverse* function. To identify that a function and its inverse belong to each other this can be expressed in the name by writing $x = f^{-1}[y]$. Lines generally have an inverse. For example, $y = f[x] = 3 + 2\, x$ has inverse $f^{-1}[y] = (y - 3) / 2$. When you make a graph of $f[x]$ and flip the (transparent) paper then you have the graph of $f^{-1}[y]$.

4.2.4 Construction of a line by means of a table

With the formula for a line we can calculate some points, locate these on the graph, and then use a ruler to draw the line through those points. It takes only two points to find a line. However, we might add some more just for checking. The calculations can best be put in a table for easy overview.

- For the line $y = 3\,x - 1$ and arbitrary points.

$$\begin{pmatrix} \text{Effect} & -1 & 2 & 5 & 29 \\ \text{Cause} & 0 & 1 & 2 & 10 \end{pmatrix}$$



`CauseEffectTable[`*expr*, *x*, *lis*`]`	for functional *expr* with cause *x* and *lis* of causes
`CauseEffectTable[`*f*, *lis*`]`	for function *f* and *lis* of causes

The table gives the effect in the top row. This fits the general layout that the cause is on the horizontal axis and the effect is on the vertical axis. Use text, formula, table and graph to discuss topics.

Homework: Try the applet again. Clarify how your understanding has changed compared with the former time. Explain how it is with cause and effect. Can you use a line to determine the co-ordinates or can you use the co-ordinates to determine the line ? Explain what method the applet uses. (It does not allow you to put balls there.)

4.2.5 The general formula for a line

The general formula for a line is: $p\ x + q\ y = r$. This formula is better than the functional form since it deals with $s = \infty$. When we have two lines then we get three possibilities: (1) parallel, (2) intersect, (3) overlap.

Special cases are (prove this):

- Vertical lines appear when we take $p = 1$ and $q = 0$, for these give $x = r$.
- Horizontal lines appear when $p = 0$ and $q = 1$, for these give $y = r$. Show this for two points.
- If $q \neq 0$, then $y = r / q - p / q\ x = c + s\ x$.

For the following two lines, can you say what is special about them ?

$$2x + 3y = 5$$
$$4x + 6y = 10$$

We say that the lines associated with these formulas overlap or are the same line. If we multiply the first line with 2 then we get the second line. We can multiply both sides of the general formula with a constant $c \neq 0$. Then we get $c\,p\,x + c\,q\,y = c\,r$. We can write this as $P\,x + Q\,y = R$. If $q \neq 0$, then $Q \neq 0$. Then $c = r / q = R / Q$ and $s = -p / q = -P / Q$. It appears that there is no difference between these lines though the formulas seem to suggest that there might be a difference.

PM. A proof voor vertical lines is: Take arbitrary points {3, 1} and {3, 5}. Then

$$p\,3 + q\,5 = r$$

$$p\,3 + q\,1 = r.$$

Thus $0 + q\,4 = 0$ from subtraction, or $q = 0$. The first equation then gives $p\,3 = r$. The equation of the line thus is $p\,x = r = p\,3$ and that gives $x = 3$.

PM. When you draw a line you might say that you draw an infinity of lines all at the same time, namely all overlapping; though it is more accurate that it still is the same line in different shapes.

4.3 Implications

Important implications are that: (1) a point and a slope give a line, (2) two points give a line, with constant and slope, (3) two lines may solve into a point, (4) a line through the origin has slope only and gives proportionality.

4.3.1 A point and a slope give a line

Suppose that we want a line with slope -2 through {3, 4}. If $\{x, y\}$ is on the line and we take {3, 4} as the origin then the line goes through $\{x - 3, y - 4\}$, as a ray with slope -2:

$$(y - 4) / (x - 3) = -2$$

$$y - 4 = -2\,(x - 3)$$

$$y = -2\,x + 10.$$

Check that the point fits the line: 4 = -2 * 3 + 10.

Let us write Δx for "a step by x" and Δy for "a step by y". Δ is Greek capital delta, or D, and stands for *difference*. Then numerically $\Delta y = s\,\Delta x$ and this can be transformed into $s = \Delta y / \Delta x$.

4.3.2 Two points give *s* and *c*

We have already been calculating slopes and constants in the applet but let us now make it official: how it is done and why it works. Let us determine the line through {4, -3} and {10, 5}. The steps are as follows, also for the formal case.

4.3.2.1 The basic algorithm

Explanation	*Example*	*Theory*
Two points are given	$\{4, -3\}$ and $\{10, 5\}$	$\{a, b\}$ and $\{u, v\}$
Select one as the origin	$\{0, 0\}$ and $\{10-4, 5+3\}$	$\{0, 0\}$ and $\{u-a, v-b\}$
The slope of a ray	$s = (5+3)/(10-4) = 4/3$	$s = (v-b)/(u-a)$
$\{x, y\}$ is also on the line	$\{x-4, y+3\}$	$\{x-a, y-b\}$
$\{x, y\}$ gives a slope too	$s = (y+3)/(x-4)$	$s = (y-b)/(x-a)$
Solving the equal slopes	$y+3 = s(x-4) = 4/3(x-4)$	$y-b = s(x-a)$
Rewriting	$y = 4/3x - 16/3 - 3$	$y = sx - sa + b$
The constant	$c = -16/3 - 3 = -25/3$	$c = b - sa$
The result	$\mathbf{y = 4/3x - 25/3}$	$y = sx + c$
Checking on $\{4, -3\}$	$-3 = 4/3 * 4 - 25/3$	

4.3.2.2 With a computer routine

- Above gives the explanation. PM. This routine breaks down for a vertical line.

```
TwoPointsToLine[x, {10, 5}, {4, -3}]
```

$$\frac{4x}{3} - \frac{25}{3}$$

$$\left(1 + \frac{1}{3}\right)x + \left(-8 - \frac{1}{3}\right)$$

4.3.2.3 With the functional form

The above is not the only method. Another algorithm is to take the general formula of the line $y = c + s\,x$ and then substitute the two points {4, -3} and {10, 5}. This gives a system of two equations with the two unknowns s and c:

$$\begin{pmatrix} 5 = c + 10\,s \\ -3 = c + 4\,s \end{pmatrix}$$

Substracting the two equations gives (5 + 3) = s (10 - 4). Again s = 4/3 and then either equation gives c.

4.3.2.4 With a table

Above procedure can be put into a table. Δ means the difference. Put y above x

both in the table and in Δy / Δx.

$$\left\{ \begin{array}{c|ccc} & A & B & \Delta \\ \hline y & 5 & -3 & 8 \\ x & 10 & 4 & 6 \end{array}, \quad \begin{array}{l} s = \Delta y / \Delta x \to \frac{4}{3} \\ c = yA - s * xA \to -\frac{25}{3} \end{array} \right\}$$

4.3.3 Two intersecting lines

We turn above problem upside down and consider the question how to calculate the point of intersection of two lines. If at least one of the lines is vertical or horizontal then the question is relatively easy.

- When both are horizontal or both are vertical and they do not overlap then there is no intersection. The equations are not consistent.
- If one line is horizontal or vertical then the solution is straightforward: we substitute the particular value and solve for the unknown.
- When no line is horizontal or vertical then there is more algebra. It is simplest to first normalize to the functional form so that $y = c + s\ x$ now with constant and slope given.

4.3.3.1 The basic algorithm

Example lines are:

$$\begin{pmatrix} y = 4 - 2\,x \\ y = -5\ +\ 7\,x \end{pmatrix}$$

Subtraction gives 0 = (4 + 5) + (-2 - 7) x, thus $x = 1$. Then $y = 2$.

4.3.3.2 Something worth of note

Suddenly, though, you may notice something peculiar. Let us compare with the problem in the former subsection. What do you see ?

$$\begin{pmatrix} y = 4 - 2\,x \\ y = -5\ +\ 7\,x \end{pmatrix} \quad \text{compared to} \quad \begin{pmatrix} 5 = c + 10\,s \\ -3 = c + 4\,s \end{pmatrix}$$

We reorder into the same form:

$$\begin{pmatrix} 4 = y\ +\ 2\,x \\ -5\ =\ y - 7\,x \end{pmatrix} \sim \begin{pmatrix} 4 = c\ +\ 2\,s \\ -5\ =\ c - 7\,s \end{pmatrix} \quad \text{compared to} \quad \begin{pmatrix} 5 = c + 10\,s \\ -3 = c + 4\,s \end{pmatrix}$$

There are two lines with two unknowns. This is the same structure. It does not matter how the variables are labelled. The conclusion is that we can use the algorithm above ! In fact, we used it !

Why does this work ? Well, the line $p\ x + q\ y = r$ holds for $\{x, y\}$ but also for the coefficients $\{p, q\}$. An example is in economics. Your income $Z = h\ w$ consists of hours worked h and your wage w. You spend it on a quantity q consumed at price p and some savings S. Then $S = h\,w - p\,q$. We can describe this situation in terms of prices $\{p, w\}$ or quantities $\{q, h\}$. If we know the prices we can solve for the quantities; if we know the quantities we can solve for the prices.

We must make this translation: $x \to s$, $y \to c$. To prevent us from getting really confused we now use upper case letters. The coefficients are the slope S and constant C per line, and the X and Y are proper $\{x, y\}$ again.

$$\left\{ \begin{array}{c|ccc} & A & B & \Delta \\ \hline C & 4 & -5 & 9 \\ S & 2 & -7 & 9 \end{array}, \begin{array}{l} X = \Delta C / \Delta S \to 1 \\ Y = CA - X * SA \to 2 \end{array} \right\}$$

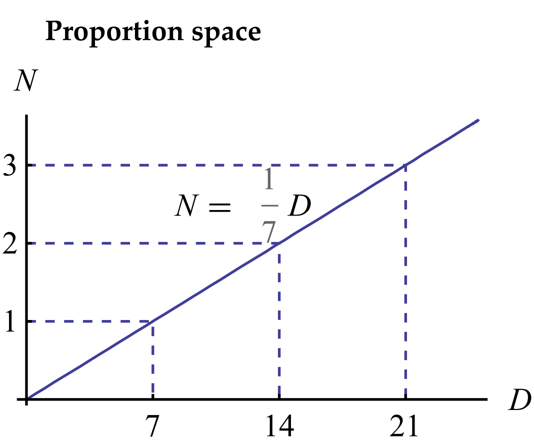
Thus the same answer with $x = 1$ and $y = 2$.

When you are new to the subject then this is basically something to be only aware of. For actual calculations you probably work best by using the terms that you are familiar with. Gradually, though, the structural identity becomes a subject of study itself.

4.3.4 Proportionality

A proportion (ratio) is a point in proportion space. Proportion is two-dimensional while a fraction is a one-dimensional number. We already discussed ratio or proportion above. The discussion there was hindered by the lack of a graphical display since we had not yet presented lines in the two-dimensional plane.

Let numerator N and denominator D be in proportional ratio of 1 to 7, then $N : D :: 1:7$, and we can write the fraction $f = N / D = 1 / 7$, or $N = 1/7\, D$. Such proportional relations always give rays through the origin.

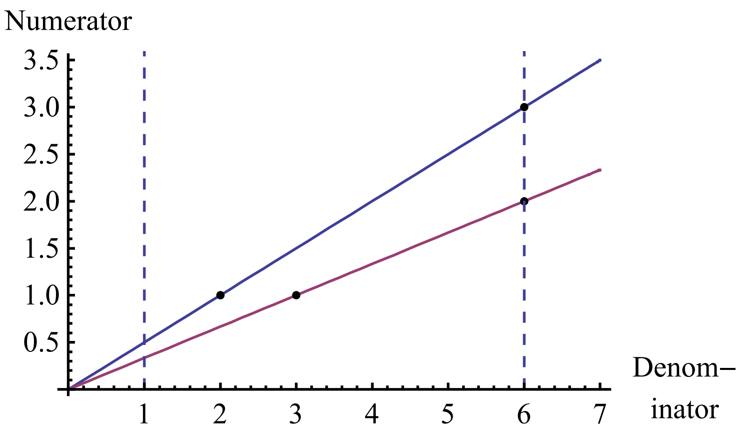


A proportion $N : D$ is two-dimensional $\{D, N\}$ while fraction N / D is a single number, namely the slope of the ray. Proportion space $\{D, N\} = \{D, f\,D\} = \{1, f\}D$ has a reverse writing order of fraction $f = N / D$. In the sense by Euclid, proportion would be the equivalence of ratios, and thus proportion would be his word for our ray as the collection of all equivalent points. Instead we have found that it is more efficient to refer to number and function, within the context of numbers and functions in general.

We have a preference to express a ratio in a unit of D = 1. A ratio of apples to oranges of 1 : 7 is preferably expressed as oranges to apples of 7 : 1, inverting the plane. This preference may conflict with the modern convention on the cause & effect choice of co-ordinates. If $N : D :: 1 : 7$ stands for the effect : cause ratio and if indeed 1 unit of effect requires 7 units of cause, then we may say this in the reverse as 7 : 1 for cause to effect, but we would retain above diagram.

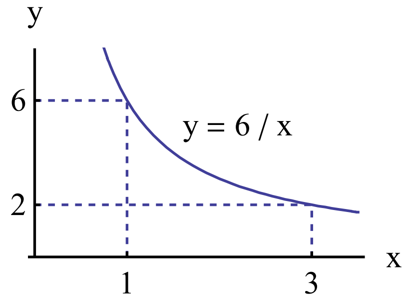
A conventional way to learn about fractions is from cutting up pies and cakes. Adding up fractions can become an intricate matter in that manner. A perhaps simpler and clearer way but requiring 2D graphics is the method shown in the figure below. Fraction ½ can be denoted by the slope of two steps to the right and 1 up. Fraction 1/3 is three steps to the right and one up. Adding them can be done by taking a common multiple of steps, say 6. Extending the lines gives us 3/6 + 2/6 or 5/6 as the total. If we look at the vertical at 1 instead then we would get decimal fractions. Note that the labels on the axes matter.

Proportion space: Adding 1/2 and 1/3



PM. Given all this discussion about proportionality and linear relations it is useful to give an example that is not linear. A good example is inverse proportionality. Let the product of x and y result into a fixed number, e.g. $x\ y = 6$. This can be written as $y = 6 / x$ so that y is proportional to $1 / x$ or the inverse of x.

- When you want a room to have an area of 6 m^2 then you have still some choice as to length and width.



4.4 Dynamic quotient

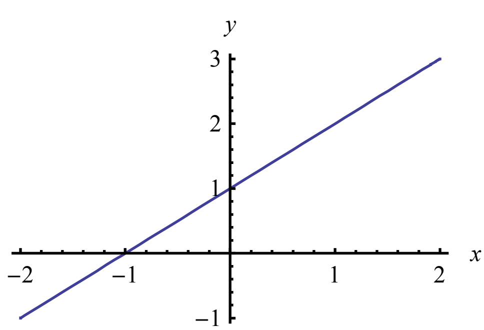
Above we defined the dynamic quotient y // x. In that discussion we could not yet use a graph since the system of co-ordinates had not been defined yet. The dynamic quotient plays a key rol in calculus so it is useful to return to it, now better equipped.

An expression like $(1 - x^2) / (1 - x)$ would be undefined at $x = 1$ but the natural tendency is to simplify to $1 + x$ and not to include a note that $x \neq 1$ since there is nothing in the context that suggests that we would need to be so pedantic. Standard graphical routines also skip the undefined point (see the graph below). Traditional teaching and math exam practice is to use the division $g[x] / f[x]$ as a hidden code that must be cracked to find where $f[x] = 0$. Students fail the exam if they do not crack that code. Rather the reverse applies: that such undefined points must be explicitly provided if those values are germane to the discussion.

- A dynamic quotient assumes variables and domain flexibility. A warning and a dot in the graph is required if we want $x = 1$ specifically be excluded.

Simplify[(1 − x^2) / (1 − x)]

$x + 1$



We have discussed the division on $3\ x = 5\ x$. A classic example of the inappropriateness of division by zero is $(x - x)\ (x + x) = x^2 - x^2 = (x - x)\ x$, where division by $(x - x)$ causes $x + x = x$ or $2 = 1$. This indeed is another good example that, indeed, we should never divide by zero. Thus distinguish between:

- creation of a quotient such as putting "/" or "//" between "$(x - x)\ (x + x)$" and "$(x - x)$"; here quotes indicate the literal expressions and not their simplifications
- handling of a quotient such as $(x - x)\ (x + x)$ (/ or //) $(x - x)$ once it has been created.

The first is the great sin that creates such nonsense as 2 = 1, the second is only the application of the rules of algebra. In this case, the algebraic rules tell us that $x - x = 0$, which is a constant and not a variable. Simplification generates a value Indeterminate, and this would hold for both / and //. In comparison the static quotient $a\ (x + x) / a$ generates $2x$ for $a \neq 0$ and is undefined for $a = 0$. However, the dynamic quotient $a\ (x + x)\ //\ a$ renders $2x$, and we would be committed to it as our working hypothesis, also if later in the deduction we would meet a value $a = x - x = 0$: at such a point in the deduction we can re-evaluate the case and determine whether this value is germane to the discussion and whether the dynamic quotient was wrongly applied or not.

Once the idea is clear, there is no need to be very strict about always writing "//" and we might simply keep on writing "/".

4.5 What analytic geometry is about

4.5.1 Paper and mind

There are geometric shapes on paper and those conceptualized by you in your mind. The following table distinguishes relevant aspects. We denoted the number line with $\mathbb{R}$. This also gives us the formal definition of the plane as the two dimensional product of two real axes $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$. We include Euclid's axiomatic development of (synthetic) geometry.

Drawing	*Mind*	*The Elements*	*Analytic formulas*
Visible dot	Point without size	Axioms	$\{x,\ y\}$
Visible line	Only length and no width	...	$p\,x + q\,y\ =\ r$ or $\mathbb{R}$
Piece of paper	Plane without thickness	...	$\mathbb{R} \times \mathbb{R}\ = \mathbb{R}^2$

The distinctions can also be denoted as $\mathbb{R}^0$, $\mathbb{R}^1$, $\mathbb{R}^2$ (, ...). The formula for the line $p\ x + q\ y = r$ in the table above can be recognized as a single $\mathbb{R}$ but with an orientation in the plane. Though we can find points on a line it is not quite accurate to say that the line "consists" of points: since there is a change in dimension and the line introduces a new concept in the discussion: continuity.

Euclidean geometry consists of the first three columns (with arithmetic) while analytic geometric concerns all columns. Euclidean geometry regards only the second column as the true result. Once the ladder of *The Elements* has been climbed it can be thrown away. Analytic geometry regards the first column as a useful stepping stone but no standard for proof. The formal developments in the last two columns differ from our imagination of space but have value of their own - since the creation of a formal system is a fair achievement of mankind as well.

Euclid: http://aleph0.clarku.edu/~djoyce/java/elements/elements.html

4.5.2 Looking back and ahead

With this understanding of what analytic geometry is about, it can be enlightening to look back at your learning process in this chapter. (1) The distinction between horizontal, vertical and diagonal lines is at a Level 0 of understanding. The material helps you to grow aware of what you actually already know. You start linking lines to co-ordinates and numbers. (2) The second step is description, sorting and classification of the lines that we have found, and this generates Level 1. Your understanding is helped by the fact that we already gave names to the lines at Level 0. (3) The third step consists of giving formulas for the separate lines. This is far removed from an abstract formal development so this is rather informal. You get a lot of help from the book since otherwise you would not know what the idea is. Writing the axes as $y = 0$ and $x = 0$ is not something that students conceive of naturally. This is Level 2. At this level you know how formulas look like, you proceed in trying to find a general formula, and you have a general notion of slope and constant. (4) The final level is Level 3 of formal deduction, with the crown in $p\ x + q\ y = r$. This formula is again presented to you since it is a very abstract insight that hardly anyone will conceive of by himself or herself. The key point is that you are able to understand its generality and how it relates to the specific cases seen before.

These levels of understanding use the same words but in different meanings. Also deduction and proof have different functions. These are sublanguages within a language, and people speaking these sublanguages will not understand each other (unless you are trained at a higher level to see whether someone is still at a lower level).

The introduction of $p\ x + q\ y = r$ at this stage has two advantages: (1) You are aware that there is a single formula, and you are not lost in the curious distinction between the function $y = s\ x + c$ and the vertical $x = r$ that does not satisfy that formula. (2) We have a foundation for the later discussion of systems of equations. There is a small disadvantage in that the general formula may not be used much at this stage. For practical calculations and plotting you are more likely to make separate use of $x = r$, $y = c$ and $y = s\,x + c$. This might also be how the mind works, since we try to imagine what a formula stands for. Accept this disadvantage and await the rewards that come later.

5. Circle

5.1 Distance and radius

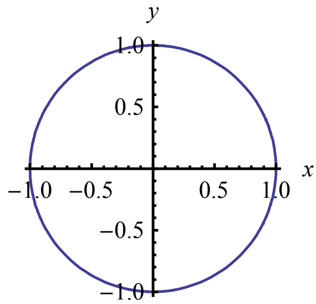
5.1.1 What defines the plane

Regard a point $v = \{x, y\}$ that lies at some distance from the origin {0, 0}. For example the point {3, 4}. The plane has been drawn on a blackboard and the teacher holds a coin at that point {3, 4}. Letting go, the coin drops down to {3, 0}. Then the teacher picks it up and puts it on {0, 0}, and then pockets it. He invites students to give him more coins so that he can show more examples of what a Manhattan distance is. Streets in Manhattan tend to form a perpendicular grid and you have to make corners to be able to go sideways. The distance travelled by the coin then is $x + y$ or in the example $3 + 4 = 7$. However, when we regard a plane without obstacles then we can go as the bird flies. We take this as the formula for the distance in the geometric plane:

$$|v| = |\{x, y\}| = \sqrt{x^2 + y^2}$$

This is called the Euclidean distance measure. For {3, 4} we get $\sqrt{9 + 16} = \sqrt{25} = 5$. This $|v|$ is also called the *absolute value* or *modulus* of v. It is the length of the line section from v to the origin. Another example: Let $P = \{5, y\}$ and suppose that we know that it lies at a distance of 10 from the origin. What is the value of y ? Well, $|P| = 10$ thus $25 + y^2 = 100$ and $y = \pm 5\sqrt{3}$. The abundance of the square root sign in mathematics derives from this definition of distance.

The Euclidean distance measure and the circle are related. A circle was defined as the collection of all points that are at the same (given) distance from its center. This given distance is called the radius of that circle. The distance measure thus seems like nothing new. However, the news is the link of the distance measure to the system of co-ordinates. We already divided the circle in quadrants and we have discussed the unit circle. But this is the first time that we draw a graph with a system of co-ordinates. This is really something else.



A first consequence of using co-ordinates is that we get another formula when we move the center to the point $\{a, b\}$.

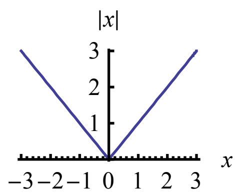
- If we want a circle around a point $\{a, b\}$ then this is similar to moving a point to the origin. In this new formula, the point $\{x, y\} = \{a, b\}$ gives $r = 0$ and thus $\{a, b\}$ must be the center.

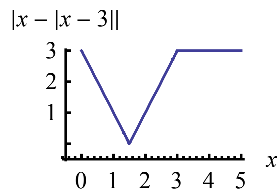
$$r = \sqrt{(x-a)^2 + (y-b)^2}$$

A second consequence is that the distance between points v and u can be found as $|v - u|$. Namely $|v - u| = |\{x, y\} - \{a, b\}| = |\{x - a, y - b\}|$ etcetera.

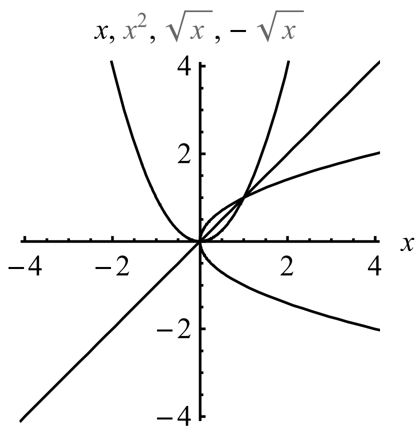
The following graphs explain more about the properties of the radius.

- Left $|x|$. On the right $|x - |x - 3||$ creates a square root sign.





- The mirror image of x^2 over the line $y = x$ gives the two root solutions. The intersections are at 0 and 1.



5.1.2 Euclid and Pythagoras

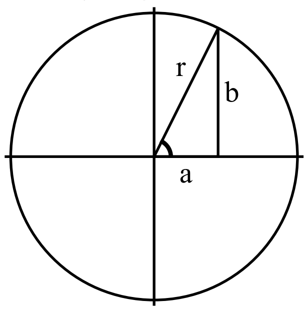
There is the paradoxical situation that we may take great pains to prove something that from another point of view is merely a matter of definition.

The Pythagorean Theorem is commonly expressed in terms of sides *a*, *b* and *c*. For the circle $c = r$. Then we get:

- Pythagoras convinces us that we have to *prove* that $c^2 = a^2 + b^2$
- For a distance we now *define* that $c^2 = a^2 + b^2$

The solution to this paradox is that Euclid used other axioms than we now do for the distance. Though Pythagoras (ca. 572 - 500 BC) lived before Euclid (around 300 BC), we can say in a figure of speech: Given the Euclidean axioms Pythagoras has to prove his Theorem. Once he got the proof he could define the circle. Without the proof he might define the circle but then would have to prove that it really exists. That said, in analytic geometry it is easier to work the other way around. Starting with formulas is a fast way to get up and running. Using distance we can define parallel lines as lines that have equal distance. With distance the circle arises naturally. The notion of distance is crucial for the Euclidean plane. We surmise that Euclid relied on a notion of distance too by using the compass.

- The Pythagorean Theorem holds by definition.



What remains in all this is our notion of Euclidean space: a notion of straightness of lines and flatness of the plane that might derive from everyday experience but that essentially is a concept of the mind, and essentially a definition.

`CircleDefinedByPythagoras[`*a*`, `*b*`, `*opts*`]`

the circle with center {0, 0} through {a, b}. The radius is given by the Pythagorean theorem r = Sqrt[$a^2 + b^2$].
For the angle α we have b = r Yur[α] and a = r Xur[α] ,
such that $\text{Yur}[\alpha]^2 + \text{Xur}[\alpha]^2 = (b/r)^2 + (a/r)^2 = 1$

5.1.3 Do we not have a definition for the slope ?

In above plot of the Pythagorean Theorem: take a hard look at the sides a and b of the triangle. Refresh what you know about the circle. With co-ordinates, the graph above gets a new meaning. As r makes an angle, it is also a line with an angle. Do we not have a definition for the slope ? Yes, it is $s = b / a$. We can write $x = a$ and $y = b$. Let us look closer what this means in a system of co-ordinates.

5.2 Functions *X* and *Y*

5.2.1 Points on the unit circle

The unit radius circle, thus with radius $r = 1$, will be called the unit circle. Points on the unit circle better have the suffix that they are on the unit radius circle, thus x_{ur} and y_{ur}. For convenience they will be labelled $\{X, Y\}$ as well, thus $X = x_{ur}$ and $Y = y_{ur}$. We use these capital labels throughout this book. When one co-ordinate is known then we know the other one. Note the two solutions $Y = \pm\sqrt{1 - X^2}$.

A point $v = \{x, y\}$ other than the origin always has a partner point on the unit circle. Two functions come into play. They are essentially one function because of the dependence, but with the double (±) solution it is best to keep track of them both.

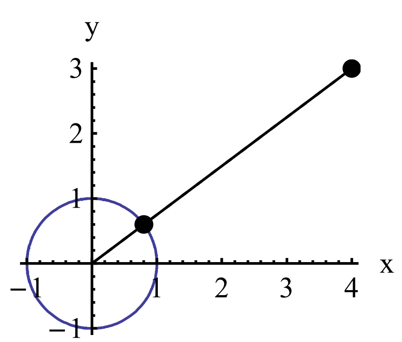
$$X_v = X[v] = x / |v| \qquad Y_v = Y[v] = y / |v|$$

The X and Y procedure is called "normalization". The meaning is a move to the unit circle along a ray to the origin.

- Normalizing from the point {4, 3} to the unit circle.

```
XandY[{4, 3}]
```

$\left\{\frac{4}{5}, \frac{3}{5}\right\}$



Consider the following example question. For point R on the unit circle we know $X = ½$. Point Q is at a distance of 10 from the origin and has the same direction as

R. What are the values of x and y ? Well, $|Q| = 10$ thus $x^2 + y^2 = 100$. We also know that $\{x, y\} = |Q| \{X, Y\} = 10 \{X, Y\}$. We find $Y^2 = 1 - X^2 = ¾$ so that $Y = \pm ½ \sqrt{3}$. Dus $\{x, y\} = \{5, \pm 5\sqrt{3}\}$. (Thus $Q = P$ from the earlier example.) (It was tricky to say "Point R" when there must be two.)

5.2.2 A measure for slope and direction

Let us summarize what we have for a point $\{x, y\}$:

1. We had the notions of the angle α and the arc φ for geometric shapes.
2. We had the notion of the Pythagorean Theorem $r^2 = a^2 + b^2$.
3. Now we strictly define X and Y for the unit circle so that there is no confusion with other $\{x, y\}$ on the plane. Because of the location of the unit circle at the center of the system of co-ordinates this gives a systematic treatment for all points in the plane.
4. In § 1.5.1 and 2 we defined various notions for the unit circle. These then hold only for X and Y, unless there is invariance due to proportionality.
5. The slope is proportional: $s = y / x = Y / X$. We might set up a system on the slopes and forget about α and φ. But if we want to know the circumference of the circle then this is equivalent to showing an interest in α and φ since those are arcs by their very nature. Coffee cups and soda cans tend to be somewhat roundish for a reason.
6. When we focus on angle α and arc φ then this leads us to the subject of trigonometry.

Before we proceed with trigonometry it is better to first consider vectors. For two reasons: (1) it is conceptually easier to work in the system of co-ordinates with X_v and Y_v, (2) we can there prove the key theorem of analytic geometry.

PM. We proved the Pythagorean Theorem when discussing triangles. Then we saw it again when discussing the circle. Then we used it in a system of co-ordinates to define distance. Now we are going to re-use it in a system with vectors. Don't think that our creativity in using the theorem in some disguise or other stops there.

6. Vector

Vectors aren't that difficult. Pierre van Hiele who was a celebrated researcher on the didactics of mathematics was a strong proponent that they are taught in elementary school. When we have a point $\{a, b\}$ and a point $\{x, y\}$ then the novel idea is that we add these two and get $\{a + x, b + y\}$. That is basically it. It is addition of more things at the same time. Let us count the numbers of pens and pencils that each kid has, but separately. That Van Hiele did not succeed in getting his proposal accepted has more to do with the training of elementary school teachers than with the difficulty of the subject. The discussion below will be a bit more difficult since we will not only do addition but also multiplication and we will also develop why it works and why it is mathematically sound. We start out a bit more complex than with points, namely with arrows.

6.1 Arrows have a direction

Walking on the bumpy grasslands of the flat plane, Alice met The Sulking Arrow. "Where are you heading ?" she asked it. "I am sorry, I don't know," it answered, indeed sulkingly. "I lost my head. If you really want to know, you must go find it, and ask it yourself." (Free after Lewis Caroll.)

6.1.1 Definition

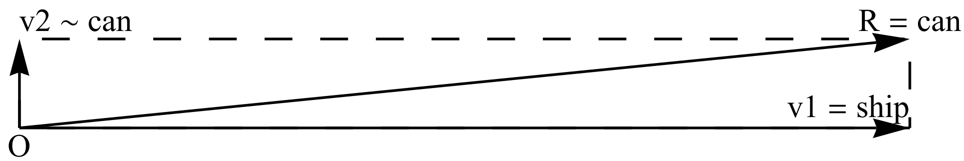
Consider a soda can on a deck of a ship. In 10 seconds it rolls 7 meters from port to starboard. In those 10 seconds the ship itself has sailed 67 meters. People on the ship may see only the movement of the can on the ship. A landbased observer sees a combined movement. The object of discussion now is how we could best handle this kind of case.

Let us consider two points $P = \{x, y\}$ and $Q = \{a, b\}$. We can draw an arrow that starts from P and the arrow head ending in Q. We shall call that arrow a *vector* and write $v = \{P, Q\}$.

For example the vector from {0, 1} to {3, 4} is $v = \{\{0, 1\}, \{3, 4\}\}$.

The ship moves along the horizontal axis from {0, 0} to {67, 0}, and this will be vector v1. If the ship would be at rest then the soda can moves along the vertical axis across the deck from {0, 0} to {0, 7}, and this will be vector v2. The resultant movement is R. After 10 seconds the can is at position {67, 7}.

- The ship moves a distance of v1, the soda can a distance of v2 on the ship, and the soda can has a resulting movement of R seen by a landbased observer O. In the same time, those 10 seconds, the soda can moves over a greater distance, and thus it must move faster than the ship.



While the earlier discussion used points, we now have arrows, as combinations of points. The news is that we now have a model for motion. Co-ordinates are static, vectors are dynamic. What are the properties of such arrows ?

`TwoVectorsPlot[v, w]`	shows the vectors and their resultant determined by adding the coordinates

Option Label controls the three labels. Option Select chooses from First, Last or All: for display with numbers and decomposition or not.

6.1.2 Properties of vectors in general

Vectors clearly have these properties:

- Length: The length of vector $v = \{Q, P\}$ is $|v| = |P - Q|$. This is just the distance between the begin and end points.
- Direction: We have a slope and $\{X, Y\}$ values. The latter are given by $X[P - Q]$ and $Y[P - Q]$ and $s = Y[P - Q] / X[P - Q]$.

Above we already performed an addition of vectors. In our example the vectors were perpendicular but we can define in general:

- Addition: When vector $v = \{P, Q\}$ and $w = \{Q, S\}$ so that the end of the first is the start of the other, then the resultant is $r = \{P, S\}$.

We now drop the notion of vectors in general and switch to the vectors that are special since they all start at the origin - like actually in above example.

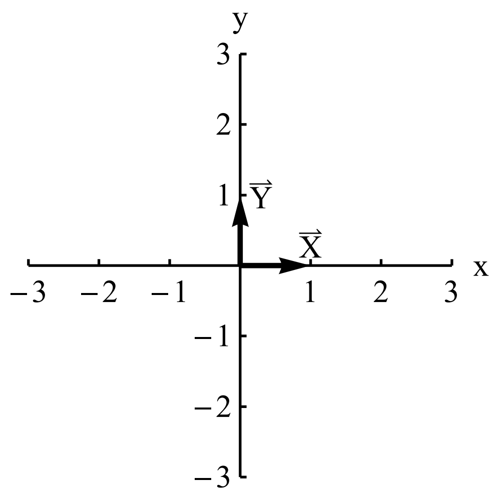
6.1.3 Vectors that originate from the origin

Apart from vectors in general there are the vectors that are special since they all start at the origin. There are four key vectors that can be reduced to two independent ones - and we use capitals because of the unit circle:

- $\vec{X}$ = {{0, 0}, {1, 0}} the unit vector to the right (forwards)
- $\vec{Y}$ = {{0, 0}, {0, 1}} the unit vector perpendicular up
- -$\vec{X}$ = {{0, 0}, {-1, 0}} the unit vector to the left (backwards, in reverse)
- - $\vec{Y}$ ={{0, 0}, {0, -1}} the unit vector perpendicular down

We define addition such that v = {{0, 0}, {x, y}} can be written as $v = x\,\vec{X} + y\,\vec{Y}$.

This is the unit vector plot. E.g. locate {3, 2} = $3\,\vec{X} + 2\,\vec{Y}$.



`UnitVectorPlot[]`	shows the two unit vectors
`UnitVectorRule[]`	for replacing the names with the actual unit vectors

It appears to be awkward to work with general vectors that have different beginnings. It is easier to work with vectors from the origin. This also allows a simplification. We only record the endpoint, e.g. {{3, 4}}. Even, we can drop the additional brackets, as long as we clearly state *vector* v = {3, 4} and *point* p = {3, 4}.

Though vectors from the origin do not link up (i.e. that the end of one is not the start of another) we can still define addition. The key example is the case of the ship and the soda can. In continuous time the origin keeps shifting so that the vectors are linked at the origin indeed. For the calculation it suffices to consider the beginning and end of the process.

- When we have vector $v = x\,\vec{X} + y\,\vec{Y}$ and $w = a\,\vec{X} + b\,\vec{Y}$ then we can define addition as follows. Subtraction is adding the negative vector.

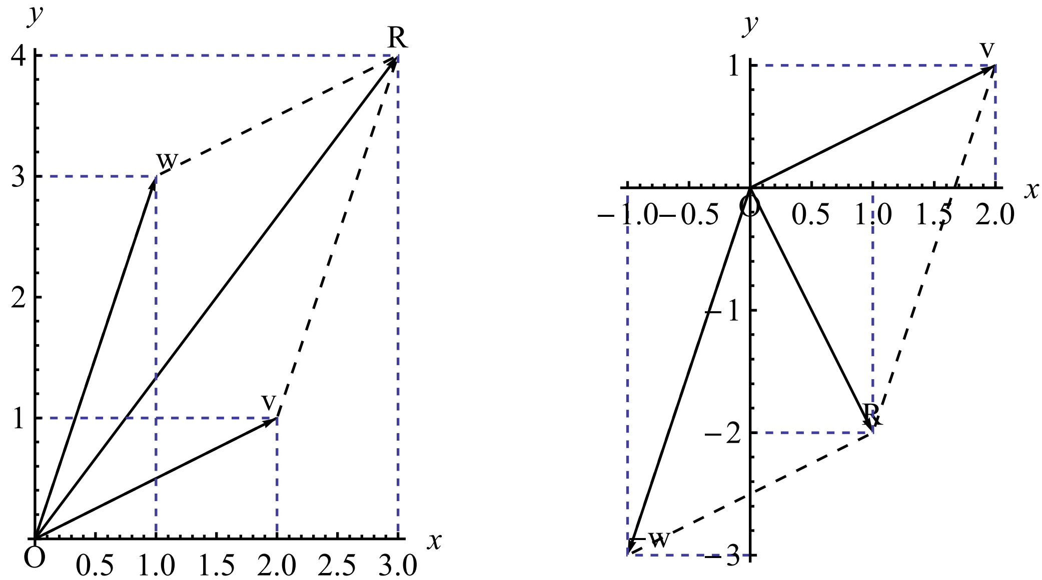
```
v + w == (x X⃗ + y Y⃗) + (a X⃗ + b Y⃗);
```

$$v + w = (a + x)\,\vec{X} + (b + y)\,\vec{Y}$$

- This substitutes the unit vectors {1, 0} and {0, 1}.
  $v + w = \{a + x, b + y\}$

It appears immaterial whether we add vectors starting at the origin or have them link at end and start. Both views give the parallellogram.

- On the left, vectors v = {2, 1} and w = {1, 3} give R = {3, 4}. On the right, subtracting v = {2, 1} and w = {1, 3} is adding the negative, $R = v - w$ = {1, -2}.



See http://www.slu.edu./classes/maymk/SketchpadApplets/AddVectors.html.

6.1.4 A key formula

Using the earlier normalization for the co-ordinates $X_v = X[v] = x \,/\, |v|$ and $Y_v = Y[v] = y \,/\, |v|$ for points on the unit circle, we find:

$$v = \{x, y\} = |v|\ (X_v\, \overrightarrow{X} + Y_v\, \overrightarrow{Y}\ )$$

This formula clarifies that X_v and Y_v are *values* of co-ordinates but not quite the co-ordinates themselves. The co-ordinates arise from the application to the unit vectors.

Defining $\overrightarrow{X_v} = X_v\, \overrightarrow{X}$ and also for Y gives a somewhat shorter form :

$$v = \{x, y\} = |v|\ (\overrightarrow{X_v} + \overrightarrow{Y_v}\ )$$

We can use either form depending upon what is handy.

6.1.5 Multiplication gives counterclockwise rotation

Above we have defined addition. Can we also find something like multiplication ?

Looking at a wheel that is turning we can count how often it turns around. The unit of measurement then is a single turn. Values are e.g. a half turn or a quarter turn. A positive turn is counterclockwise, a negative turn is clockwise.

The quarter turn is special since a point that originally was horizontally aligned with $\vec{X}$ now becomes vertically aligned with $\vec{Y}$. We use this property to define multiplication of vectors.

Definition of multiplication:

$\vec{Y}\,\vec{X} = \vec{X}\,\vec{Y} = \vec{Y}$ or $\{0, 1\}\,\{1, 0\} = \{1, 0\}\,\{0, 1\} = \{0, 1\}$ or a quarter turn upwards

$\vec{Y}\,\vec{Y} = -\vec{X}$ or $\{0, 1\}\,\{0, 1\} = \{-1, 0\}$ or a quarter turn down to the left again

$\vec{X}\,\vec{X} = \vec{X}$ or no change.

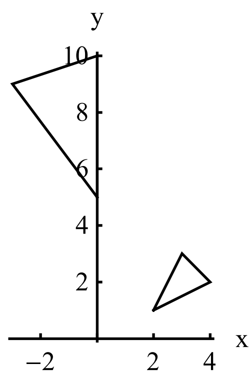
When we turn an arbitrary point $\{x, y\} = x\,\vec{X} + y\,\vec{Y}$ with $\vec{Y}$ then we get (using the two formats):

$\vec{Y}\left(x\,\vec{X} + y\,\vec{Y}\right) =$	$\{0, 1\}\,(x\,\{1, 0\} + y\,\{0, 1\}) =$
$x\,\vec{Y}\,\vec{X} + y\,\vec{Y}\,\vec{Y} =$	$x\,\{0, 1\}\,\{1, 0\} + y\,\{0, 1\}\,\{0, 1\} =$
$x\,\vec{Y} - y\,\vec{X} =$	$x\,\{0, 1\} + y\,\{-1, 0\} =$
$\{-y, x\}$	$\{-y, x\}$

When we multiply two arbitrary points, or turn $\{x, y\}$ with $\{a, b\}$:

$\left(a\,\vec{X} + b\,\vec{Y}\right)\left(x\,\vec{X} + y\,\vec{Y}\right) =$	$\{a, b\}\,\{x, y\} =$
$a\,x\,\vec{X}\,\vec{X} + a\,y\,\vec{X}\,\vec{Y} + b\,x\,\vec{Y}\,\vec{X} + b\,y\,\vec{Y}\,\vec{Y} =$	$(a\,\{1, 0\} + b\,\{0, 1\})\,(x\,\{1, 0\} + y\,\{0, 1\}) =$
$a\,x\,\vec{X} + a\,y\,\vec{Y} + b\,x\,\vec{Y} - b\,y\,\vec{X} =$	$a\,x\,\{1, 0\}\,\{1, 0\} + \ldots + b\,y\,\{0, 1\}\,\{0, 1\} =$
$(a\,x - b\,y)\,\vec{X} + (a\,y + b\,x)\,\vec{Y} =$	$(a\,x - b\,y)\,\{1, 0\} + (a\,y + b\,x)\,\{0, 1\} =$
$\{a\,x - b\,y, a\,y + b\,x\}$	$\{a\,x - b\,y, a\,y + b\,x\}$

Regard a triangle, for example, with corners $P = \{2, 1\}$, $Q = \{4, 2\}$ and $R = \{3, 3\}$. When we turn this with $\{1, 2\}$ then we get $P' = \{0, 5\}$, $Q' = \{0, 10\}$, $R' = \{-3, 9\}$. We can assume that the sides of the triangle move with the corners. The general shape remains the same but there are both *rotation* and *enlargement*.



PM. This example has been crafted for simplicity in calculating and graphing: two points are on a ray through the origin and the multiplication gives an outcome on the vertical axis, thus with $x = 0$.

`VectorProduct[` $\{a, b\}, \{x, y\}$`]`	gives $\{a\,x - b\,y, a\,y + b\,x\}$
`VectorProductGO[` *object*, *p*`]`	the geometric transformation of *object* with point *p*
`VectorProductPlot[` *object*, *p*`]`	plots the latter
`PointToTFMatrix[`*p*`]`	gives *q* for GeometricTransformation[*object*, *q*]

6.1.6 Properties of multiplication and rotation

A property is that the lengths are simply multiplied: $|v\,w| = |v|\;|w|$.

Thus, where above triangle was enlarged, it was because we did not just rotate but also multiplied the vector lengths. Let us prove and understand this property first.

6.1.6.1 Proof that the distance of a product is the product of the distances

With $v = \{x, y\}$ and $w = \{a, b\}$ we get $v\,w = w\,v = \{a\,x - b\,y,\;\; a\,y + b\,x\}$ from above. Application of Pythagoras gives:

$$|v\,w|^2 = (a\,x - b\,y)^2 + (a\,y + b\,x)^2$$

$$= (a^2\,x^2 - 2\,abxy + b^2\,y^2) + (a^2\,y^2 + 2\,abxy + b^2\,x^2)$$

$$= a^2\,x^2 + b^2\,y^2 + a^2\,y^2 + b^2\,x^2$$

$$= (a^2 + b^2)\,x^2 + (a^2 + b^2)\,y^2$$

$$= (a^2 + b^2)\,(x^2 + y^2) = \,|v|^2\,\,|w|^2$$

6.1.6.2 Consequence for the unit circle

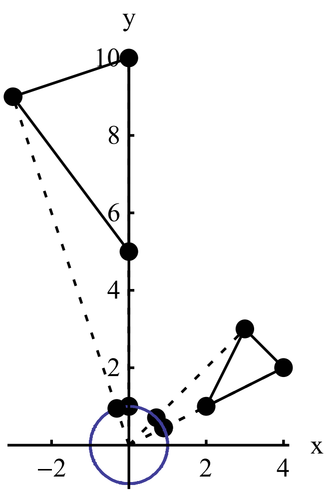
By consequence: When points on the unit circle are multiplied with each other then the resultant remains on the unit circle.

If $P = \{X, Y\}$ and $Q = \{A, B\}$ are on the unit circle then $|P| = |Q| = 1$. Then above property gives $|P\,Q\,| = 1$.

Alternatively stated: When we divide $v\,w = w\,v = \{a\,x - b\,y,\;\; a\,y + b\,x\}$ by $|v\,w| = |v|\,\,|w|$ then we get points on the unit circle. Multiplication of those gives an outcome that is again on the unit circle. Namely:

$$\frac{v\,w}{|v\,w|} = \frac{w\,v}{|v\,w|} = \left\{\frac{a}{|w|}\,\frac{x}{|v|} - \frac{b}{|w|}\,\frac{y}{|v|},\;\; \frac{a}{|w|}\,\frac{y}{|v|} + \frac{b}{|w|}\,\frac{x}{|v|}\right\}$$

This is a key discovery. Something does not change: we stay on the circle. This means that multiplication of vectors consists of (1) multiplying the lengths, and (2) rotating on the unit circle. The following graph displays the two aspects.



An open question remains what exactly happens on that circle. We have derived the new co-ordinates, but is there a handy interpretation ?

6.1.6.3 The form that relates to functions *X* and *Y*

This important result can be shown in another way, using the normed values. Expressing vectors in both length and co-ordinates on the unit circle then we get:

$$v = \{x, y\} \;= |v|\;(\overrightarrow{X_v} + \;\overrightarrow{Y_v})$$

$$w = \{a, b\} = |w|\;(\overrightarrow{X_w} + \;\overrightarrow{Y_w})$$

$$v\,w = w\,v = |v|\;(X_v\,\vec{X} + Y_v\,\vec{Y})\quad |w|\;(X_w\,\vec{X} + Y_w\,\vec{Y})$$

$$= |v|\;|w|\;(\,(X_v\,X_w - Y_v\,Y_w)\,\vec{X}\;+\;(Y_v\,X_w\;+X_v\,Y_w)\,\vec{Y}\,)$$

thus $\quad X_{v\,w} = X_v\,X_w - Y_v\,Y_w$

$$Y_{v\,w} = Y_v\,X_w + X_v\,Y_w$$

This is a form that more clearly links up with the co-ordinate functions.

When we multiply more than two vectors then these rules result in rather complicated expressions. Below we will find a way for simplification.

2D demonstrations by Roger Germundsson:
http://demonstrations.wolfram.com/author.html?author=Roger+Germundsson.

6.2 The key theorem of analytic geometry

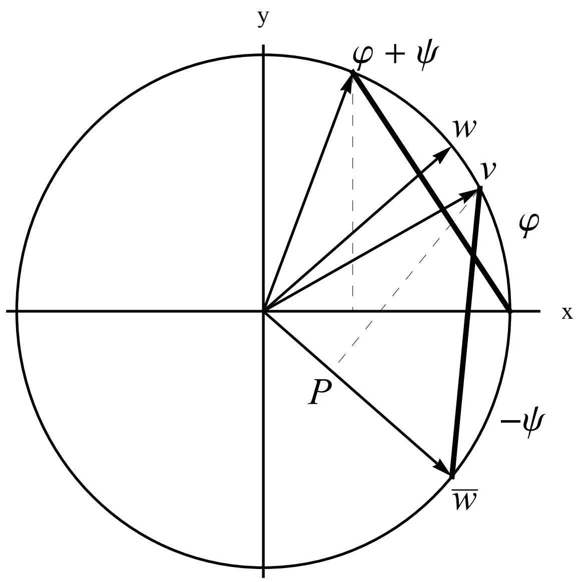
6.2.1 Vector multiplication means addition of angles

The key theorem of analytic geometry is: vector multiplication on the unit circle is the addition of angles.

We can derive the theorem directly using angles, thus using the unit circumference circle with $r = 1 / \Theta$ and using angles or turns indeed. The formulas however are more transparent on the unit circle with $r = 1$. The theorem can also be formulated so: vector multiplication means the addition of arcs. Later, when we calculate surfaces and the changes in surfaces then it appears that these arcs are important too. There is good reason to get used to them. To prove the theorem we use the following drawing on the unit circle. See the explanation below.

Call the origin O, note that all vectors have length 1 and that all arcs are measured from {1, 0}. We can identify:

1. $v = \{X, Y\}$ is the first vector. It creates an arc φ.
2. $w = \{A, B\}$ is the second vector. It creates an arc ψ.
3. $\overline{w} = \{A, -B\}$ is called the conjugate. It creates an arc $-\psi$.
4. The key triangle is $Ov\overline{w}$. The arc has absolute length $\varphi + \psi$. Projection of v on $\overline{w}$ gives point P. We write $p = |P|$.
5. Rotating this key triangle to a horizontal position we get the point (labelled $\varphi + \psi$) where the arcs are properly added (not just the absolute value). The projection P is rotated too, and appears to be at the x-co-ordinate of that point.



6. Hence, when we calculate p then we have determined what it means when arcs are added. We already had the result of vector multiplication above. If and only if these two results are the same then we have proven the theorem stated at the beginning of this section.

7. Denote $h = |v - P|$ for the dashed height of the key triangle, and $q = |v - \overline{w}|$ for the thick chord.

8. Pythagoras gives $|v|^2 = h^2 + p^2 = 1$ and $q^2 = h^2 + (1 - p)^2$.

9. Thus $q^2 = h^2 + (1 - p)^2 = 1 - p^2 + (1 - p)^2 = 1 - p^2 + (1 - 2p + p^2) = 2\,(1 - p)$

10. We also know the length of q from applying Pythagoras on v and $\overline{w}$. Thus $q^2 = (X - A)^2 + (Y - (-B))^2 = (X^2 - 2\,A\,X + A^2) + (Y^2 + 2\,B\,Y + B^2) = 2 - 2\,A\,X + 2\,B\,Y$ since both $X^2 + Y^2 = 1$ and $A^2 + B^2 = 1$.

11. Combining the two results allows us to eliminate q^2 and we get $2\,(1 - p) = 2 - 2\,A\,X + 2\,B\,Y$ which solves into $p = A\,X - B\,Y$.

12. Above we had found $X_{v\,w} = X_v\,X_w - Y_v\,Y_w$. This is exactly the same. Thus $X_{v\,w} = p$ and multiplication of vectors on the unit circle gives the same co-ordinates as the addition of the separate arcs. Q.E.D.

PM. Check what happens with an obtuse angle.

`AGKeyTheoremPlot[φ, ψ]`	the diagram of the key theorem of analytic geometry

6.2.2 The Arc function and what we have derived for it

For a point $\{X, Y\}$ on the unit circle it is natural to define the length of the arc (starting from $\vec{X}$) as a function of the co-ordinates.

$$\varphi = \text{Arc}[\{X, Y\}] = \text{Arc}[\{x, y\}] = \text{Arc}[v]$$

We have derived:

$$\text{Arc}[v\ w] = \text{Arc}[v] + \text{Arc}[w] = \varphi + \psi$$

When we divide these arcs by $\Theta = 2\pi$ then we get the turns around the circle, using $\alpha = \varphi / \Theta$ and $\beta = \psi / \Theta$.

$$\text{Turn}[v\ w] = \text{Turn}[v] + \text{Turn}[w] = \alpha + \beta$$

This is the main contribution of vector analysis. Since the focus now shifts to angles it is time for consequences like trigonometry.

6.3 When is it a vector space ?

When does a vector space apply ? The only rule is: when it works.

With one basket with {3, 3} apples and oranges and another with {2, 4} then in total there are {5, 7} apples and oranges. A vector space.

In proportion space though with the ratios {2, 1} and {3, 1}, and thus the fractions 1/2 and 1/3, then the addition {5, 2} is interesting and represents 2/5 but is no adequate outcome if your plan is to arrive at {6, 5}.

A good other example are recovered exponents (logarithms) since those can be added.

Part III. Consequences

We have seen the basic ingredients of analytic geometry: line, circle and vector, found and proven the key theorem, and now proceed to what essentially are consequences. Of course there will be new insights but in another respect these are variations on the themes.

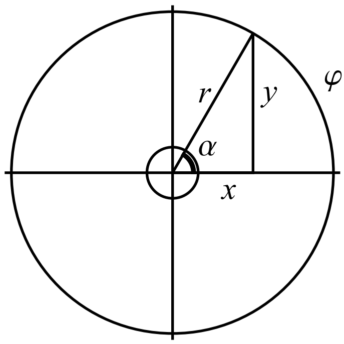
1. We develop the properties of $X_v = x_{\mathrm{ur}}$ and $Y_v = y_{\mathrm{ur}}$.
2. The complex plane is a different way to write the vectors.
3. The calculation of Θ is historically exiting (though nowadays routine).
4. Linear algebra gives no news for 2D but opens the road to higher dimensions.

Though these are variations on the themes it still appears that simply writing things a bit different may cause a new ray of light. Like speaking another language.

7. $\mathbb{X}$ and $\mathbb{Y}$

7.1 Introduction

The graph for the Pythagorean Theorem will provide for clarity of discussion.



1. Given the right angle we have $x^2 + y^2 = r^2$.
2. $s[\alpha] = y / x$ is the slope.
3. $\varphi = \alpha\,\Theta$
4. When α and r change then φ, x and y change.
5. When x and y change then φ, α and r change.

For point 5 we have defined the Arc function:

$$\varphi = \text{Arc}[\{X, Y\}] = \text{Arc}[\{x, y\}] = \text{Arc}[v]$$

And we have derived the key theorem of analytic geometry:

$$\text{Arc}[v\ w] = \text{Arc}[v] + \text{Arc}[w] = \varphi + \psi$$

$$\text{Turn}[v\ w] = \text{Turn}[v] + \text{Turn}[w] = \alpha + \beta$$

Due to the dependency: if we invert 5 then we must find 4.

We will work with φ and Arc since the property of $r = 1$ is handy. The issue is two-dimensional in Arc but in the following we are going to use one-dimensional functions ArcX and ArcY. We first proceed quickly and then later correct for shortcuts made along the way.

7.2 *X* and *Y* as functions of the arc

Let us look at the first quadrant. Then for each X there is an arc φ, and for each arc φ there is a X. The same for Y, that can be found by Pythagoras. That means that there are functions ArcX and ArcY and that these have inverses. Since v is twodimensional $\{x, y\}$ and φ is a single figure, there is no confusion writing $X_v = X_\varphi$.

$$\varphi = \text{ArcX}[\ X_v = x / r\ ] \quad \text{so that} \quad X_v = x / r = \text{ArcX}^{-1}[\varphi\ ] = X_\varphi$$

$$\varphi = \text{ArcY}[\ Y_v = y / r\ ] \quad \text{so that} \quad Y_v = y / r = \text{ArcY}^{-1}[\varphi\ ] = Y_\varphi$$

This looks like an abundance of names for $\{X, Y\}$ but they help to express shifts in focus. X_v expresses the dependence upon the two-dimensional vector, X_φ the dependence upon the arc. Then the key theorem of analytic geometry can be restated in the following form, and in a frame to express its importance:

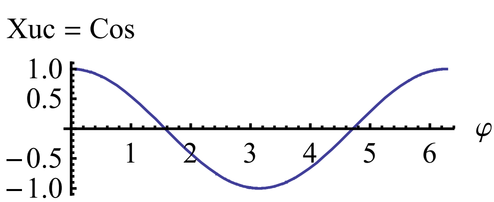
$$\boxed{X_{v\,w} = X_v\,X_w - Y_v\,Y_w = X_\varphi\,X_\psi - Y_\varphi\,Y_\psi = X_{\varphi+\psi}}$$

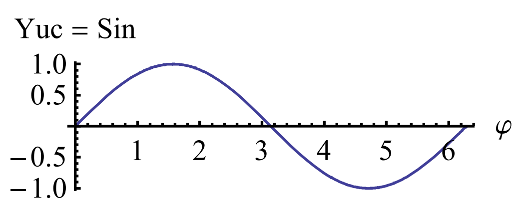
Historically, the function X_φ that expresses the dependence of the x-co-ordinate upon the arc is denoted as Cos[φ], so that $x / r = X_\varphi$ = Cos[φ]. And Y_φ is denoted as Sin[φ]. Cos and Sin are called cosine and sine from Latin *sinus* for angle, corner, vertex. These historical names are somewhat unattractive since they do not directly refer to the co-ordinates that we are dealing with. See if you can get used to them:

$$X_{v\,w} = X_v\,X_w - Y_v\,Y_w = \text{Cos}[\varphi]\ \text{Cos}[\psi] - \text{Sin}[\varphi]\ \text{Sin}[\psi] = \text{Cos}[\varphi\ +\ \psi]$$

When we extend this relation from the first quadrant to the whole plane then we get the following graph. For plotting it is useful to have proper axes to that we can see that the Y_φ or sine wave starts at 0 with a slope of 1 or an angle of 1 / 8 or 45°.

- Cos and Sin are the X and Y values on the unit circle given by the arc there.





See http://demonstrations.wolfram.com/IllustratingTrigonometricCurvesWithTheUnitCircle/

7.3 *X* and *Y* as functions of the angle

Also for the angles it appears useful to have different symbols to express the shifts in focus. The co-ordinate is first related to the vector but can also be seen as a function of the angle. Since the suffix of X has been exhausted we take another kind of X symbol. We can write $X_v = x_{\text{ur}}[\alpha] = \mathbb{X}[\alpha]$ called Xur and $Y_v = y_{\text{ur}}[\alpha] = \mathbb{Y}[\alpha]$ called Yur, where α is on the angular circle or the Unit Circumference Circle, and x_{ur} and y_{ur} are on the Unit (Radius) Circle. Straightforwardly $\alpha = \varphi / \Theta$ so that:

$$X_v = x / r = \mathbb{X}[\alpha] = \mathbb{X}_\alpha \qquad\qquad Y_v = y / r = \mathbb{Y}[\alpha] = \mathbb{Y}_\alpha$$

The key theorem can now be formulated neatly in its proper form for angles:

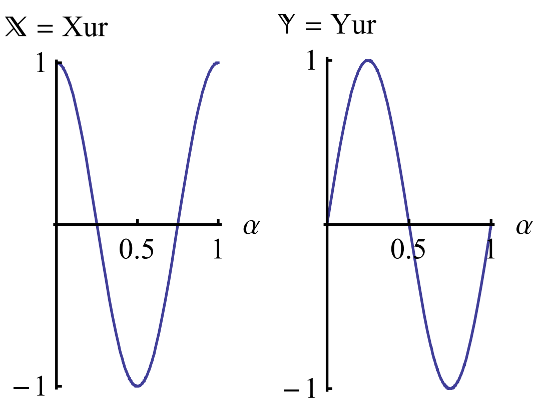
$$\boxed{X_{v\,w} = X_v\, X_w - Y_v\, Y_w = \mathbb{X}_\alpha\, \mathbb{X}_\beta - \mathbb{Y}_\alpha\, \mathbb{Y}_\beta = \mathbb{X}_{\alpha+\beta}}$$

This expresses that we are dealing with (a) co-ordinates and (b) dependence upon angles expressed in turns, (c) multiplication of vectors is addition of angles.

Plotting gives the same wave form as above but on the 0 to 1 domain for angles.

- These are the co-ordinates on the unit circle as functions of the angle.

  $\mathbb{X}_\alpha = x_{\text{ur}}[\alpha] = \text{Cos}[\alpha\, \Theta]$ and $\mathbb{Y}_\alpha = y_{\text{ur}}[\alpha] = \text{Sin}[\alpha\, \Theta]$ for angle α.



7.4 The algorithm

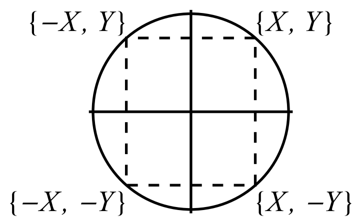
The algorithm thus is: when we multiply two vectors with known co-ordinates normalized on the unit circle then we can find the result first of all from $X_{v\,w} = X_v\, X_w - Y_v\, Y_w$ while $Y_{v\,w}$ can be found from Pythagoras or from $Y_{v\,w} = Y_v\, X_w + X_v\, Y_w$.

But we are free to express everything also in terms of the addition of angles and translate these back into co-ordinates. This can be handy for more multiplications like for cyclic functions as waves or regular polygons.

7.5 There are four quadrants

7.5.1 Pythagoras again

There is a tricky issue. We took a shortcut that needs to be resolved. Due to Pythagoras and his squares we get plusses and minusses. One X gives two possible Y's and one Y gives two possible X's. We can find unique values when we keep track of the positive or negative signs of the variables, i.e. in what quadrant they occur. Then we can translate the two-dimensional Arc[v] function to the one-dimensional ArcX and ArcY functions that depend upon the X and Y values.



There thus are the following dependencies:

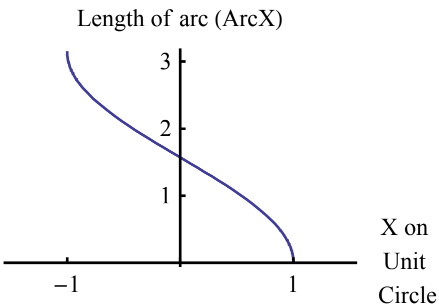
Dependency	$\mathbb{X}$	$\mathbb{Y}$
Mirror over horizon	$\mathbb{X}[-\alpha] = \mathbb{X}[\alpha]$	$\mathbb{Y}[-\alpha] = -\mathbb{Y}[\alpha]$
Mirror over vertical	$\mathbb{X}[1/2 - \alpha] = -\mathbb{X}[\alpha]$	$\mathbb{Y}[1/2 - \alpha] = \mathbb{Y}[\alpha]$
Jump over diameter	$\mathbb{X}[\alpha + 1/2] = -\mathbb{X}[\alpha]$	$\mathbb{Y}[\alpha + 1/2] = -\mathbb{Y}[\alpha]$
Positive quarter turn	$\mathbb{X}[\alpha + 1/4] = -\mathbb{Y}[\alpha]$	$\mathbb{Y}[\alpha + 1/4] = \mathbb{X}[\alpha]$
Negative quarter turn	$\mathbb{X}[\alpha - 1/4] = \mathbb{Y}[\alpha]$	$\mathbb{Y}[\alpha - 1/4] = -\mathbb{X}[\alpha]$

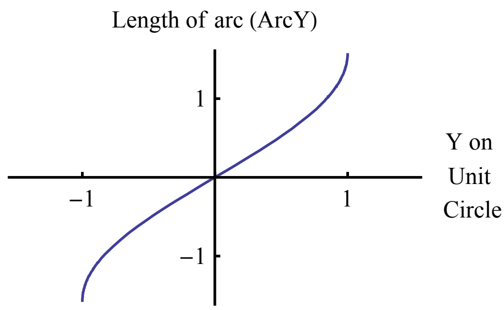
When the arcs are expressed in terms of only one co-ordinate, we get:

Arc[v] = Arc[{x, y}] = ArcX[$X_{\{x,\ y\}}$, given Sign[y]] = ArcY[$Y_{\{x,\ y\}}$, given Sign[x]]

We can apply these functions in two ways. The first way is to know in what quadrant we are and then apply the proper subfunction. The other way is to leave the quadrant somewhat unspecified and then work in a default manner, with proper inverses in a limited range.

- When we don't know the signs we can determine a unique value over a limited range. For X we assume that $Y \geq 0$ with range {0, 1/2} Θ. For Y we assume $X \geq 0$. To get a continuous funtion for Y we assume that a negative value means a negative arc. The range then is {-1/4, 1/4} Θ. The outcomes can differ from when we have full information.





`ArcLength[{x, y}]`	gives the length of arc on the unit circle when the point is projected there along a ray through the origin. The arc length is measured counterclockwise from {1, 0} and the range is 0 to Θ
`ArcX[x (, n)]`	assumes x = X[{x, y}]. If y < 0 then n = Negative can be used.
`ArcY[y (, n)]`	assumes y = Y[{x, y}]. If x ≥ 0 then n = NonNegative, otherwise Negative.
`ArcLengthPlot[(s)]`	plots. If label s = XandY then the variants with NonNegative and Negative are used

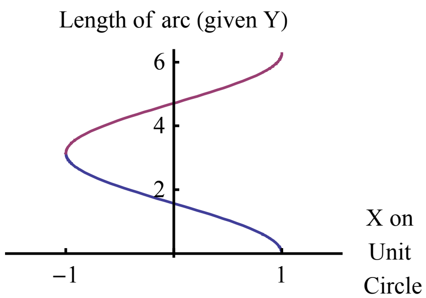
Arc is implemented with the name ArcLength to emphasize what it does. With ArcX and ArcY this is fairly obvious.

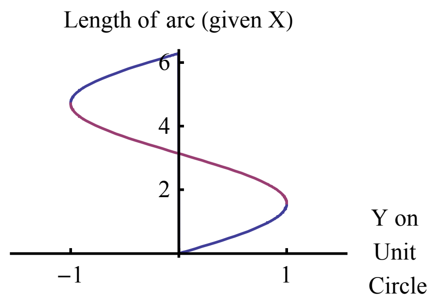
Since $Y = \pm\sqrt{1 - X^2}$ it matters in what quadrant the values are.

7.5.2 Unifying the plane

Unifying the plane again, we get the following graph. For example, when $X = 1/2$ then there are potentially two values for the arc. When we look at those two values in the plot for Y then we find two values for Y that generate the same arc, one positive Y and one negative Y. Iff we know the sign of Y then we know the proper arc length.

- The points $\{X, Y\}$ are on the unit circle so that X and Y are between -1 and 1. We let X move from 1 to -1 and then $Y = \pm\sqrt{1 - X^2}$ gives two values. If we know the signs of X and Y then all fits.





These are not functions but correspondences. They are transforms of the Cos and Sin functions, with the X and Y axes interchanged, turned but also mirrored.

7.6 The polar and angle planes

A point in space $v = \{x, y\} = r\,\{\mathbb{X}_\alpha, \mathbb{Y}_\alpha\} = r\,\{X_\varphi, Y_\varphi\} = r\,\{\text{Cos}[\varphi], \text{Sin}[\varphi]\}$ thus has different representations.

Alongside the two-dimensional plane for $\{x, y\}$ there is a plane for $\{r, \varphi\} = \{|v|, \text{Arc}[v]\}$ that uses both the Euclidean length of the vector and the length of the arc. These are called the *polar plane* and *polar co-ordinates*. Another plane uses length and angle $\{|v|, \alpha, 1\}$, called the angle or UMA plane with likewise co-ordinates (unit measure around), where it is useful to include the parameter for the unit range so that it is not confused with the polar plane with parameter Θ that is not stated by default.

7.7 Taking stock

It is useful to look back at our steps and the results that came about. Remember why we do all of this.

For vectors it was easy to define addition. Next we introduced multiplication and looked at the implications. We want to understand it and if possible we want to find an easier expression.

We noted that the unit circle with its easy radius plays a key role. We distinguished the length of a vector and a normalization to the unit circle. When multiplying those normalized vectors, the normalized result remained on the unit circle. The idea arose that multiplication of vectors meant the addition of their arcs. We proved that this is indeed the case: the key theorem of analytic geometry.

The relationship between co-ordinates and arc was expressed in φ = ArcX[x / r] = ArcY[y / r] functions (for the first quadrant), that take a co-ordinate and generate an arc. Their inverses take an arc and generate a co-ordinate, i.e. the functions traditionally called Cos and Sin.

In review and summary: We have succeeded in finding a transparant interpretation. With the key result of analytic geometry we now have found the following relationships between co-ordinates, turns, angles and arcs (where an angle is based upon arcs too) and slopes as well. There is an abundance of notations but these allow us to express the shift in focus:

1. The length of $z = v\ w$ is given by $|z| = |v\ w| = |v|\ |w|$.

2. The arc of z is given by Arc[z] = Arc[$v\ w$] = Arc[v] + Arc[w] = $\varphi + \psi$.

3. The horizontal co-ordinate is $X[z] = X_v\ X_w - Y_v\ Y_w$.

4. Moving from two dimensions to one dimension we get:

4a. φ = ArcX[X_v] = ArcCos[X_v] and thus inversely $X_v = X_\varphi$ = Cos[φ].

4b. φ = ArcY[Y_v] = ArcSin[Y_v] and thus inversely $Y_v = Y_\varphi$ = Sin[φ].

5. Thus we have Cos[$\varphi + \psi$] = Cos[φ] Cos[ψ] - Sin[φ] Sin[ψ].

6. $\mathbb{X} = x_{\text{ur}}$ and $\mathbb{Y} = y_{\text{ur}}$ use angle $\alpha = \varphi / \Theta$ with $\Theta = 2\ \pi$.

7. Then $X_v = x / |v| = x_{\text{ur}}[\alpha] = \mathbb{X}_\alpha$ = Cos[φ] and $Y_v = y / |v| = y_{\text{ur}}[\alpha] = \mathbb{Y}_\alpha$ = Sin[φ].

8. Thus we have $X_{v\,w} = X_v\ X_w - Y_v\ Y_w = \mathbb{X}_\alpha\ \mathbb{X}_\beta - \mathbb{Y}_\alpha\ \mathbb{Y}_\beta = \mathbb{X}_{\alpha+\beta}$.

9. The slope is given by $s_v = y / x = Y_v / X_v = \mathbb{T}[\alpha]$ = Tan[φ] on the line $x = 1$.

A bit confusing in this development are the historical names of Cos and Sin that do not express their relation to the unit circle and the specific co-ordinates. Due to the path that we followed it ought to be established firmly in your mind though that those Cos and Sin functions are only the co-ordinates on the unit circle.

PM. It is an option to rebaptise Cos and Sin into Xuc and Yuc as all their action is on the Unit Circle, but this likely causes confusion with Xur and Yur, and use of Cos and Sin gives better access to the traditional literature and webpages.

PM. Traditional books and websites write Cos[φ] = x / r and suggest that this is a definition of Cos. However, it is an equation to solve. A definition like $f[x] = x^2$ has x on the right hand side, and there is no φ on the right hand side of Cos[φ] = x / r. At best it is an inverse definition, but confusingly this is hardly ever explained. It will be the way how in the historical past numerical tables were generated but it is inverted from how we use the functions. The proper expression is $x = r$ Cos[φ] where the co-ordinate is calculated from the arc and the radius. Also φ = ArcX[x / r] = ArcCos[x / r] = $\text{Cos}^{-1}[x / r]$ calculates the arc from the co-ordinate on the unit circle.

PM. Since computer programs implement Cos and Sin functions by tradition, the transforms to angles are:

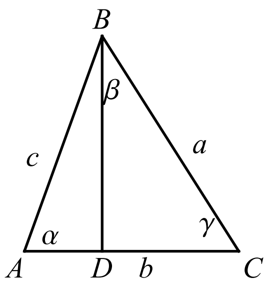
$$\alpha = \varphi \,/\, \Theta = \mathrm{ArcX}[X_v = x \,/\, r] \,/\, \Theta \quad \text{so that} \quad X_v = x \,/\, r = \mathbb{X}[\alpha] = \mathrm{ArcX}^{-1}[\varphi] = \mathrm{Cos}[\varphi]$$

$$\alpha = \varphi \,/\, \Theta = \mathrm{ArcY}[Y_v = y \,/\, r] \,/\, \Theta \quad \text{so that} \quad Y_v = y \,/\, r = \mathbb{Y}[\alpha] = \mathrm{ArcY}^{-1}[\varphi] = \mathrm{Sin}[\varphi]$$

7.8 Appendix

7.8.1 The xur and yur or cosine and sine rules for a triangle

The key theorem of analytic geometry uses a particular step that can be stated as a general rule for triangles: the x_{ur} or cosine rule for a triangle. Let us reproduce the situation.



1. Denote $p = AD$ and $h = BD$ (the height)
2. Pythagoras gives $c^2 = h^2 + p^2$ and $a^2 = h^2 + (b - p)^2$
3. Thus $a^2 = c^2 - p^2 + (b - p)^2 = c^2 - p^2 + (b^2 - 2bp + p^2) = c^2 + b^2 - 2bp$
4. Thus $a^2 = c^2 + b^2 - 2bp$ can be used if b, c and p are known.
5. $\mathbb{X}[\alpha] = x_{\mathrm{ur}}[\alpha] = p \,/\, c = \mathrm{Cos}[\alpha\, \Theta]$. Thus $a^2 = c^2 + b^2 - 2bc\, \mathbb{X}[\alpha]$.

The x_{ur} or cosine rule thus is $a^2 = c^2 + b^2 - 2bc\, \mathbb{X}[\alpha]$.

If $\gamma = 1/4$ then $p = b$ and the rule reduces to Pythagoras (line 4).

There is also a rule for y_{ur} or sine. With $\mathbb{Y}[\alpha] = h \,/\, c$ and $\mathbb{Y}[\gamma] = h \,/\, a$, elimination of h gives, more in general, also including β:

$$\frac{\mathbb{Y}(\alpha)}{a} = \frac{\mathbb{Y}(\beta)}{b} = \frac{\mathbb{Y}(\gamma)}{c}$$

PM. A teacher's problem in teaching is a student's problem in learning, and it is useful for students to be aware of some choices made here. Consider: triangles can have any direction while those in the unit circle are oriented in four limited directions. If you would have learned trigonometry for triangles of any orientation then you might have to unlearn and adjust for the system of co-ordinates. Or the other way around. This book tries to give a balance. The chapter on geometry is low key on the calculation of angles and sides: there is little need for that there. These cosine and sine rules are only introduced here since one of them shows up as a step in the proof for the key theorem of analytic geometry. But the rules can be used for angles of any orientation. This book likes you to be aware of these choices so that you can better deal with orientation. This approach seems much better than first provide training on calculating all kinds of angles and sides, and then actually unlearn again for application to the unit circle.

7.8.2 Doubling and halving angles

With $\mathbb{X}_{\alpha+\beta} = \mathbb{X}_\alpha \mathbb{X}_\beta - \mathbb{Y}_\alpha \mathbb{Y}_\beta$ we take $\alpha = \beta$ so that $\mathbb{X}_{2\alpha} = \mathbb{X}_\alpha{}^2 - \mathbb{Y}_\alpha{}^2$. With $\mathbb{X}_\alpha{}^2 + \mathbb{Y}_\alpha{}^2 = 1$ we get $\mathbb{X}_{2\alpha} = \mathbb{X}_\alpha{}^2 - (1 - \mathbb{X}_\alpha{}^2)$ or $\mathbb{X}_{2\alpha} = 2\mathbb{X}_\alpha{}^2 - 1$.

Halving gives $\alpha \to \beta/2$ and $\mathbb{X}_\beta = 2\mathbb{X}_{\beta/2}{}^2 - 1$ or $\mathbb{X}_{\beta/2} = \sqrt{\left(1 + \mathbb{X}_\beta\right)/2}$. Thus:

$$\mathbb{X}_{2\alpha} = 2\mathbb{X}_\alpha{}^2 - 1$$

$$\mathbb{X}_{\alpha/2} = \sqrt{(1 + \mathbb{X}_\alpha)/2}\,.$$

The method where a result of a calculation is used as the input for a next round of calculation is called "recursion". In human culture there are many instances of the use of repetition and repeated application, so this use is not particularly new. It becomes mathematics though when we make a systematic analysis of it.

- Starting at $\mathbb{X}_{1/4} = 0$ and repeatedly halving the angle. The first four steps:

$$\left\{0, \frac{1}{\sqrt{2}}, \frac{\sqrt{2+\sqrt{2}}}{2}, \frac{1}{2}\sqrt{2+\sqrt{2+\sqrt{2}}}\right\}$$

In the limit L, the result no longer changes and $L = \sqrt{(1 + L)/2}$ gives $L = 1$.

7.8.3 Dependency in the unit circle

Above we had a table of the dependencies in the four quadrants. There are different formats, and evaluation in *Mathematica* kindly verifies.

- For the angles.

	$\mathbb{X}$	$\mathbb{Y}$
Mirror over horizon	$\mathbb{X}[-\alpha] = \mathbb{X}[\alpha]$	$\mathbb{Y}[-\alpha] = -\mathbb{Y}[\alpha]$
Mirror over vertical	$\mathbb{X}\left[\frac{1}{2} - \alpha\right] = -\mathbb{X}[\alpha]$	$\mathbb{Y}\left[\frac{1}{2} - \alpha\right] = \mathbb{Y}[\alpha]$
Jump over diameter	$\mathbb{X}\left[\alpha + \frac{1}{2}\right] = -\mathbb{X}[\alpha]$	$\mathbb{Y}\left[\alpha + \frac{1}{2}\right] = -\mathbb{Y}[\alpha]$
Positive quarter turn	$\mathbb{X}\left[\alpha + \frac{1}{4}\right] = -\mathbb{Y}[\alpha]$	$\mathbb{Y}\left[\alpha + \frac{1}{4}\right] = \mathbb{X}[\alpha]$
Negative quarter turn	$\mathbb{X}\left[\alpha - \frac{1}{4}\right] = \mathbb{Y}[\alpha]$	$\mathbb{Y}\left[\alpha - \frac{1}{4}\right] = -\mathbb{X}[\alpha]$

- Translated to Cos[α Θ], then simplify. The mirror over the horizon is directly true without simplification.

True	True
$\cos\left(2\pi\left(\frac{1}{2}-\alpha\right)\right) = -\cos(2\pi\alpha)$	$\sin\left(2\pi\left(\frac{1}{2}-\alpha\right)\right) = \sin(2\pi\alpha)$
$\cos\left(2\pi\left(\alpha+\frac{1}{2}\right)\right) = -\cos(2\pi\alpha)$	$\sin\left(2\pi\left(\alpha+\frac{1}{2}\right)\right) = -\sin(2\pi\alpha)$
$\cos\left(2\pi\left(\alpha+\frac{1}{4}\right)\right) = -\sin(2\pi\alpha)$	$\sin\left(2\pi\left(\alpha+\frac{1}{4}\right)\right) = \cos(2\pi\alpha)$
$\cos\left(2\pi\left(\alpha-\frac{1}{4}\right)\right) = \sin(2\pi\alpha)$	$\sin\left(2\pi\left(\alpha-\frac{1}{4}\right)\right) = -\cos(2\pi\alpha)$

7.8.4 Solving equations

Solve $\mathbb{X}_\alpha^2$ - $\mathbb{X}_\alpha$ = 0. Solved by $\mathbb{X}_\alpha$ ($\mathbb{X}_\alpha$ - 1) = 0, thus $\mathbb{X}_\alpha$ = 0 or $\mathbb{X}_\alpha$ = 1.

Thus either $\alpha = 1/4 + k / 2$ or $\alpha = k$ uma for $k = 0, 1, 2, ...$

Or solve $\cos(\varphi)^2$- $\cos(\varphi)$ = 0. Solved by $\cos(\varphi)$ (1 - $\cos(\varphi)$) = 0. Cos(φ) = 0 or $\cos(\varphi)$ = 1.

Thus $\varphi = \Theta/4 + k\,\Theta/2$ or $\varphi = k\,\Theta$ rad

Or $\varphi = \pi / 2 + k\,\pi$ or $\varphi = k\,2\pi$ rad.

7.8.5 Calculating arcs

Consider points $P = \{x, 0.8\}$ and $Q = \{0.1, y\}$ on the unit circle for negative solutions of x and y. Calculate the length of the shortest arc from P to Q.

Done by ArcYur[0.8] = 0.147584. The angle to P is α = 1/2 - 0.147584 = 0.352416.

ArcXur[0.1] = 0.234058. The angle to Q is β = 1 - 0.234058 = 0.765942.

The shortest arc on the unit circle is Θ (β - α) = 2.60.

Or ArcSin[0.8] = 0.927. The arc to P is $\varphi = \pi$ - 0.927 = 2.214.

ArcCos[0.1] = 1.471. The arc to Q is $\psi = 2\pi$ - 1.471 = 4.813.

The shortest arc between P and Q is $\psi - \varphi$ = 2.60.

7.8.6 Transformations

With α and β for angles and φ and ψ for arcs we can map from angles to arcs and vice versa. We may keep the same variable names but may also change them. When we move from arcs to angles then it makes sense to replace π by Θ/2 but a separate routine is useful for the converse - since it is a separate issue and Θ is useful for arcs too.

For $x_{ur}[\alpha]$ there is Cos[φ]. The angle function is represented by (1) Xur that translates to Cos on the spot, (2) xur that is symbolic and displays as $\mathbb{X}$, and (3) of course we may directly type $\mathbb{X}$. Similarly for y_{ur} and Sin and t_{ur} and Tan.

- From symbolic cos to symbolic x_{ur} while keeping the variable.

Cos[ψ + Pi/4] // ArcToAngle

$$\mathbb{X}\left(\frac{\psi}{\Theta}+\frac{1}{8}\right)$$

- It is more convenient to have standard replacement of names.

SetOptions[ArcToAngle, Angle → {α, β, γ}, Arcs → {φ, ψ, ϑ}];
SetOptions[AngleToArc, Angle → {α, β, γ}, Arcs → {φ, ψ, ϑ}];

Cos[ψ + Pi/4] // ArcToAngle

$$\mathbb{X}\left(\beta+\frac{1}{8}\right)$$

Cos[2 φ + 3 ψ + Pi/6] // ArcToAngle

$$\mathbb{X}\left(2\,\alpha+3\,\beta+\frac{1}{12}\right)$$

- If k is a unit counter then it should not be replaced.

Cos[φ + k Pi/6] // ArcToAngle

$$\mathbb{X}\left(\alpha+\frac{k}{12}\right)$$

Result // AngleToArc // ToPi

$$\cos\left(\frac{\pi\,k}{6}+\varphi\right)$$

- If ϑ is not a unit counter then its coefficient is a proper coefficient.

Cos[φ + ϑ Pi/6] // ArcToAngle

$$\mathbb{X}\left(\alpha+\frac{\gamma\,\Theta}{12}\right)$$

Result // AngleToArc // ToPi

$$\cos\left(\varphi+\frac{\pi\,\vartheta}{6}\right)$$

`FromPi[`*expr*`]` `ToΘ[`*expr*`]`	replaces π in expr with $\Theta/2$
`FromΘ[`*expr*`]` `ToPi[`*expr*`]`	replaces Θ in expr with $2\,\pi$

`ArcToAngle[`*expr*`]`	transforms Cos, Sin and Tan into xur, yur and tur, and Pi into Θ/2
`AngleToArc[`*expr*`]`	transforms xur, 𝕏, yur, 𝕐, tur and 𝕋 into Cos, Sin and Tan. It leaves Θ: see ToPi

ArcToAngle and AngleToArc have the same input formats and orders.

Option Simplify → True is default. Option Arcs can contain variable names for arcs, and Angle (without s) for the corresponding angles (defaults { }) and then those are substituted, with appropriate $\varphi = \alpha\,\Theta$.

`AngleToArc[Rule]`	the replacement rules
`AngleToArc[`*expr*, α, φ`]`	substitutes too
`AngleToArc[`*expr*, α, φ, β, ψ, ...`]`	for alternating sequence
`AngleToArc[`*expr*, {α, β}, {φ, ψ}`]`	for collected lists (that might also be in the options

ArcToAngle and AngleToArc have the same input formats and orders.

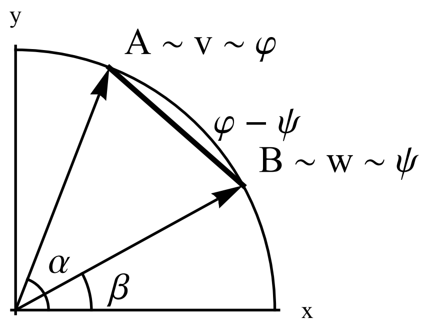
`𝕏[ ... ]`	used as a label for xur, see ArcToAngle
`xur[`*expr*`]`	displays as 𝕏[expr], see ArcToAngle
`Xur[`α`]`	gives Cos[Θ α] for angle α measured in Unit Turn or Unit Measure or Meter Around (UMA). It gives the x value on the circle with unit radius. Xur[α] equals the ratio of the horizontal value to the radius. Let $\Theta = 2\,\pi$, then Θ UMA = 1 radian
`ArcXur[`*x*`]`	gives the angle α = ArcCos[x] / Θ such that Xur[α] = x

The same for Yur and Tur. 𝕏 is `\[DoubleStruckCapitalX]`.

7.8.7 A corollary

Given $\mathbb{X}_{\alpha+\beta} = \mathbb{X}_\alpha\,\mathbb{X}_\beta - \mathbb{Y}_\alpha\,\mathbb{Y}_\beta$ it is straightforward to find $\mathbb{X}_{\alpha-\beta} = \mathbb{X}_\alpha\,\mathbb{X}_\beta + \mathbb{Y}_\alpha\,\mathbb{Y}_\beta$, using the properties that $\mathbb{X}_{-\beta} = \mathbb{X}_\beta$ and $\mathbb{Y}_{-\beta} = -\,\mathbb{Y}_\beta$.

It can be a good exercise to directly prove it, though, using below graph. Point *A* associates with angle α, vector v and arc φ. Point *B* associates with angle β, vector w and arc ψ. The chord *AB* is the side of a triangle given by the radii but also by Pythagoras.



- Xur rule and unit radius: $AB^2 = 1 + 1 - 2 * 1 * 1 * \mathbb{X}_{\alpha-\beta} = 2 - 2\,\mathbb{X}_{\alpha-\beta}$.
- Pythagoras: $AB^2 = (\Delta X)^2 + (\Delta Y)^2 = \left(\mathbb{X}_\alpha - \mathbb{X}_\beta\right)^2 + \left(\mathbb{Y}_\alpha - \mathbb{Y}_\beta\right)^2 = \mathbb{X}_\alpha{}^2 - 2\,\mathbb{X}_\alpha \mathbb{X}_\beta + \mathbb{X}_\beta{}^2 + \mathbb{Y}_\alpha{}^2 - 2\,\mathbb{Y}_\alpha \mathbb{Y}_\beta + \mathbb{Y}_\beta{}^2 = 2 - 2\,(\mathbb{X}_\alpha \mathbb{X}_\beta + \mathbb{Y}_\alpha \mathbb{Y}_\beta)$
- Combining these gives $\mathbb{X}_{\alpha-\beta} = \mathbb{X}_\alpha\,\mathbb{X}_\beta + \mathbb{Y}_\alpha\,\mathbb{Y}_\beta$

7.8.8 A technical note

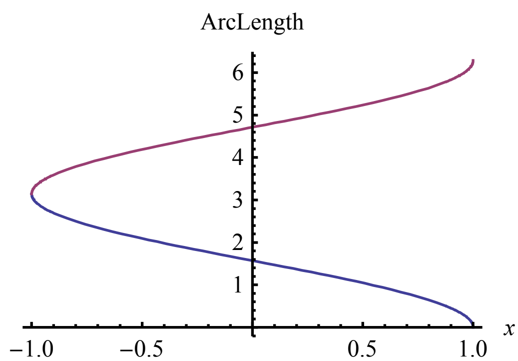
The arc functions rely on standard routines in *Mathematica* such as for complex numbers that are discussed below.

- These are the definitions.

```
See[ ArcLength[{x, y}],   ArcX[x],   ArcY[y] ]
```

$-\pi\,(1 - \text{sgn}(y))\,\text{sgn}(y) + \arg(x + i\,y) \qquad \cos^{-1}(x) \qquad \sin^{-1}(y)$

```
Plot[{ArcLength[{x, Sqrt[1 − x^2]}], ArcLength[{x, − Sqrt[1 − x^2]}]
    }, {x, −1, 1}, PlotRange → All, AxesLabel → {x, ArcLength}]
```



This again demonstrates that the complexity of trigonometry derives mainly from pushing a 2D problem into 1D functions.

8. The complex plane

8.1 Notation of the above as complex number

The key theorem showed that trigonometry is an essential part of vector analysis. We will now rewrite these results into the form of "complex numbers". Basically you will learn nothing new. Except, that this new way of writing may simplify some steps in deductions. Complex numbers are much used and it is very useful when you understand the notation. The notation of complex numbers can be seen as more efficient than the notation we have used for the vectors. Eventually however the vector notation can be generalized better into more-dimensional systems while the notation of complex numbers gets a bit stuck to 2D or 3D problems. Complex numbers are handy and elegant and educational like horseback riding but also limited in trying to reach the stars. An advantage of complex numbers is that square roots of negative numbers always have solutions. It will be a reason why *Mathematica* uses complex numbers as the standard.

8.2 The imaginary number

A simplification of $x \overrightarrow{X} + y \overrightarrow{Y}$ is to write $\overrightarrow{X} = 1$ and $\overrightarrow{Y}$ as $\imath$. (Or *i* or *I*.) We get:

$$v = \{x, y\} = x \overrightarrow{X} + y \overrightarrow{Y} = x + \ y \, \imath$$

Where we said horizontal and vertical, now x is called the *real* part and y is called the *imaginary* part. When we take $x = 0$ then $v^2 = y^2 \, \imath^2 = y^2 \, \overrightarrow{Y} \, \overrightarrow{Y} = - \, y^2 \, \overrightarrow{X} = -y^2$. We can divide by any $y \neq 0$ and then get $\imath^2 = -1$. Since $\sqrt{-1}$ has not been defined yet, we may solve $\imath^2 = -1$ as $\imath = \pm\sqrt{-1}$. It is standard to define $\imath = \sqrt{-1}$. In the historical past mathematicians were only familiar with real numbers and roots of nonnegative real numbers. Hence the root of -1 struck them as a rather imaginary phenomenon. Whence the historical name of $\imath$ as *the* imaginary number. Numbers with $\imath$ are called imaginary numbers in general and this plane thus is called the imaginary plane.

Working with $\imath = \sqrt{-1}$ is tricky. Regard $1 = \sqrt{1} = \sqrt{-1 \, * \, -1} = \sqrt{-1} * \sqrt{-1} = \imath^2 = -1$. If you do not know the proper theory on vectors above you will be flabbergasted. This outcome can be understood from the notion that we need two rotational steps to get from 1 to -1. The third equality sign then is improper. This may be easily overlooked so beware your steps.

Though working in the imaginary plane can be tricky it has the advantage of formulas that are a bit easier to write than $x\overrightarrow{X} + y\overrightarrow{Y}$. People also get used to a language. Hence it has remained a popular approach.

A quick result is the vector product that we used in the key theorem of analytic geometry, that we have learned to recognize as the x_{ur} or cosine rule:

$$(a + b\, i)(x + y\, i) =$$

$$a\, x + a\, y\, i + b\, x\, i + b\, y\, i^2 =$$

$$a\, x - b\, y + (a\, y + b\, x)\, i$$

It was an option for this present book to develop analytic geometry starting with complex numbers. This option was rejected for the reason that i is a bit vague when you first meet it. Starting with $\{x, y\} = x\overrightarrow{X} + y\overrightarrow{Y}$ is a more solid way to really grind in the notion of the two axes. And a base for future n-dimensionality. The objective of this book is to allow you to understand analytic geometry, in a quick and easy and accurate but also deep and solid manner. Once you understand it you choose your own notation.

Given our knowledge of vectors the following properties are straightforward. We might have developed these properties in vector notation but once you master complex numbers then you will be happy to use the current derivations.

8.3 Absolute value and conjugate

There are two definitions that matter.

- The absolute value or modulus. *Mathematica* presumes that x and y may be unknown complex numbers and thus leaves the expression as it stands instead of producing Pythagoras as NRadius does.

$$\begin{pmatrix} \text{Abs}[v] & \text{Abs}[x + i\, y] & \text{Abs}[4 + 3\, i] & \text{NRadius}[\{x, y\}] \\ |v| & |x + i\, y| & 5 & \sqrt{x^2 + y^2} \end{pmatrix}$$

- The conjugate of $v = x + y\, i$ is $\overline{v} = x - y\, i$. *Mathematica* uses v^*. The conjugate is the number mirrored over the horizontal axis. When we do not specify what are real numbers then they are not recognized as such.

```
See[Conjugate[v], Conjugate[x + y i], Conjugate[4 + 3 i]]
```

$v^* \qquad x^* - i\, y^* \qquad 4 - 3\, i$

```
FullSimplify[ Result , Assumptions → {{x, y} ∈ Reals}]
```

v^* $\quad x - i\,y \quad$ $4 - 3\,i$

A key to complex numbers is this property of multiplication. The product of these two complex expressions gives the square of the absolute value, which is a real number again.

```
(x + y i) (x − y i) // Simplify
```

$x^2 + y^2$

We now have sufficient material to reproduce the Pythagorean Theorem.

- The conjugate is important because $v\,\bar{v} = |v|^2$. It is a real number too. Thus $v\,\bar{v} = (x + y\,i)\,(x - y\,i) = x^2 - y^2 i^2 = x^2 + y^2 = |v|^2$. Also $|v| = |\bar{v}|$.

```
(4 + 3 i) (4 + 3 i)*
```

25

```
Conjugate[v] v // FullSimplify
```

$|v|^2$

- The conjugate of a product is the product of conjugates: $\overline{v\,w} = \bar{v}\,\bar{w}$.

```
((a + b i) (x + y i))*
```

$(a^* - i\,b^*)(x^* - i\,y^*)$

The length of the product is the product of the lenghts: $|\,v\,w\,| = |v|\,\,|w|$. Above we have spent quite some time and formulas in deriving this but now it goes fast using the conjugates: $|v\,w|^2 = v\,w\,\overline{v\,w} = v\,\bar{v}\,\,w\,\bar{w} = |v|^2\,\,|w|^2$.

- Expressed with roots instead of squares.

```
Abs[ (a + b i ) (x + y i )] // FunctionExpand
```

$|a + i\,b|\,|x + i\,y|$

Division is a neat result. In a quotient $v\,/\,w$ we multiply both positions with the conjugate of the denominator. The denominator then reduces to a real number and we get a standard complex number again. Thus $v\,/\,w = (v\,\,\bar{w})\,/\,(w\,\bar{w}) = (v\,\,\bar{w})\,/\,|w|^2$.

- *Mathematica* is so powerful that it can leave this as it stands.

```
(x + y i) / (a + b i) // FunctionExpand
```

$$\frac{x + i\,y}{a + i\,b}$$

- Enforcing the simplification for numerator and denominator.

```
Division[ Simplify[Times[##]] &, (x + y𝕚) / (a + b𝕚), {a − b𝕚}]
```

$$\frac{(a - i\, b)\,(x + i\, y)}{a^2 + b^2}$$

- *Mathematica* however takes complex numbers as simpler.

```
Simplify[ Result , Assumptions → {{a, b, x, y} ∈ Reals}]
```

$$\frac{x + i\, y}{a + i\, b}$$

The length of a ratio is the ratio of the lengths:

- $|v / w| = |v\ \overline{w} / (w\, \overline{w})| = |v\ \overline{w}| / |w|^2 = |v|\ |\overline{w}| / |w|^2 = |v| / |w|$.

```
Abs[(x + y𝕚) / (a + b𝕚)] // FunctionExpand
```

$$\frac{|x + i\, y|}{|a + i\, b|}$$

8.4 Real and imaginary parts, Re and Im

Re[ ] takes the real part and Im[ ] takes the imaginary part, i.e. the coefficient of i.

- In $x + y\, i$ variables x and y may be complex numbers too.

```
See[ Re[x + y𝕚],      Im[x + y𝕚] ]
```

$\mathrm{Re}(x) - \mathrm{Im}(y) \qquad \mathrm{Im}(x) + \mathrm{Re}(y)$

```
Simplify[ Result , Assumptions → {{x, y} ∈ Reals}]
```

$x \qquad y$

- Pythagoras. Application of FullSimplify would generate True.

```
Abs[v] ^2 == Conjugate[v] v  // FunctionExpand
```

$\mathrm{Im}(v)^2 + \mathrm{Re}(v)^2 = v\, v^*$

The real part is halfway on the line between v and its conjugate.

- The real part.

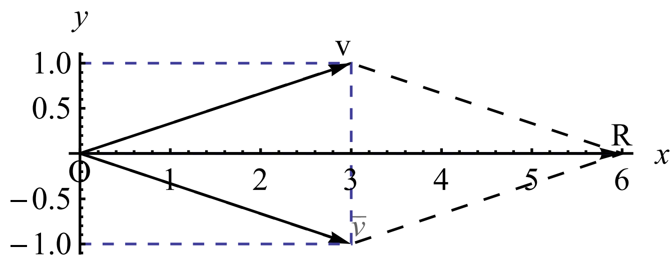
```
Re[v] == (v + v*) / 2
```

$$\mathrm{Re}(v) = \frac{v^* + v}{2}$$

Simplify[Result /. v → x + y i, Assumptions → {{x, y} ∈ Reals}]

True

- For the vector {3, 1}, the standard addition with the conjugate {3, -1} gives {6, 0} and halfway we find Re[{3, 1}].



The imaginary part is halfway on the difference between v and its conjugate. NB. The imaginary part is a real number and thus is on the horizontal axis.

- Take half and divide by i to eliminate it from the numerator.

Im[v] == (v − v*) / (2 i)

$\text{Im}(v) = -\frac{1}{2} i \left(v - v^*\right)$

Simplify[Result /. v → x + y i, Assumptions → {{x, y} ∈ Reals}]

True

Take for example $v = 3 + i$.

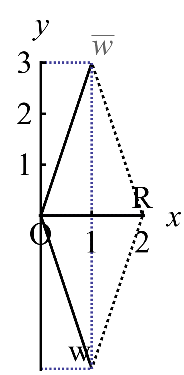
First take $w = -i\,v$ and $z = i\,\overline{v}$.

$w = -i\,v = -3\,i + 1.$

$z = i\,\overline{v} = i\,(3 - i) = 3\,i + 1 = \overline{w}.$

Then $w + z = w + \overline{w} = \{2,\ 0\}$.

Im[{3, 1}] is halfway from $w + \overline{w} = \{2,\ 0\}$.



8.5 Argument function, Arg

The argument Arg of a complex number is the arc from {1, 0} counterclockwise. Values for a negative y get a negative arc too.

- The key theorem of analytic geometry again. The argument of a product is the sum of arguments.

Arg[v w] == Arg[v] + Arg[w];

- Application to a specific number requires explanation that the coefficients are reals. Note the Floor function.

PowerExpand[Arg[(*a* + *b* i) (*x* + *y* i)] , Assumptions → {{x, y, a, b} ∈ Reals}]

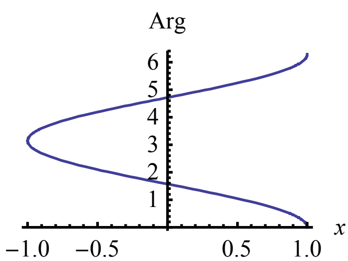
$$2\pi\left\lfloor -\frac{\arg(a+i\,b)}{2\pi}-\frac{\arg(x+i\,y)}{2\pi}+\frac{1}{2}\right\rfloor+\arg(a+i\,b)+\arg(x+i\,y)$$

- ArcLength above is from 0 to Θ. Arg is from -π to π.

See[Arg[x + y i], Arg[1 + i], Arg[1 − i], ArcLength[{1, −1}]]

$\arg(x+i\,y)$ $\quad \frac{\pi}{4}$ $\quad -\frac{\pi}{4}$ $\quad \frac{7\pi}{4}$

Plot[y = $\sqrt{1-x^2}$; {Arg[x + y i], 2 Pi + Arg[x − y i]}, {x, −1, 1}, AxesLabel −> {x, Arg}]



8.6 The polar form

We get another format for the polar form:

$$v = |\mathrm{v}|\ (\mathbb{X}[\alpha] + \mathbb{Y}[\alpha]\ i)$$

$$v = |\mathrm{v}|\ (\mathrm{Cos}[\varphi\ ] + \mathrm{Sin}[\varphi]\ i)$$

- The key theorem of analytic geometry.

***v w* == *r* (Cos[φ] + Sin[φ] i) ∗ *s* (Cos[ψ] + Sin[ψ] i)**

$v\,w = r\,s\,(\cos(\varphi)+i\sin(\varphi))\,(\cos(\psi)+i\sin(\psi))$

assump = Assumptions → {{*r*, φ, s, ψ} ∈ Reals};
Simplify[PowerExpand[Result , assump], assump]

$v\,w = r\,s\,(\cos(\varphi+\psi)+i\sin(\varphi+\psi))$

8.7 Euler's form

8.7.1 What it is and why it works

The recovered exponent of a product is the sum of the recovered exponents, or rex[x^a y^b] = a rex[x] + b rex[y]. This property has the same structure as the multiplication of vectors and addition of arcs. Thus exponents can be used to represent arcs. A base can be chosen that is agreeable with various properties: and this appears to be the number e = 2.71828....

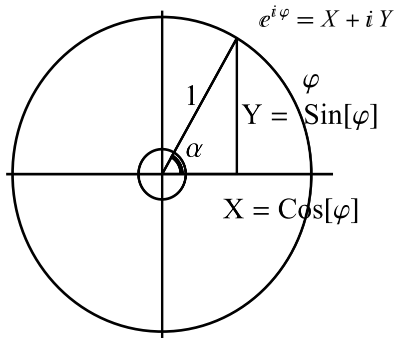
Later we will discuss calculus. This is a good place to make a point that you may not understand yet. Teachers will, students may not. It is better to put it here where it belongs than discussing it later where it will be a bit out of context. Once you understand calculus you can look back at this short discussion, where it then is at the right place.

The pivot around which this then works is e^i = Cos[1] + Sin[1] $i \approx$ 0.540302 + 0.841471 i, thus the co-ordinates of an arc of 1 radian. Arcs φ raise this pivot to the power, thus $(e^i)^\varphi$, and the radius is multiplied with it. Thus: (1) a real part c in the exponent gives the radius $r = e^c$ and (2) the imaginary part in the exponent gives the co-ordinates on the unit circle. Together:

$$e^{c + i\varphi} = e^c e^{i\varphi} = r\, e^{i\varphi} = r(\text{Cos}[\varphi] + \text{Sin}[\varphi]\, i)$$

With $c = 0$ there is the unit circle, and $\varphi = 1$ holds for 1 radian. In general it still is the point $v = \{x, y\} = r\,\{\text{Cos}[\varphi], \text{Sin}[\varphi]\}$ with $r = |v|$.

- $r = 1$ and $\varphi = 1$ give e^i = Cos[1] + Sin[1] $i \approx$ 0.540302+ 0.841471 i



- Euler's definition, true by definition, but codifying the key theorem.

```
Simplify[r ⅇ^(ⅈ φ) == r (Cos[φ] + Sin[φ] ⅈ), Assumptions → {{r, φ} ∈ Reals}]
```

True

PM. Above we saw how 1 = -1 could be proven with improper steps. Let us now manipulate φ in radians to show that any complex number on the unit circle must be 1.

$$\left\{e^{i\varphi} = \left(e^{2 i \pi \varphi}\right)^{\frac{1}{2\pi}}, \quad \Rightarrow \quad, e^{i\varphi} = \left(e^{i\,2\pi}\right)^{\frac{\varphi}{2\pi}} = 1\right\}$$

Result // PowerExpand

$$\{\text{True}, \quad \Rightarrow \quad, e^{i\varphi} = 1\}$$

8.7.2 Key theorem of analytic geometry

- The key theorem of analytic geometry now has a very basic form.

v w == (r $e^{i\varphi}$) ∗ (s $e^{i\psi}$)

$$v\, w = r\, s\, e^{i\varphi + i\psi}$$

Euler's form is a definition and subsequently we require an existence proof that indeed the arcs can be added. See the existence proof in §6.2. Thus the Euler form is basically an efficient way of stating the key theorem. It does a bit more by joining up complex numbers with exponential numbers.

- The key theorem of analytic geometry in some different formats.

(r $e^{i\varphi}$) ∗ (s $e^{i\psi}$) == *r* (Cos[φ] + Sin[φ] i) ∗ *s* (Cos[ψ] + Sin[ψ] i)

$$r\, s\, e^{i\varphi + i\psi} = r\, s\, (\cos(\varphi) + i \sin(\varphi))\, (\cos(\psi) + i \sin(\psi))$$

Simplify[Result, Assumptions → {{*r*, s, φ} ∈ Reals}]

True

8.7.3 Euler's form in angles

Straightforward:

- Euler's form of the complex number as function of angle $\alpha = \varphi / \Theta$:

$$\boxed{v / r = e^{i\,\alpha\,\Theta} = \mathbb{X}(\alpha) + i\, \mathbb{Y}(\alpha) = \cos(\alpha\,\Theta) + i \sin(\alpha\,\Theta)}$$

PM. The following is a bit of fun. A full circle gives $e^{i\Theta}$ = Cos[Θ] + Sin[Θ] i = 1 + 0 i, thus the unit vector on the horizontal axis. If we take $e^{i\Theta}$ = 1 + 0 i = 1 and thus erroneously multiply out i, then it seems that $(e^{i\Theta})^{\alpha} = 1^{\alpha} = 1$ so that there is not much of a formula to work with here. To remove the mystery: this is like the proof above that 1 = -1. We must distinguish between an operator and the arithmetic for reals. The outcome 1 is the accurate value when $\varphi = \Theta$. The only thing that we do now is rescaling. The notation as imaginary number and as a power of e is fundamentally shorthand for Cos[... Θ] $\overrightarrow{X}$ + Sin[... Θ] $\overrightarrow{Y}$.

9. Approximation of Θ

9.1 Noblesse oblige

It is not altogether clear why this book bothers to look into the approximation of Θ. *Mathematica* has Θ already to great accuracy. Books and websites abound (e.g. Jon and Peter Borwein). If you are interested in that then you are in paradise already.

Still, in a book on analytic geometry and calculus there is a sense of obligation to say something.

1. There is the issue whether Θ is really dimensionless.
2. A primer is a primer. You might like some material for practicing. Below applies Viètes method with steps that each apply the Pythagorean Theorem. It is very basic. It avoids a formal development of polygons. There is no use of the trigonometric functions. Thus it can be appreciated as an exercise to get some more familiarity with geometric reasoning. Naturally, it uses the X and Y values on the unit circle and therefor the trigonometric functions are implicit. It is the duty of trigonometry to convince and teach students that it is useful to make those explicit. Discussing the calculation of Θ might suit that.
3. Having seen the proof may help to understand the formulas for circumference and area. The proof establishes proportionality.
4. It might be a good stepping stone to later understand calculus on the circle.
5. The calculation uses recursion. Archimede in his calculation of π already used recursion. It is a fairly nice way to grow aware of the issues involved.

PM. Use of the symbol Θ in *Mathematica* happens to be awkward since it has not been programmed in the N[ ] function. Routines ToΘ, FromΘ, ToPi and FromPi replace vice versa.

```
N[{Θ, 2π, π}, 20]
{Θ, 6.2831853071795864769, 3.1415926535897932385}
```

```
(expr = 5/6 Θ // FromΘ) == (expr // ToΘ)
```

$$\frac{5\pi}{3} = \frac{5\Theta}{6}$$

Independently, Bob Palais also judged the selection of π over Θ to be a historical error.
See Palais, R. (2001a), "π Is wrong!", The mathematical intelligencer, Vol 23, no 3 p7-8.

9.2 The dimension of Θ

Our comments on the dimension of Θ and the choice of the unit of measurement are:

1. It is doubtful whether $\Theta = 2\pi$ is dimensionless. We can define a ratio but a movement around is a different dimension than a one-dimensional radius or diameter. Moving around requires two dimensions.
2. In geometry we frequently do not specify a unit of measurement. This only becomes relevant for practical application when the engineers take over from the mathematicians. This however is a different issue. What now is at issue is that an angle is different from a line, even though they are in the same plane. Originally we took an angle as a section of the plane, later we replaced this with measurement along an arc. In the plane a meter one way is the same as a meter in another direction, but here we deal with a change in direction.
3. The proper approach is to use the Unit Turn as the unit measure for turns around. A turn is a Unit Measure Around (UMA) as provided by the angular circle.
4. When we measure the circumference or how the circle rolls one cycle then a point P on the angular circle indeed moves over the distance of that UMA. When the engineers step in and we require a dimension then we already have a unit of measurement in the Meter. The distance rolled by the circle then is measured in Meters. This means that the point P also has moved one Meter. If the measure is the Meter then we keep the label UMA in a different interpretation. In that engineering world the unit circle with $r = 1$ Meter also has a radian of 1 Meter, so for radians the principle of assigning 1 to 1 Meter applies too.
5. Using (a) dimensionless α as a ratio or (b) the unit of measurement Unit Turn (UMA) or (c) the UMA-read-as-Meter causes a great simplification of various formulas. Formulas clutter with π but the symbol can now disappear in many cases.

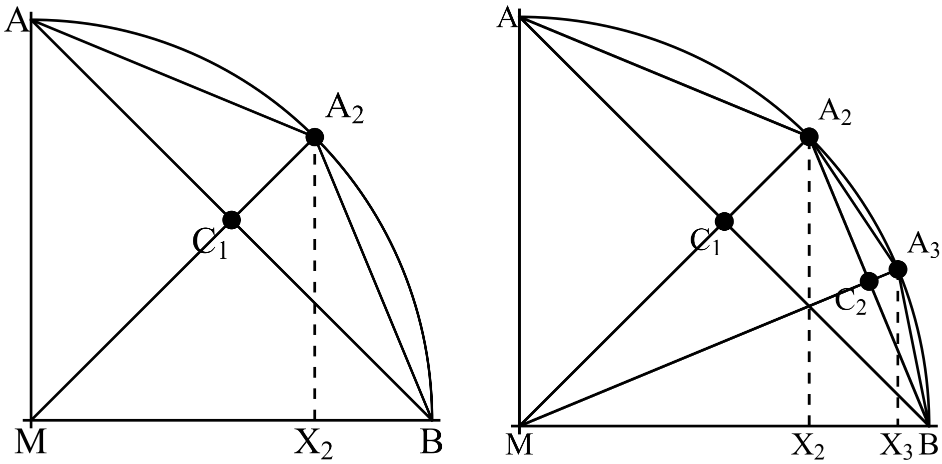
6. Using a half circle and π as a standard is curious since it is more logical to use a whole circle. Note the shift in denominator when radian = arc / **radius** but π = (full arc) / **diameter**. Practiced users use the label 2π for whole turns but curiously not in their unit of measurement. Always dividing and renormalizing on the spot remains awkward for didactics. Students think that they calculate fractions of π while they should focus on turns. Students tend to punch 2π into the calculator which is not the objective. (Now they will be tempted to punch Θ / 2 (if the symbol is there) but with above didactics that is less likely to happen.)
7. The use of π comes with the convention to use positive turns to π and negative turns to $-\pi$. The advantage of this is nonexistent since we can also say 1/2 and -1/2 for angles and $\Theta/2$ and $-\Theta/2$ for arcs, with the advantage that the latter are readily understood.
8. Though teaching must be effective in teaching to a standard too, it is useful to keep in mind that some standards may be less effective in themselves too.

9.3 The actual approximation

9.3.1 The method

Check out the graph below. We rely on geometric insight of halving of an angle and Pythagoras but also use vector addition and the slope of a line. Above in §7.8.2 we already derived a repetitive relation on halving an angle but we need a bit more for the sides of a triangle.

1. We look at the first quadrant and multiply the result by 4. We take B fixed. The startpoint is at $A = A_1$, and we take ever shorter chords towards B: A_1, A_2, ...
2. Pythagoras tells us that $AB = \sqrt{2}$. Multiply by 4.
3. Halving the angle gives us A_2 on the circle and the intersection C_1 with AB. Since the triangles between the two radii are isosceles, the lines are perpendicular and C_1 is halfway. Thus C_1 = ½ ({0, 1} + {1, 0}) = { ½, ½ }.



4. The line through MA_n has formula $y = s\,x$. With $X^2 + Y^2 = 1$ we get $X^2 + (s\,X)^2 = 1$ or the co-ordinates on the circle are $X = 1 / \sqrt{1 + s^2}$ and $Y = s / \sqrt{1 + s^2}$.
5. The co-ordinates of C_1 give $s = 1$ and thus $A_2 = \{1 / \sqrt{2}, 1 / \sqrt{2}\}$.
6. Repeating, the length $A_2B = \sqrt{(1 - X)^2 + Y^2} = \sqrt{2 - 2 \Big/ \sqrt{1 + s^2}}$ in general and for $s = 1$ we have $A_2B = \sqrt{2 - \sqrt{2}}$. Multiply by 4 * 2 = 8.
7. Repeating, we bisect the angle again and find C_2 and A_3. Vector addition and halving gives $C_2 = ½\,(A_2 + B\,)$, we find the slope s, the co-ordinates of A_3, the length of $A_3\,B$ and we multiply by 4 * 4 = 16.
8. Repeat while you want and the estimate of Θ is $2^{n+1} A_n\,B$ when we start at $n = 1$.
9. But first simplify. Intermediate steps can be eliminated because there is a regularity. With $f[0] = 0$ and $f[n] = \sqrt{2 + f[n-1]}$ we find $A_n\,B = \sqrt{2 - f[n-1]}$. We can also prove that $f[n] = 2$ in the limit since then $f[n] \approx f[n - 1]$ and see $a = \sqrt{2 + a}$, so that we do not cross the line and new additions $A_n\,B$ are practically zero. Archimede gave a lower and upper bound but given this limit and current computer power this current method may do well. The End.

See http://demonstrations.wolfram.com/VietesNestedSquareRootRepresentationOfPi/

9.3.2 The result

The following routine ΘPiApproximation takes above steps and has not been optimized since, indeed, *Mathematica* does a much better job.

- A_n is on step n and its estimate of Θ on step n + 1. In the third step the formal output does not fit on the page but numerically at the sixth step the first two digits already look familiar.

$$\begin{pmatrix} \Theta & \pi & \{X, Y\} & \text{Slope} & \text{Step} \\ - & - & \{0, 1\} & \infty & 1 \\ 4\sqrt{2} & 2\sqrt{2} & \left\{\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right\} & 1 & 2 \\ 8\sqrt{2-\sqrt{2}} & 4\sqrt{2-\sqrt{2}} & \left\{\frac{\sqrt{2+\sqrt{2}}}{2}, \frac{1}{\sqrt{2\left(2+\sqrt{2}\right)}}\right\} & -1+\sqrt{2} & 3 \end{pmatrix}$$

Θ	π	{X, Y}	Slope	Step
–	–	{0., 1.}	∞	1
5.65685	2.82843	{0.707107, 0.707107}	1.	2
6.12293	3.06147	{0.92388, 0.382683}	0.414214	3
6.24289	3.12145	{0.980785, 0.19509}	0.198912	4
6.2731	3.13655	{0.995185, 0.0980171}	0.0984914	5
6.28066	3.14033	{0.998795, 0.0490677}	0.0491268	6

The algorithm and the actual numbers clarify some points:

1. The accuracy depends upon the accuracy of taking roots.
2. Convergence and speed are tricky. We see 6.27 flip to 6.28 and perhaps if we take many more steps it will flip to 6.29 ... Archimede's upper boundary approach then is useful. But at a cost of speed.
3. Advances in trigonometry allow more compact specifications. If we use a regular polygon and trigonometric tables are created with algebraic expansion then we get (expressing everything in degrees to prevent circularity in radians, and using surface - and the reason why this works is that the slope becomes the arc and Sin at 0 has a slope of 1 so that 360° is translated in Θ):

```
Limit[ n * Sin[360 Degree / n ] ,  n → Infinity] // N
6.28319
```

4. The Gauss-Legendre algorithm then is impressive.

9.3.3 A key point

The method reaffirms that the circumference is proportional to the radius, thus *Cir* = r Θ. Also, we use triangles that have *Sur* = $h\ w$ / 2. Since the height is r and the cumulative width is given by the circumference, circle surface is *Sur* = r (r Θ) / 2 = ½ r^2Θ.

9.3.4 Slope space ?

You were first introduced to lines where slopes were the key issue and then subsequently you are taken to angular space where slopes play a minor role. Is this a form of cruelty ?

Part of our art of mathematics is to recognize patterns and options to formalize. In above algorithm we found the co-ordinates on the circle as $X = 1 / \sqrt{1 + s^2}$ and $Y = s / \sqrt{1 + s^2}$. The slope $s = t_{ur} = y_{ur} / x_{ur}$ = Tan = Sin / Cos. We haven't discussed this but the points on the circle can be found not only by the angle but also with the tangens without explicitly referring to the angle. The values of the tangens can be found using the line $y = s\ x$ and the line $x = 1$. The drawback is that there are two points on the circle that have that slope. We may consider to insert a sign of x and then create Slope Space:

$\{x, y\} \leftrightarrow$ Polar[r, φ] $\leftrightarrow$ Angular[r, α, 1] $\leftrightarrow$? Slope[r, $s = y/x = t_{ur}[\alpha]$, Sign[x]]

Is this really a Space, with proper operations of addition and multiplication ? If we add two slopes then what does this operation mean in $\{x, y\}$ space ? It means that we add the values on the line $x = 1$. We vertically stack mountains instead of neatly fitting the angles like in a cake. In normal space this kind of stacking can occur. When we add two linear functions we take the average slope though (cf. $y = 3\ x + 1$ and $y = 2\ x - 3$). To create a Slope Space we also need a bit more, like the rule that multiplication of vectors is addition of angles. For these rules to work, the slope of y/x must for example be equivalent to twice its half slopes, for example $y/x = y / (2x) + 2y/(4x)$, that is, the point {x, y} must be producible from points $\{2x, y\}$ and $\{4x, 2y\}$. There are also the r's. This is getting complex ... For now, let us close this short exploration of other Spaces.

We draw two conclusions: (1) The concept of a Space is important, and it is great that we have both normal Euclidean Space and Polar or Angular Space. (2) We started out with lines and spent a lot of attention to the slopes of lines. It is great that we see slopes again in trigonometry. It is a disappointment though that slopes do not play the dominant role in this Space. It appears that angles take over. However, later in calculus slopes gain in importance again.

A fine example http://demonstrations.wolfram.com/TheCelestialTwoBodyProblem/

See also Palais, R. (2001b), "The Natural Cosine and Sine Curves", JOMA
http://mathdl.maa.org/mathDL/4/?pa=content&sa=viewDocument&nodeId=483

10. Linear algebra

10.1 Dimensions

The complex plane is impressive but for three or more dimensions $x + y\, i + z\, j + ...$ becomes less tractible than the earlier notation where all unit vectors are specified, thus $x\,\vec{X} + y\,\vec{Y} + z\,\vec{Z} + ...$ For then each dimension is treated the same and we prevent paradoxes like 1 = -1.

We already have touched upon 3D or 4D problems, for example the mapping of Euclidean into polar space. We will not develop n-dimensional space. We will remain in the two dimensions of the vector space that we already discussed, thus $x\,\vec{X} + y\,\vec{Y}$. What we can do however is develop the notation of linear algebra. The news is that this is an efficient notation that allows us to derive and express results in a compact manner.

People in the past have been solving systems of equations a lot. Over time they invented an efficient way of notation and solving. The following is their invention. The efficiency gain exists for 3D and higher. There is a loss for 2D, but even there we gain in conceptual transparancy since there is new light on what we are doing.

It would have been an option to introduce this notation when presenting vectors and before discussing vector multiplication that causes rotation. It was better however to first master the key theorem of analytic geometry and there the vector notation was sufficient. The imaginary plane added to our understanding. We can now collect the findings and move one step further.

10.2 Inner product

Regard the general formula for a line again: $p\,x + q\,y = r$. In this equation we see the vector $v = \{x, y\}$ but also the coefficients $w = \{p, q\}$.

We already had multiplication of vectors as rotation. Let us now define another kind of multiplication, called the *inner* product or the *dot* product. We can write the general formula for the line in a really compact manner.

Definition of *inner product*: $r = p\,x + q\,y \quad \Leftrightarrow \quad r = w\,.\,v = \{p, q\}\,.\,\{x,\ y\}$

As you see, this is just another way to write what we already know.

- Check for yourself that this holds.

```
r == {2, 3} . {3, 4}
```

$r = 18$

We denote vectors with Roman letters and numbers with Greek letters. Such numbers λ and μ are also called scalars (as scale is also a ladder). Then λv is a multiple λ of v, an expansion along a ray from the origin. Properties for the inner product are:

1. Linear along rays: $(\lambda v) . (\mu w) = (\lambda \mu)\ v.w$.
2. We can switch order ("commutative"): $v.w = w.v$.
3. Eliminating brackets in "distributive" manner: $v . (w + u) = v.w + v.u$.
   With (2) we can derive $(v + w)(u + z) = v.u + v.z + w.u + w.z$
4. For two nonzero vectors $v.w = 0$ if and only if they are perpendicular.
5. The inner product with itself gives the squared length: $v.v = |v|^2$.

- Writing r^2 instead of r it is the Pythagorean Theorem again.

```
r^2 == {a, b} . {a, b}
```

$r^2 = a^2 + b^2$

10.3 Collecting parameters

10.3.1 A system of inner products

Two lines intersect in a point. Above in §4.3.3 we had this systems of equations:

$$\begin{pmatrix} y = 4 - 2x \\ y = -5 + 7x \end{pmatrix} = \begin{pmatrix} 4 = y + 2x \\ -5 = y - 7x \end{pmatrix}$$

- Write these equations as inner products and solve them.

```
Equation[1] =   4 == {2, 1} . {x, y}
```

$4 = 2x + y$

```
Equation[2] =   -5 == {-7, 1} . {x, y}
```

$-5 = y - 7x$

```
Solve[{Equation[1], Equation[2]}, {x, y}]
```

$\{\{x \to 1,\ y \to 2\}\}$

We have solved these equations step by step before but now the objective is to understand this solution in terms of inner products.

10.3.2 Matrix of coefficients

With two equations there are two sets of parameters. Let us make that compact again. We collect all parameters in a table, called *matrix*.

Definition: A matrix is a list of lists of equal length. A matrix has rows and columns. If the number of rows is equal to the number of columns then the matrix is called square. When we have as many equations as unknowns then we get a square matrix. A matrix can contain zeros for elements that are absent.

- This is the matrix of our system of equations. We can write the rows below each other but also next to each other.

```
A == {{2, 1}, {-7, 1}}
```

$$A = \begin{pmatrix} 2 & 1 \\ -7 & 1 \end{pmatrix}$$

10.3.3 Inner product of matrix and vector

With $v = \{x, y\}$ for the unknown variables take $w = \{4, -5\}$ for the outcome. Our new definition for compact writing is:

Definition: The inner product for vectors can be extended for a matrix A and a vector v such that a new vector is created as $w = A \,.\, v$. The condition is that each row of matrix A has the same length as the vector v, so that there is a proper inner product for each row.

- This product of matrix and vector generates another vector.

```
{4, -5} == {{2, 1}, {-7, 1}} . {x, y}
```

$$\{4, -5\} = \{2\,x + y,\ y - 7\,x\}$$

- Writing the equality for each element.

```
Thread[Result]
```

$$\{4 = 2\,x + y,\ -5 = y - 7\,x\}$$

Writing $w = A \,.\, v$ gives a compact form for a system of equations. It separates the variables from the coefficients. The variables are not that interesting. We can work directly on the coefficients.

10.4 Solving by means of the inverse

10.4.1 Inverse matrix or matrix inverse

We know that there is a solution for the unknowns, namely $v = \{x, y\} = \{1, 2\}$. A major step is now to presume that alongside to matrix A there is a matrix B that gives that very solution.

Definition: If $w = A \,.\, v$ and if there is a B such that $B \,.\, w = v$, and if this holds for arbitrary v and w, then B is called the *matrix inverse* of A. In that case $v = B \,.\, w$ solves $w = A \,.\, v$.

- Matrix A and its inverse B.

$$A = \begin{pmatrix} 2 & 1 \\ -7 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} \frac{1}{9} & -\frac{1}{9} \\ \frac{7}{9} & \frac{2}{9} \end{pmatrix}$$

- Application to our problem: $w = A \,.\, v$ solves into $v = B\, w$.

```
{4, −5} == {{2, 1}, {−7, 1}} . {x, y};
```

$$\{x, y\} = \text{Hold}\left[\begin{pmatrix} \frac{1}{9} & -\frac{1}{9} \\ \frac{7}{9} & \frac{2}{9} \end{pmatrix} . \{4, -5\}\right]$$

```
Result // ReleaseHold
```

$$\{x, y\} = \{1, 2\}$$

Another value of {4, -5} would give another solution for $\{x, y\}$ without the need for additional solving of equations. The remainder of our discussion is on understanding the theory and on how to find the inverse.

10.4.2 Constructing the inverse

How to find the inverse ? There are three steps:

1. We calculate the so-called “determinant” or Det[A] = $|A|$. If Det = 0 then the equations are dependent, the lines overlap or are inconsistent. For example x = 1 and x = 5 do not have a point of intersection. Then there is no solution, and we stop: see the next subsection). If Det ≠ 0 then we proceed.
2. We switch the coefficients in the matrix to create the “adjoint” matrix, or Adj[A].
3. We divide the latter by the determinant. The inverse then is Adj[A] / $|A|$.

§10.7 below gives a geometric display of the determinant. It is the surface of the parallelepiped generated by the rows of the matrix. Like a vector has a size measure called length, two vectors have a size measure called surface. That geometric display fits better there than here, but if you are interested you can already look there. If the surface is zero then we obviously cannot divide by it.

- Let us take a general matrix.

```
mat = {{a, b}, {c, d}}
```

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

- Step 1: The determinant. It uses the cross-product of the elements of the matrix.

```
det == Det[mat]
```

$$\text{det} = a\,d - b\,c$$

- Step 2: The adjoint. We trade places for *a* and *d,* and give opposite signs to *b* and *c.*

```
adj == Adjoint[mat]
```

$$\text{adj} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

- Step 3: The inverse. Divide adjoint by determinant.

```
adj / det == Inverse[mat]
```

$$\frac{\text{adj}}{\text{det}} = \begin{pmatrix} \frac{d}{a\,d-b\,c} & -\frac{b}{a\,d-b\,c} \\ -\frac{c}{a\,d-b\,c} & \frac{a}{a\,d-b\,c} \end{pmatrix}$$

- For our example, check that Det[*A*] = 2 * 1 - 1 * (-7) = 9

```
mat = {{2, 1}, {-7, 1}};
See[Det[mat], Adjoint[mat], Inverse[mat]]
```

$$9 \qquad \begin{pmatrix} 1 & -1 \\ 7 & 2 \end{pmatrix} \qquad \begin{pmatrix} \frac{1}{9} & -\frac{1}{9} \\ \frac{7}{9} & \frac{2}{9} \end{pmatrix}$$

10.4.3 Solving linear equations

As said, our solution was:

```
{x, y} == Inverse[{{2, 1}, {-7, 1}}] . {4, -5}
```

$$\{x, y\} = \{1, 2\}$$

Mathematica has a specific routine for linear equations with square matrices.

- Taking the inverse and multiplying with the known vector in a single step.

```
LinearSolve[{{2, 1}, {-7, 1}}, {4, -5}]
```

{1, 2}

Nonsquare matrices arise in two cases.

1. There are not enough equations, e.g. when the determinant is zero and there appears a dependence so that equations must be eliminated. Then we can solve for only part of the unknowns. The variables give a solution space.
2. There are too many equations. If there is a contradiction then we are stuck. Otherwise we drop those that must be overlapping and proceed with the square problem.

These cases can be handled by matrix algebra. That leads too far for our purposes though. In those cases we can use the general Solve routine.

- Let us solve the first equation for x. It appears to be a linear function y. If there is more information then a precise solution may be found.

```
Solve[y == 4 - 2 x, {x}]
```

$$\left\{\left\{x \to \frac{4-y}{2}\right\}\right\}$$

10.5 Inner product for matrices

We proceed with the theory of linear algebra.

10.5.1 Identity matrix

Property: If $w = A \, . \, v$ and $B \, . \, w = v$ then actually $v = B \, . \, (A \, . \, v)$ and $w = A \, . \, (B \, . \, w)$.

Till now we did not apply the dot product to two matrices only. Let us see whether we can extend the definition to two matrices. Steps are:

- Let us rewrite the brackets $v = B \, . \, (A \, . \, v)$ into $v = (B \, . \, A) \, . \, v$.
- Let $(B \, . \, A) = I$ for some $I = \{\{a, b\}, \{c, d\}\}$.
- Then $v = I \, . \, v$.
- This must hold for any v. Let us try $v = \{1, 0\}$ and $\{0, 1\}$.
- Using $v = \{1, 0\}$.

```
{1, 0} == {{a, b}, {c, d}}.{1, 0}
```

$\{1, 0\} = \{a, c\}$

```
Thread[Result]
```

$\{1 = a,\ 0 = c\}$

- Using $v = \{0, 1\}$.

```
{0, 1} == {{a, b}, {c, d}}.{0, 1}
```

$\{0, 1\} = \{b, d\}$

```
Thread[Result]
```

$\{0 = b,\ 1 = d\}$

- Thus - and check that for $v = \{x, y\}$ that $v = I\ .\ v$.

```
IdentityMatrix[2]
```

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

We can do the same for $A\ .\ B$.

Definition: If an inverse B exists then $B\ .\ A = A\ .\ B = I$, called the *identity matrix*.

```
mat = {{2, 1}, {-7, 1}};   Inverse[mat] . mat
```

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Definition: The inverse of matrix A will be written A^{-1}. Thus $A^{-1}\ .\ A = A\ .\ A^{-1} = I$. For the determinant we find $|I| = 1$. NB. We can now use the symbol B for any matrix again and not just the inverse of A.

10.5.2 Matrix product

What will be a practical definition for the dot product for matrices $B\ .\ A$? Consider the rows and columns separately and write B = {BRow1, BRow2} and A = {AColumn1, AColumn2}. There is a system of four inner products, namely each a row of B and a column of A.

$$\begin{pmatrix} \text{BRow1.AColumn1} & \text{BRow1.AColumn2} \\ \text{BRow2.AColumn1} & \text{BRow2.AColumn2} \end{pmatrix}$$

Lower case letters read a bit better than upper cases now, so we switch from A and B to a and b. In finer detail, consider all elements and write $a_{i,j} = a[i, j]$ for the element in matrix a in row i and column j. Writing the dot products of the rows and columns out in detail:

$$\left\{a = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix},\ b = \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix}\right\}$$

$$a.b = \begin{pmatrix} a_{1,1}\, b_{1,1} + a_{1,2}\, b_{2,1} & a_{1,1}\, b_{1,2} + a_{1,2}\, b_{2,2} \\ a_{2,1}\, b_{1,1} + a_{2,2}\, b_{2,1} & a_{2,1}\, b_{1,2} + a_{2,2}\, b_{2,2} \end{pmatrix}$$

- Check that the dot product in this manner gives the identity matrix.

$$\text{Hold}\left[\begin{pmatrix} \frac{1}{9} & -\frac{1}{9} \\ \frac{7}{9} & \frac{2}{9} \end{pmatrix} . \begin{pmatrix} 2 & 1 \\ -7 & 1 \end{pmatrix}\right]$$

- The dot product of a matrix and its adjoint gives the determinant on the diagonal. Clearly, when we divide this by the determinant we get the identity matrix.

$$\begin{pmatrix} a\, d - b\, c & 0 \\ 0 & a\, d - b\, c \end{pmatrix} = \text{Hold}\left[\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} . \begin{pmatrix} a & b \\ c & d \end{pmatrix}\right]$$

PM. Again, see §10.7 below for a geometric explanation of the determinant.

10.5.3 Properties of determinants

We give two properties without proof, that you however could check for our two-dimensional matrices.

Without proof we take $|A \,.\, C\,| = |A|\ |C\,|$ and hence $|A^{-1}| = 1 \,/\, |A|$.

Without proof we take $|r\ A\ | = r^n\ |A|$ where n is the size of the square matrix, or the number of rows or columns. Thus $|\text{Adj}[A]| = |A^{-1} * |A|\ | = |A|^n\ |A^{-1}| = |A|^n / |A| = |A|^{n-1}$.

Definition: The transposed or dual matrix arises when we interchange rows and columns. Then the determinant does not change.

$$\left\{\begin{pmatrix} a & b \\ c & d \end{pmatrix} \to \text{Transpose} \to \begin{pmatrix} a & c \\ b & d \end{pmatrix}, \text{Det} \to a\, d - b\, c\right\}$$

10.6 An algorithm for n = 2

For $n = 2$ the methods of §4.3.2 and 4.3.3 seem most efficient but it is instructive to see how matrix algebra would work.

For $n = 2$ the adjoint has determinant $|A|$. Indeed, the adjoint in our example has determinant 9 too. Thus it is an option to directly write the adjoint matrix from the system of 2 equations and use that for the determinant. The calculation $v = A^{-1} \,.\, w$ can be written as $u = \text{Adj}[A] \,.\, w \,/\, |A|$. If we use the inverse matrix only once for the solution at a single point then it is faster to do $\text{Adj}[A] \,.\, w$ and only then divide by $|\text{Adj}[A]|$ (for $n = 2$) since this divides only two numbers instead of four.

Let us take the problem of §4.3.2:

$$\begin{pmatrix} 5 = c + 10\,s \\ -3 = c + 4\,s \end{pmatrix}$$

```
adj = Adjoint[{{10, 1}, {4, 1}}]
```

$$\begin{pmatrix} 1 & -1 \\ -4 & 10 \end{pmatrix}$$

```
See[temp = adj . {5, −3}, det = Det[adj]]
```

{8, −50}   6

```
res = temp / det
```

$$\left\{\frac{4}{3}, -\frac{25}{3}\right\}$$

Compare this to §4.3.2. Counting the number of steps then the simple method there scores better (where all steps have equal effort on a pocket calculator).

	Plus	Minus	Times	Divide	Total
Δy / Δx		2		1	3
yA − s ∗ xA		1	1		2
Total	0	3	1	1	5

	Plus	Minus	Times	Divide	Total
Adjoint			2 signs		2
Det		1	2		3
adj ∗ coefs	2		4		6
Final				2	2
Total	2	1	8	2	13

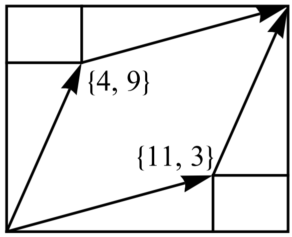
A small advantage for Chapter 4 is that *c* has coefficient 1 so that the equations are easy to subtract. This advantage however also arises for the determinant for mental calculation. You will be more comfortable with the step by step approach as long as you do not have experience with matrix algebra. When working in 3D and higher, computers quickly take over so that much does not matter.

10.7 The geometry of the determinant

10.7.1 Area of a parallepiped

The matrix {{*a*, *b*}, {*c*, *d*}} contains two row vectors {*a*, *b*} and {*c*, *d*} that together span a parallelepiped. When we draw a diagram of this, we find that the parallelepiped is contained in a rectangle with sides (*a* + *c*) and (*b* + *d*) which are the column sums of the matrix.

- TwoVectorsPlot essentially gives the same plot but here we put a box around it.



The total area of the rectangle is given by $(a + c)(b + d)$ while the area of the parallelepiped can be found by subtraction of the triangles and small rectangles, thus $(a + c)(b + d) - 2\ b\ c - 2 * (\frac{1}{2} a\ b) - 2 * (\frac{1}{2} c\ d) = a\ d - b\ c$. This latter value is the determinant of the matrix.

$$(a + c)(b + d) - 2\,b\,c - 2\left(\frac{1}{2}\,a\,b\right) - 2\left(\frac{1}{2}\,c\,d\right) \;\text{// Simplify}$$

$a\,d - b\,c$

Det[{{*a*, *b*}, {*c*, *d*}}]

$a\,d - b\,c$

`ShowDet[]`	plots defaults {11, 3} and {4, 9}
`ShowDet[ (Label,) {{a, b}, {x, y}}]`	plots the two rows as vectors. Label for a better graph

10.7.2 Kinds of data

How we interprete the geometry depends very much upon the kind of data we have. When we measure the size of a window then we understand what it means when we say that it is too small and must be made 10% larger. When we have categorized people according to their religion or political party then it is dubious to say that catholics are halve protestants, or that some party is the “average”. These kinds of measurement scales arise:

1. Nominal scale: just names, labels, categories or bins. For example religion. Mathematically we use only equality and counts.
2. Ordinal scale: there is some ordering or ranking. For example exam results (A to F). Cumulative counts have a meaning.
3. Interval scale: differences have meaning but there is no natural zero value. For example degrees Celsius, created by equal distances on a thermometer. We cannot say that 10 degrees is twice as warm as 5 degrees, but the difference of 5 degrees is the same length at 10 or 30 degrees. Mathematically we can use

addition.

4. Ratio scale: ratios have meaning here. Length, time, mass, angle. Degrees Kelvin, since 0 degree Kelvin has absolute meaning. A unit of measurement is taken from the phenomenon itself. Mathematically we can also use multiplication.

With some creativity we can try to work around these categories of measurement scales. A notable example is the Arpad Elo rating in chess, the same system found independently by Georg Rasch for tests on reading skills. In chess we see winners and losers, and it seems that we can only rank them in their scores. However, when players of a different strength meet then we may calculate a probability of winning. The ratings can be updated after the outcome. In that way the ranking can be transformed into a rating, an interval scale. The same approach allows comparing competence or skill with challenge. If the skill is too large then there is boredom and if the challenge is too large there is stress. There is "flow" if there is adequate match. The difference in skill and challenge generates the probability of succeeding on a test, which again generates an interval scale.

We now consider measures of association. When we have vectors $\{a, b\}$ and $\{c, d\}$, how can we compare them ? With nominal data we may still use the determinant, with a ratio scale we may do more.

10.7.3 A measure of association

A statistician asks 100 men and 100 women whether they are on a diet or not. The results for these nominal data can be put in a 2 by 2 table. This kind of table is called a contingency table. In a test the effect is put in the rows. The subjects carry the cause and are put in the columns. The test is whether the subjects have a disease or not (in this case whether they have a diet). Contingency tables generally are presented with table-headings and border-sums.

- The relevant data are in the core and generate the border sums.

$$\begin{pmatrix} a & b & a+b \\ c & d & c+d \\ a+c & b+d & a+b+c+d \end{pmatrix}$$

- Assumed frequencies of men and women dieting or not. The behaviour of the groups differs. Can we express the degree of difference in a measure ?

	Men	¬ Men	Tested
Dieting	20	64	84
¬ Dieting	80	36	116
Sum	100	100	200

We can use the determinant as a measure of association.

- When we take the ratio of the areas $cr = (a\ d - b\ c) / ((a + c)(b + d))$ then we find a number between -1 and 1.
- Row sums differ from column sums. A determinant $a\ d - b\ c$ holds for the dual (transposed) matrix too, giving a row ratio *rr*.
- Since there are two ways of looking at the matrix a more robust measure is the geometric average $\sqrt{cr * rr}$. The numerator remains $a\ d - b\ c$ but the denominator becomes $\sqrt{((a + c)(b + d)(a + b)(c + d))}$. This gives us a "standardized volume (surface) ratio".

$$\frac{a\,d - b\,c}{\sqrt{(a + b)\,(a + c)\,(b + d)\,(c + d)}}$$

```
Result /. Thread[{a, b, c, d} -> {20, 64, 80, 36.}]
```

−0.445742

A diagonal matrix with $b = c = 0$ gives outcome +1 and with $a = d = 0$ gives outcome -1. Nominal data have no natural order, but one cannot avoid an order of presentation and the sign of the association measure reflects that. In this example there is a negative association meaning that the rising diagonal gets relatively more weight. Men incline to not-dieting, women incline towards dieting. So much was obvious, but if we now have other data we can compare the inclinations.

10.7.4 Statistical independence means zero association

Let p be the fraction of men in the total number of observations. (Here p = 50%.) Let t be the fraction of all observations that satisfies the test. (Here t = 42%.) When the distribution of men over those who satisfy the test and those who do not is the same as the overall distribution, then this must necessarily also hold for the women. Then the test (habit of dieting) is no different in the two subgroups. In that case the variables sex and eating habit are called statistically independent and there is zero association between the variables.

Thus *algebraic dependence* means that our measure of association shows zero association, and means that there is *statistical independence*. The statistical independence causes the algebraic dependence.

- We construct a table with statistical independence and verify the 0.

```
mat4 = PrTable[t, p] * N;
Simplify[CorrelationPr2By2[mat4], Assumptions → {t ≥ 0, p ≥ 0}]
```

0

	Men	¬ Men	Tested
Dieting	$N\,p\,t$	$-N\,(p-1)\,t$	$N\,t$
¬ Dieting	$-N\,p\,(t-1)$	$N\,(p-1)\,(t-1)$	$N - N\,t$
Sum	$N\,p$	$N - N\,p$	N

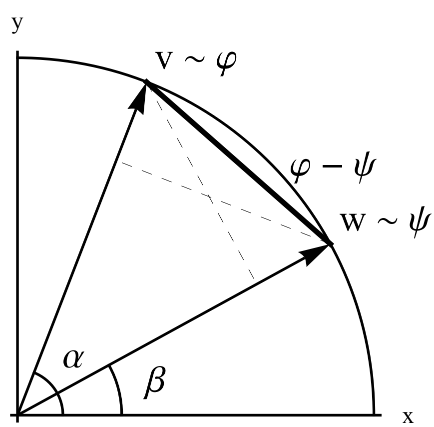
`CorrelationPr2By2[` *`matrix`*`]`	gives the measure of correlation for a contingency table of two binary nominal variables (correlation and not just association)

Let Cause be the column variable and Effect the row variable. In logic, the variables take values {1, 0}. Here it is better to take {1, -1} so that equal numbers of observations give a zero mean. Output then is the normal Pearson CorrelationPr[{1, 1, -1, -1}, {1, -1, 1, -1}, {n11, n21, n12, n22}].

10.8 The geometry of the inner product

Let vectors v and w not necessarily on the unit circle be associated with angles α and β and arcs φ and ψ. Three dots give a plane. Thus with the origin and two vectors from the origin we have a plane and we can apply plane geometry.

- Let $u = v - w$ and check that we have a triangle with those three.
- The xur or cosine rule gives $|u|^2 = |v|^2 + |w|^2 - 2\,|v|\,|w|\,\mathbb{X}[\alpha - \beta]$
- $u.u = (v - w) . (v - w) = v.v + w.w - 2\,v.w$ or $|u|^2 = |v|^2 + |w|^2 - 2\,v.w$
- Elimination gives $v.w = |v|\,|w|\,\mathbb{X}[\alpha - \beta]$
- $\mathbb{X}[\alpha - \beta] = v . w / (|v|\,|w|)$
- or the inner product of the vectors normalized to the unit circle is the xur of the angle between the vectors, also seen as the projection of either vector onto the other.



- The inner product of the normalized vectors equals the x_{ur} or cosine of the angle between these vectors. Use XandY or otherwise explicit:

```
{11, 3}.{4, 9} / (NRadius[{11, 3}] NRadius[{4, 9}]) // N
```

0.632267

With $v = \{x, y\}$ and $w = \{a, b\}$ we thus get two values for the x_{ur} or cosine: either from $\mathbb{X}_{\alpha+\beta}$ when the vector product $a\ x\ -\ b\ y$ applies (i.e. the product in the complex plane), or as $\mathbb{X}_{\alpha-\beta}$ when the inner product $a\ x + b\ y$ applies (see §7.8.7). Addition of angles means subtraction of the y-terms, subtraction of angles means addition with the y-terms.

When writing this book it was conceivable to start the discussion with linear algebra and the inner product and then show that it means projection and subtraction of angles. The choice was made to start with vector products and show that these are the addition of angles, since this links up a bit better with the complex plane and the analogy with (recovered) exponents. Materially it should not matter since subtraction is addition of a negative. The inner product arises as the multiplication with the complex conjugate.

10.9 The geometry of correlation

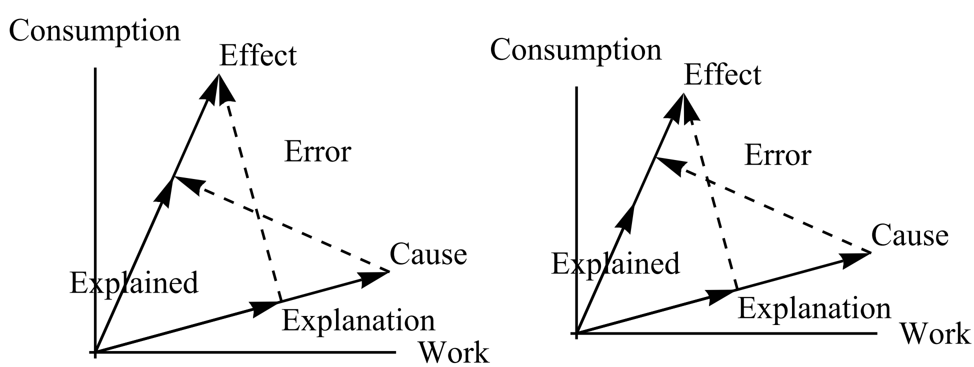
10.9.1 Correlation is not cause

Let us take the $savings = h\ w\ -\ p\ q$ model where income is hours times wage and spending is price times quantity. Let our wage be \$1 per hour and let us plan to work 11 hours and consume \$3 so that anticipated savings are \$8. Winning a lottery makes us change behaviour, we work only 4 hours and our consumption rises to \$9. A statistician does not know about that lottery and has to make do with what is observed. Statistically there is a record of plans {11, 3} and effects {4, 9}. A hypothesis is that the plans are causal factors for behavioural outcome. The statistician calculates a correlation coefficient between plans and realizations of 63%. The variation in the effects is explained by 63% by the variation in the causes but there is still is some sizeable error due to unknown factors.

The method followed is called regression:

- The effect is split in an explained part and an error.
- This split is achieved by a perpendicular projection P of effect on presumed cause. The projection is interpreted as the *explanation* by the cause. The cause contributes by a factor μ = |Explanation| / |Cause| using the absolute lengths of the vectors. Thus $P = \mu$ Cause.

- The vector from *P* to the effect is the error. The *effect* is a vector addition of the *explanation* and the *error*. The perpendicularity means that all error influences are fully independent of the explanation along the line of the cause.
- The ratio of the explanation to the effect is called the *correlation coefficient*. It ranges between -1 and 1. Geometrically it is the Xur (cosine) of the angle between the vectors of cause and effect. A value of 1 means that cause and effect overlap and there is a perfect explanation. The squared value of the correlation coefficient is called the coefficient of determination, and it gives the share of the variation that has been explained.
- The error ratio is the Yur. The squares of correlation and error ratio add up to 1.
- When the vectors of cause and effect are normalized to the unit circle then the explained part of the effect and the explanation by the cause are just as long, and both give the Xur value of the angle inbetween. When the vectors are not normalized then the Xur is only given by the said projection *P* of effect on presumed cause.
- The correlation coefficient between plans and realization is 63%. The left is normalized to the unit circle, the right is not normalized.



	{Cause → {11., 3.}, Effect → {4., 9.}}
|Effect|	9.849
|Explanation|	6.227
|Cause|	11.402
|Error| = |Effect − μ Cause|	7.63
μ = |Explanation| / |Cause|	0.546
R = Correlation = |Explanation| / |Effect|	0.632
ErrorRatio = |Error| / |Effect|	0.775
R^2 = Correlation^2	0.4
Correlation^2 + ErrorRatio^2	1.

Cause and effect are determined by the model that we make. The notion of *model* is key here. It is hard if not impossible to know reality but we can deal with models. What we see as a cause depends upon our perception of what is happening, thus the model that we design. Correlation is no cause. A model of cause and effect suggests to us to calculate a correlation coefficient but that is only a part in testing and modelling. A low correlation may cause us to reject a hypothesis but may also confirm a suspicion and cause a search for confounders and reasons for the errors.

- The inner product of the normalized vectors equals the x_{ur} or cosine, thus $v\,.\,w\,/\,(|v|\;\;|w|)$ = $\mathbb{X}[\gamma]$ for the angle γ between the two vectors. Use XandY or normalize explicitly:

```
{11, 3}.{4, 9} / (NRadius[{11, 3}] NRadius[{4, 9}]) // N
```

0.632267

NB. For vectors of length > 2 we take the difference from the mean first, thus v - Mean[v].

`ProjectionPlot[` {a, b}, {x, y}`]`	plots with cause {a, b} and effect {x, y}. 2 D only, defaults {11, 3} and {4, 9}
`ProjectionPlot[Table,` {a, b}, {x, y}`]`	gives the explanatory table

10.9.2 Two points give a line

The two points {11, 3} and {4, 9} that we used on planned and realized work and consumption may also be put on a line.

- The routine that we used before.

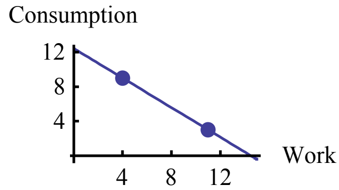
```
line = TwoPointsToLine[x, {11, 3}, {4, 9}]
```

$$\frac{87}{7} - \frac{6\,x}{7}$$

- Linear estimation in *Mathematica* gives the same. Each hour of work reduces consumption by $6/7, or each hour of leisure increases it.

```
model = LinearModelFit[data = {{11, 3}, {4, 9}}, x, x] // Rationalize
```

FittedModel$\left[\boxed{\frac{87}{7} - \frac{6\,x}{7}}\right]$



This line is a weak model. We presented this only for the following reasons:

- To introduce the notion of linear estimation.
- The linear estimation routine can be used to find a line trough two points. The estimation error then is zero. The line perfectly links up the two points. Seen in this way, the correlation coefficient is 100%. When we add more points then we really start estimation.
- 100% differs from the 63% that we saw earlier. Presently we regress consumption on hours worked, earlier we looked at plans versus realizations: different cases.

The issue that we are discussing is clearly complex and multidimensional. We have four categories {hours worked, consumption} × {plan, realization}, and actually there are also the wage, price and quantity consumed, savings, and the lottery surprise. This is too much for present purposes.

Nevertheless, it is useful to have a geometric interpretation of linear regression and thus we simplify. In the following we assume a list of observations for five people, assuming that not all can win a lottery, and with recorded hours and dollars spent on consumption. These can also be hours of homework and level of pocket money, though hopefully there is no correlation there.

10.9.3 More observations

Assume that five people provide us with their data on hours worked and dollars spent on consumption on a particular day (actual and not plans). We sort the data on the hours worked (just to have them in neat order) and then fit them, with the consumption explained from the hours. It appears that only about 25% of the variation in the data is explained.

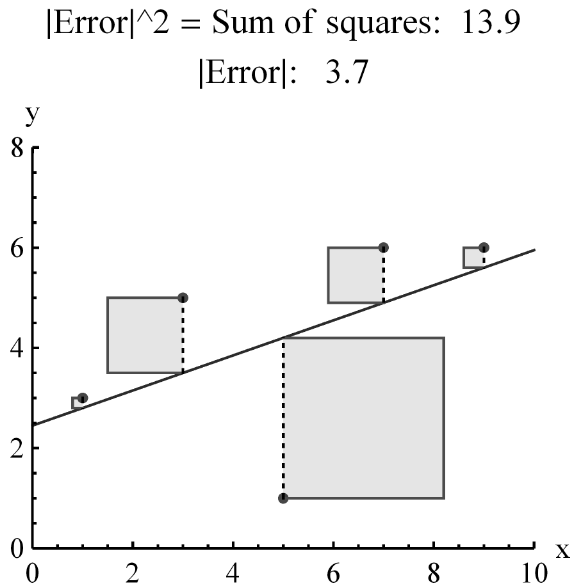
```
data = {{1, 3}, {3, 5}, {5, 1}, {7, 6}, {9, 6}}; model = LinearModelFit[ data, x, x]
```

FittedModel[$\boxed{0.35\,x + 2.45}$]

```
{r2 = model["RSquared"], Correlation → Sqrt[r2]}
```

{0.260638, Correlation → 0.510527}

In a graph we put both the dots of the observations and the estimated line. The vertical distance of a dot to the line is the estimation error e_i for that point of observation. The squared error is the square created by that distance. When we do this for all dots and we add the areas of the squares then we get the "Sum of Squared Errors", SSE = $e_1^2 + \ldots + e_n^2$. Taking the root gives us |Error|. Each possible estimated line gives squares of different sizes. The minimal value of these possible SSE gives our best estimate. This method is called "the least squares method". There is least error when the sum of squares is minimal.



See: http://demonstrations.wolfram.com/LeastSquaresCriteriaForTheLeastSquaresRegressionLine/

The least squares method follows the approach displayed in the cause and effect diagram. The minimal SSE generates an explanation as the perpendicular projection of the effect on the cause. The diagram there thus applies also for data vectors larger than 2 elements. For correlation, though: take the vector differences from their mean values.

10.9.4 The stage has been set

This concludes our present discussion of linear algebra. Discussion of regression was useful not only because of this interpretation of squared errors (the distance measure) but also because of the notion of projection and the insight that x_{ur} or cosine has more uses than in measuring angles. We have also touched upon the notion of more dimensions. Our notation of vectors and matrices would allow us to handle them, but to actually do it: that is another book. We now proceed with calculus. Later we will return to this example of estimation and apply calculus to find the minimal error.

Part IV. Calculus

We consider these types of functions:

1. Polynomials contain only powers of x thus $y = c + s\ x + a\ x^2 + b\ x^3 + \ldots$
2. Exponential functions have x in the exponent such as $y = base^{c + s\,x + \ldots}$. Their inverses are recovered exponential functions, rex[*base*, *y*].
3. Trigonometric are x_{ur} and y_{ur} depending upon an angle α.
4. Two-dimensional is $z = f[x, y]$. If x or y is held constant then it is 1D again.

The new topic of discussion is the surface between the horizontal axis and the function values. Geometry is about measurement, after all.

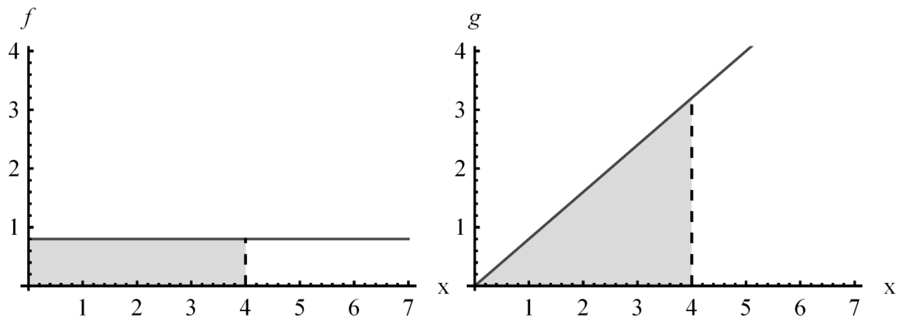
11. Polynomial

11.1 Measurement of surface

11.1.1 Rectangle and triangle

Calculus concerns the measurement of surface between a function and the horizontal axis. A key aspect is also the change in surface. The basic cases of rectangle and triangle are in the following graphs.

- The surfaces under $f[x] = 0.8$ and $g[x] = 0.8\,x$, for x over the interval [0, 4].



The surfaces under f and g for x over [0, 4] are easily calculated. The constant function f gives a rectangle 0.8 * 4 = 3.2. The ray function g gives a triangle ½ $h\,w$ = ½ (0.8 * 4) * 4 = 6.4. Let us make it more formal.

Original function	$f[x] = c$	$g[x] = s\,x$
$c = s = 0.8,\ x = 4$	0.8	3.2
Surface function	$Sur[f, x] = c\,x$	$Sur[g, x] = ½\,(s\,x)\,x = ½\,s\,x^2$
$c = s = 0.8,\ x = 4$	3.2	6.4

In this table we see the value 3.2 in two places. Scrutinizing the formulas we observe this equality for the height of the triangle h:

$$\text{when } c = s \text{ then } h = g[x] = s\,x \overset{!}{=} \mathrm{Sur}[f, x] = c\,x$$

There *exist* functions such that $c = s$. We used an example 0.8 but there are others. Our distinction between slope and constant is a bit too strict: the slope of one function (g) can be the constant of another (f). When the domain [0, 4] changes then the relation remains. Given that f and g are related with this choice of parameter $c = s$ we can express the dependence in their names and formal expression. The following definitions are useful:

- Since g gives the surface for f we call g the "primitive" and f the "derivative".
- When g is the primitive then we use a prime ′ to indicate the derivative: $f = g'$.
- It will also be clearer to use an upper case letter for a primitive function, thus F and $f = F'$. This cannot be maintained forever but it is a good convention.

The key idea is - let us not call it a theorem just yet, since we ought to develop the theory first:

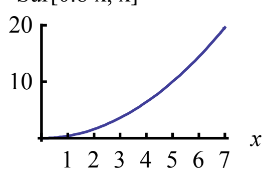
The primitive is $F[x] = s\, x = Sur[F', x]$ iff the derivative is $F'[x] = s$. (*idea*)

When we apply this idea to g and the surface for g itself then we get:

The primitive is $F[x] = ½\ s\, x^2 = Sur[F', x]$ iff the derivative is $F'[x] = s\, x$. (*idea*)

- The surface primitive for g is found to be $Sur[g, x] = ½ * 0.8 * x^2$

G[x] = Sur[0.8 x, x]



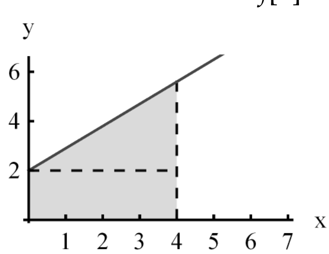
11.1.2 Triangle stacked upon a rectangle

The function $f[x] = c + s\, x$ gives a triangle stacked upon a rectangle. The triangle has $h = s\, x = f[x] - c$. Total surface is $Sur[f, x] = c\, x + ½\, (s\, x)\, x = c\, x + ½\, s\, x^2$.

The primitive is $F[x] = c\, x + ½\, s\, x^2 = Sur[F', x]$

iff the derivative is $F'[x] = c + s\, x$. (*idea*)

- The surfaces under $f[x] = 2 + 0.9\, x$, for x in [0, 4]



When functions are added then we can add their surfaces.

11.2 Distance, speed and acceleration

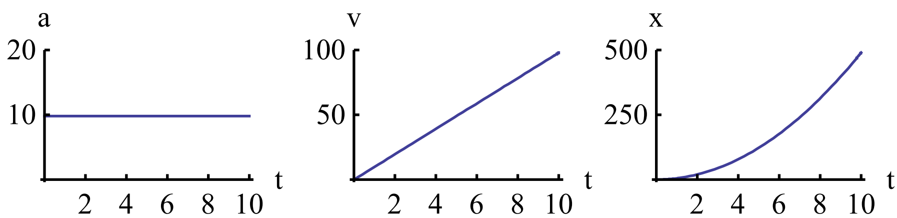
We are quite familiar with the notion of surface, so we already have a good idea what these primitives and derivatives mean. In physics we however see a new and more compelling application and interpretation.

11.2.1 Basic physics

When an object starts from position $x[0]$ with a speed $v[0]$ and gets a constant acceleration $a[t] = a$ then the same formulas and graphs arise as in the former subsections. Constant acceleration a is defined such that each unit increase of time t causes an increase in speed with a units. This is purely a definition.

$$v[t] = v[0] + a\ t$$

An observation in physics is that gravity and acceleration for small objects close to Earth can be taken constant. Take gravity with a = 9.8 [m / s^2, in meters and seconds]. Let an object be in rest and then fall. With no air friction it falls 490 meters in 10 seconds. Its speed at that moment is 98 [m / s], or 352.8 [km / h].



From $x[0] = v[0] = 0$: (a) constant acceleration, (v) $v = a\ t$, (x) the position of the object.

How do we get the quadratic expression for the position of the object ? At time t the object has moved with an average speed of $(v[0] + v[t]) / 2 = v[0] + ½\ a\ t$. The distance covered will be this average speed times the time lapsed:

$$x[t] = x[0] + (v[0] + ½\ a\ t)\ t$$

$$= x[0] + v[0]\ t + ½\ a\ t^2$$

These are precisely the kind of formulas that we looked at earlier. The model is in physics but the description is in terms of surfaces. Why a surface ? We now spot the rectangle in $a * t$. Also "distance = (average speed) * time" involves a product, and gives a surface as well. This was a bit difficult to spot when we only focus on average speed and instantaneous speed at t. Nicole d'Oresme (1323 - 1382) discovered that this product can be seen as a surface.

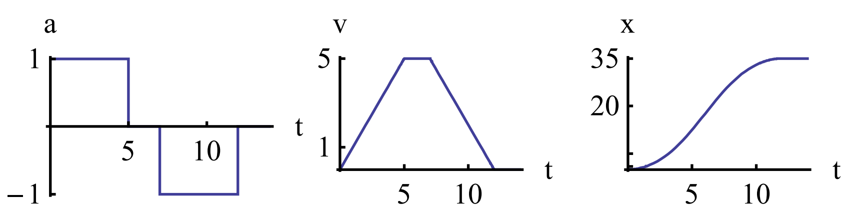
Concept	*Symbol*	*Dimension*	*Description*
Time	t	s	clicks of radioactive decay
Place	s	m	co – ordinates, Latin : situs
Distance	x	m	distance between two points; for co – ordinates $s[t]$: the distance is $\lvert s[t] - s[0] \rvert$ (Pythagoras)
Speed	v	m/s	$x / t = \lvert s[t] - s[0] \rvert / t$ is average speed, see later for instantaneous speed
Acceleration	a	$\mathrm{m} / \mathrm{s}^2$	$(v[t] - v[0]) / t$ is average acceleration, see later for instantaneous acceleration

The unit of measurement of location and distance is the meter; of time the second. To distinguish s from s it is a convention to put the dimension within [ ] brackets. Variables are in italics too.

In these measurements we calibrate to a moment $t = 0$ and conveniently take $s[0] = x[0] = 0$. What we actually use are the differences in time and place, Δt and $x = \Delta s$ so that average speed over an interval is $v = \Delta s / \Delta t$.

11.2.2 An elevator

Acceleration, speed and location of an elevator are a nice example. For five seconds it has an acceleration of 1 m/s^2. Then the acceleration drops to zero, and the elevator moves at constant speed of 18 km/h for 2 seconds. Then the elevator starts braking at -1 m/s^2. In a total of 12 seconds it moves 35 meters up.



11.2.3 Vectors

Above displays are one-dimensional. The elevator goes up and down only. These equations for acceleration, motion and place reflect a motion in one direction only. A soccerball would hopefully have at least two dimensions: not only up and down but also some distance covered.

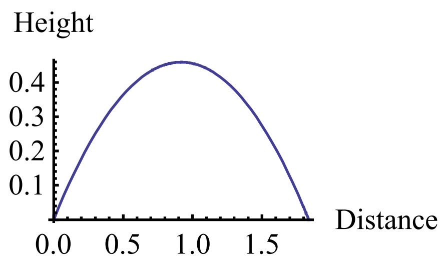
When a ball is kicked, the sideways movement will remain constant (neglecting air friction and other players). The vertical movement will be determined by gravity. The vertical movement does not determine the distance covered, for that is done by the movement sideways. The vertical movement only determines the time available, till the ball drops on the ground and is controlled by friction again. What happens is that all variables depend upon time, but that time as a variable can be eliminated, so that distance can be seen as a function of height, or in a standard form that height is a function of distance.

	speed	*distance*
vertical	v	h
horizontal	w	d

- Kick space has vertical (h and v) and horizontal (d and w) dimensions. Elimination of time t gives height h as a function of distance d.

Solve[{h == v t − 1/2 g t², d == w t}, h, t]

$$\left\{\left\{h \to \frac{2\,d\,v\,w - d^2\,g}{2\,w^2}\right\}\right\}$$



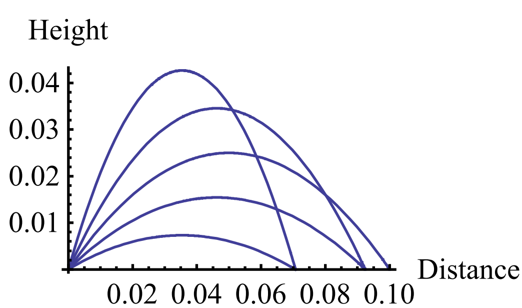
- Time t is determined by the vertical speed v and distance from w.

Solve[0 == v t − 1/2 g t², t]

$$\left\{\{t \to 0\}, \left\{t \to \frac{2\,v}{g}\right\}\right\}$$

There would actually be only one true velocity but at some angle: polar space. Vertical and horizontal speeds are projections onto the axes of Euclidean space. If we assume a proper single velocity of 1 then the horizontal speed would be given by x_{ur} and the vertical speed by y_{ur}. Maximal distance is achieved by a tangent of 1, as we can check for values of α around 1/8.

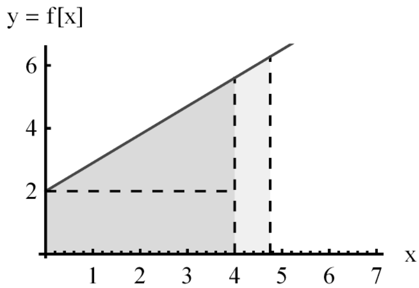
Show[Table[Height2D[Plot, Xur[α], Yur[α], 10], {α, 1/16, 3/16, 1/32}], PlotRange → All]

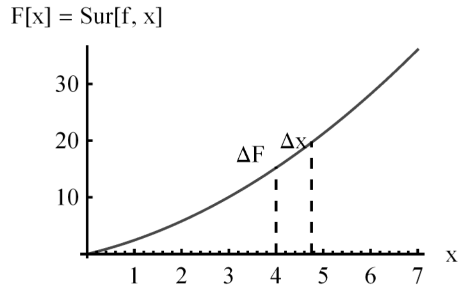


11.3 The core theory of calculus

11.3.1 A surface function and its derivative

We found a relationship between surface and derivative for two basic functions and a next question is whether we can generalize this. What about x^2 and its $Sur[x^2, x]$ function ? How does that function look like and is it easy to work with ? But x^2 is curved and the issue looks rather difficult. Well, we have two cases of a relation between a primitive and a derivative: let us see whether we can discern some pattern.





11.3.2 Stepwise development of an algorithm

In these steps we presume F known and f to be unknown. We keep an eye on what we already derived but we will not use linearity.

1. Let $F[x]$ be the surface under $y = f[x]$ from 0 till x, for known F and unknown f and y that are to be determined (note this order). For example $F[x] = x^2$ gives a surface under some f and we want to know that f.
2. We take a small step Δx. The surface area under f becomes larger. We do not know f but we know that $F[x + \Delta x]$ gives the new total surface area. (See the upper graph. Start at the right hand side with some Δx and see what it means for the left hand side.)
3. The change in surface is $\Delta F = F[x + \Delta x] - F[x]$. When $\Delta x = 0$ then $\Delta F = 0$.
4. The surface change can be approximated in various ways. Of these $\Delta F \approx y\, \Delta x$ is the simplest expression with $y = f[x]$ for the unknown function and $y \approx \Delta F / \Delta x$ its estimate. But we are not quite interested in approximation and $\Delta F / \Delta x$ seems undefined for $\Delta x = 0$.
5. The idea is: When we find an expression for $y = f[x]$ such that $\Delta F // \Delta x$ is not undefined at $\Delta x = 0$ and then set $\Delta x = 0$ then we have extracted f (while at $\Delta x =$

0 the exact surface is given by $F[x]$, known).

6. For $\Delta x \neq 0$ we simplify $u = \Delta F \,//\, \Delta x$ algebraically.
7. We try whether setting $\Delta x = 0$ gives a defined outcome. When that is the case, then we set $y = u$ as that outcome. We also expand the domain with $\Delta x = 0$.
8. We thus have the program: $y = \{\Delta F \,//\, \Delta x$, then set $\Delta x = 0\}$. It is the dynamic quotient, thus first algebra assuming $\Delta x \neq 0$ and then expanding the domain, but with the crucial added step of actually setting $\Delta x = 0$.
9. We rewrite $y = f[x] = F'[x]$ as well.
10. In summary, the derivative is $f[x] = F'[x] = \boldsymbol{d}F \,/\, \boldsymbol{d}x = \{\Delta F \,//\, \Delta x$, then set $\Delta x = 0\}$. The Δ stands for "difference" (a defined step) and the $\boldsymbol{d}$ stands for "differential" - a difference that doesn't actually exist since the step has been set to zero. This contains a seeming 'division by zero' while actually there is no such division, since we have been adjusting the domain. The expression "to differentiate" means finding the derivative.

PM. This is discussed in more detail in §15.5.

11.3.3 Check on what we already know

Let us check the derivatives that we know about.

	Linear	*Quadratic*
$F[x]$	$s\,x$	$a\,x^2$
$F[x+\Delta x]$	$s\,(x + \Delta x)$	$a\,(x + \Delta x)^2 = a\,(x^2 + 2\,x\,\Delta x + \Delta x^2)$
ΔF	$s\;\Delta x$	$a\left(2\,x\,\Delta x + \Delta x^2\right)$
$\Delta F \;//\; \Delta x$	s	$a\,(2\,x + \Delta x)$
Set $\Delta x = 0$	s	$2\,a\,x$

Applications of $\boldsymbol{d}F \,/\, \boldsymbol{d}x = \{\Delta F \,//\, \Delta x$, then set $\Delta x = 0\}$. For example $a = ½\, s$.

11.3.4 Application to a new function

Let us apply the rule to a function that we have not seen before: $F[x] = a\,x^3$.

$$a\,(x + \Delta x)^3 = a\,x^3 + 3\,a\,x^2\,\Delta x + 3\,a\,x\,\Delta x^2 + a\,\Delta x^3$$

$$\frac{a\,(x+\Delta x)^3 - a\,x^3}{\Delta x} = \frac{3\,a\,x^2\,\Delta x + 3\,a\,x\,\Delta x^2 + a\,\Delta x^3}{\Delta x}$$

$$\frac{a\,(x+\Delta x)^3 - a\,x^3}{\Delta x} = a\left(3\,x^2 + 3\,x\,\Delta x + \Delta x^2\right)$$

Extend domain and set $\Delta x = 0 \quad \Rightarrow f(x) = 3\,a\,x^2$

This answers the opening question of this section: what is $Sur[x^2, x]$? The answer is $F[x] = Sur[x^2, x] = 1/3\ x^3$. Namely, with $a = 1/3$ in the above.

In this manner the idea rises that $F[x] = Sur[\ a\ x^n, x] = \frac{1}{n+1}\ a\ x^{n+1}$ in general.

11.3.5 The constant and the switch from surface to integral

11.3.5.1 The surface under *y* = 0

What we haven't properly discussed yet is the constant function: $F[x] = c$. This may be a confusing function to substitute in. Write: $F[x] = c + 0 * x$. Then $\Delta F = F[x + \Delta x] - F[x] = c - c = 0$. The dynamic quotient $0 // \Delta x$ can be simplified to 0. There is no problem in extending the domain. Thus the derivative of a constant is 0. Thus when $F[x] = c$ then $f[x] = F'[x] = 0$.

Now a paradox arises when we want to go from the derivative to the primitive. For $f[x] = 0$ there is no surface between it and the horizontal axis. For convenience of notation we still use $F[x] = Sur[0, x] = c$, but use the abstract term "integral" for all F rather than surface. We also accept that the constant can be any c, of unknown size. This also means that we have to adapt our rules: when going in reverse direction from the derivative f to the integral $F[x] = Sur[f, x]$ then we must include an unknown C, called "the integration constant". Above we used the label "(*idea*)" but we can now observe that the "if and only if" clause only holds when we include the integration constant in the surface or, better, integral function.

11.3.5.2 The general relation

For power functions or polynomials in general (now from idea to theorem):

For $n = -1$ we still have to determine the integral of $1 / x$:
The primitive $F[x] = C + c\,x + ½\ s\,x^2 + ... + a\,x^{n+1} / (n+1) = Sur[F', x]$
iff the derivative is $F'[x] = c + s\,x + ... + a\,x^n$.

A function tells us how fast the surface under it is changing.

We can prove this by means of recursion. Let us assume that it holds for some n and then show that it holds for n+1. We have already shown it for n = 1 and 2 so from there we can work up towards infinity. Since this relationship between primitive and derivative is equivalent in either direction, we can develop the proof in the easier manner of taking the derivative, while making sure that each step is reversible. We consider only the derivative of $a\,x^{n+1} + c$ since the addition with lower terms is the same. In steps:

Assume that it holds for n:

0. We have $\frac{d}{dx}(a\,x^n + c) = \{a\,((x + \Delta x)^n - x^n) // \Delta x$, then set $\Delta x = 0\} = a\,n\,x^{n-1}$

We could quickly prove it by replacing n+1 for n in this very form: $\frac{d}{dx}(a\,x^{n+1} + c) = a\,(n+1)\,x^{n+1-1} = a\,(n+1)\,x^n$. The form persists. However, in all likelihood, the smaller steps carry more convincing power:

1. $f\,[x] = F'[x] = \boldsymbol{d}F \,/\, \boldsymbol{d}x = \frac{d}{dx}(a\,x^{n+1} + c) = \{\Delta F \,//\, \Delta x$, then set $\Delta x = 0\}$.
2. $\Delta F = a\,(x + \Delta \mathrm{x})^{n+1} - a\,x^{n+1}$
3. $\Delta F = a\,(x + \Delta \mathrm{x})^n\,(x + \Delta x) - a\,x^{n+1}$
4. $\Delta F = a\,x\,(x + \Delta \mathrm{x})^n \quad - a\,x^{n+1} + a\,(x + \Delta \mathrm{x})^n\,\Delta \mathrm{x}$
5. $\Delta F = a\,x\,(x + \Delta \mathrm{x})^n \quad - a\,x\,x^n + a\,(x + \Delta \mathrm{x})^n\,\Delta \mathrm{x}$
6. $\Delta F = a\,x\,\{(x + \Delta \mathrm{x})^n - x^n\} \quad + a\,(x + \Delta \mathrm{x})^n\,\Delta \mathrm{x}$
7. $\Delta F \,//\, \Delta x = x\,\{a\,((x + \Delta \mathrm{x})^n - x^n) \,//\, \Delta x\} + a\,(x + \Delta \mathrm{x})^n$
8. $\{\Delta F \,//\, \Delta x$, then set $\Delta x = 0\} = x\,\{a\,((x + \Delta \mathrm{x})^n - x^n) \,//\, \Delta \mathrm{x}$, then set $\Delta x = 0\} + a\,(x)^n$
9. $\quad = x\,\{a\,n\,x^{n-1}\} + a\,(x)^n = a\,(n + 1)\,x^n$
10. $f\,[x] = F'[x] = \boldsymbol{d}F \,/\, \boldsymbol{d}x = \frac{d}{dx}(a\,x^{n+1} + c) = a\,(n + 1)\,x^n$

Thus we have a proof for polynomials, and a nice example of a proof by recursion.

11.3.5.3 Definite and indefinite integral

The functions that we have been looking at all gave surfaces from 0 to x. For a general interval domain [d, u] with down and up value we find $F[u]$ - $F[d]$ as the surface enclosed. The integration constant drops out.

- A *definite* integral uses a domain. We write $Sur[F', \{x, d, u\}] = F[u] - F[d]$.
- An *indefinite* integral leaves the domain unspecified. When going from the derivative to the primitive then it is imperative to include the integration constant C.

NB. The use of the letters F and $f = F'$ has been very useful while developing this theory of calculus. In the future we tend to use smaller case letters f and $f\,'$. Indeed, a derivative can have a derivative too, called the second derivative $f\,''$ when starting from f. Or if you start high enough, a third derivative f'''. In general $f^{(n)} = \boldsymbol{d}^n f \,/\, \boldsymbol{d}x^n$.

11.3.6 Notation

The conventions are:

- Derivative, both in *Mathematica* input and in traditional form.

```
D[f[x], x] == d f[x] / d x
```

$$f'(x) = \frac{\text{DifferentialD}[f(x)]}{\text{DifferentialD}[x]}$$

- The indefinite integral or a definite for an interval from a to b. The notation contains an elongated S and reflects that surface is found by $F[x] \approx f[x]\ \Delta x$.

```
See[ Integrate[f[x], x], versus, Integrate[f[x], {x, a, b}] ]
```

$\int f(x)\, dx$ versus $\int_a^b f(x)\, dx$

- An example, where *Mathematica* in its wisdom reorders the expression. The outcome assumes that you know that a constant needs to be included.

```
Integrate[c + 2 x + x^2, x]
```

$$c\,x + \frac{x^3}{3} + x^2$$

11.3.7 Historical importance

dF / dx is a crowning achievement of mathematics. It started with Archimede and was developed by Fermat, Newton, Leibniz, Cauchy, Weierstraß. Interestingly, while Newton developed his flux method and used it to derive and check results, he wrote his “Principia” in terms of geometry as was the standard of the day. It was an initiative of Emilie du Châtelet to rewrite the analysis in proper derivatives, which version became very popular and was instrumental in winning people over both on gravity and the method of analysis. Finally mankind had sound mathematical clarity about distance and speed and changes in them.

Newton denoted the derivative with the “flux dot” $\dot{y}$. Leibniz coined the ***dF / dx*** that is more explicit on what variable is variated. For integration he created $\int$ or an elongated S standing for a summation of surface under a function.

PM. A sum with more terms like $y = x_1 + x_2 + \ldots + x_n$ can be compacted into the expression $y = \sum_{i=1}^{n} x_i$ where Σ is capital sigma or Greek S, called the “sum sign”, and where i is an index, a variable, that runs over the indicated integer domain. Leibniz’s $\int$ is a variant of Σ intended for surface components that are thought infinitely small.

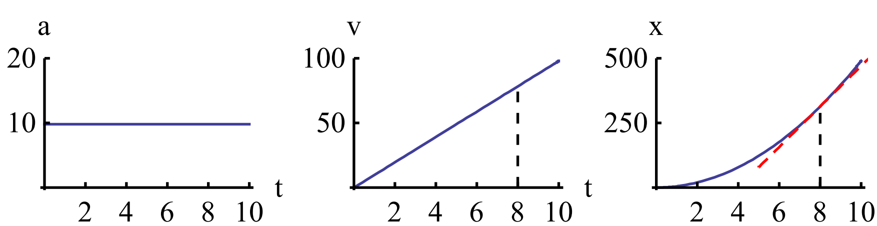
PM. More on integrals and derivatives is in §15.5. At this point the following small aspects can be mentioned. (1) Leibniz assumed that f was known and then found the F as a summation over $f[x]\ \Delta x$. His S is *summation* in Latin (summa) and not quite *surface* though he summed surface (spacium, space), later formalized in

Riemann sums of a string of $f[x]\ \Delta x$. We retain the notation. (2) The conventional approach in calculus is oriented on numbers and then uses concepts of numerical continuity and limits. Our approach relies on the algebraic or formal identity in the formulas. This is a notion of continuity too, not in the sense of numerical continuity but in the sense of 'same formula'. Limits are important but notably for $x \to \infty$ and not for $\Delta x \to 0$ in calculus. We manipulate the domain and distinguish between equation ($\Delta x = 0$) and setting to zero ($\Delta x := 0$). The whole is an algebraic operation in the creation of a formula and there is no "vanishing zero". (3) Conventional calculus first introduces the derivative and then has another chapter on finding the surface. We handle them jointly, not quite since it is smart but rather since it is proper. Primitive and derivative go hand in hand, the one cannot do without the other. We assume F known and use a single Δx to recover the f. A function gives the change of surface under it. Since we made sure that each step was logically reversible, we also can go from f to F. (4) The traditional approach starts with the derivative and the interpretation of the slope but then invites problems with "division by zero" while we focus on surface that has a multiplication. (5) It is not proper here to compare the different methods of doing calculus. For a comparison both need to be fully developed, and we develop only one here.

11.4 The derivative as a slope

11.4.1 The insight

We have learned that the slope is $s = \Delta y \,/\, \Delta x$. Focussing on the dynamic quotient $\Delta F \,//\, \Delta x$ it is a quick decision that apparently this is a slope. As x does one step to the right then F takes a vertical step. When $\Delta x \neq 0$ then we have an average slope for F but when we extend the domain with $\Delta x = 0$ then this must be the slope at that very point. The derivative is the slope of the primitive. See example §11.4.3.4 below. Using the example in physics of §12.2.1: compared to the average speed over an interval of time we now get the instantaneous speed at a point of time. E.g. $v[8]$ is the slope at $x[8]$.

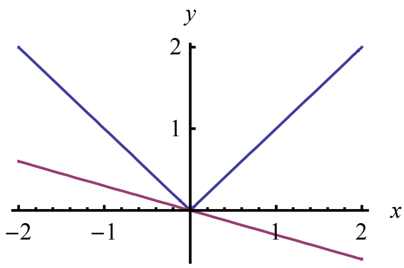


The notion of slope needs some attention however. Consider the functions $|x|$ and sign$[x]$ and their slopes, for example.

11.4.2 Derivative and slope of abs[x]

What is the slope of $|x|$ at the origin ? What are the tangent and tangent line ?

- Abs[x] and an example line through the origin.



Some observations are:

- The derivative to the left is -1 and to the right +1, which differ, so that there seems to be no overall derivative.
- When "tangent line" is defined as having the point {0, 0} in common without intersection then these can have slopes from –1 to 1.
- When "tangent line" is defined as having the "same slope" then which slope ?
- When the derivative is interpreted as a slope then the derivative of $|x|$ seems undefined, since any line through the origin might be said to be somewhat slopish. Having no intersection with $|x|$ is no strong condition since we can imagine functions with slopes where the tangent line does intersect (if tangent means "same slope"): for example a function that rises, becomes fully flat and then rises again.

The dynamic quotient helps out. For $x \neq 0$, the various combinations of $(|x + \Delta x| - |x|) // \Delta x$ give the normal result, sign[x]. For $x = 0$ the dynamic quotient gives $(|0 + \Delta x| - |0|) // \Delta x = |\Delta x| // \Delta x =$ sign[Δx]. Setting $\Delta x = 0$ gives 0. Hence in general $|x|' =$ sign[x].

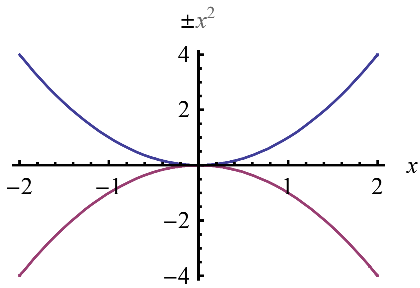
This may be seen as partly a matter of definition and partly a matter of being consistent on the notion of surface. The sign function itself may be defined as sign[x] $= x \,/\, |x|$ with an exception gate at 0. But it is also possible to define $|x| =$ sign[x] x. For surface: $|x|$ is the surface under some function, this appears to be sign[x], and the surface under this function does not change at $x = 0$.

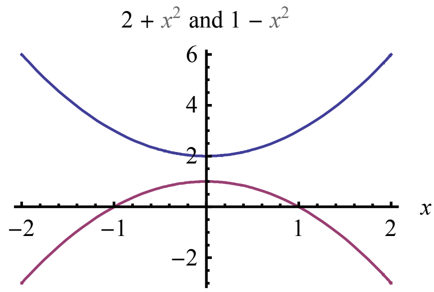
When "tangent line" is defined as having the same slope as the function then here there are only the three slopes –1, 0, 1.

11.4.3 Application to the parabola

11.4.3.1 Mirrors and balls

A quadratic function $f[x] = c + s\ x + a\ x^2$ is called a parabola (with vertical orientation). Parabolic mirrors reflect sunlight (parallel rays) to a focus. The formula also describes the path of a ball under the influence of gravity. There is a turning point and there may be intersections with a horizontal axis such as where a ball drops. The following plots give the three options: (1) one intersection, (2) no intersection, (3) two intersections.





11.4.3.2 Location of the turning point

A typical question is to find the turning point or vertex of the parabola. It can be found where the slope is zero. Thus we set the derivative equal to zero.

- For example for $c = 2$, $s = 2$ and $a = -1$.

D[$c + s\,x + a\,x^2$, x] == 0

$2\,a\,x + s = 0$

$\{\{x \to -\frac{s}{2a}\}\}$ gives $\{\{x \to 1\}\}$

11.4.3.3 Solving for intersections

Another typical question is to intersect a parabola with a line, $c + s\,x + a\,x^2 = b\,x + d$. For example the parabola gives the trajectory of a ball and the line gives the position and slope of the stands. When both coefficients of the line are zero then this is just the intersection with the horizontal axis. For nonzero values we write $(c - d) + (s - b)\,x + a\,x^2 = 0$ so that we have just another parabola, but shifted along the plane. The general solution is called the Quadratic Formula.

Solve[$c + s\,x + a\,x^2$ == 0, x]

$$\left\{\left\{x \to \frac{-\sqrt{s^2 - 4\,a\,c} - s}{2\,a}\right\}, \left\{x \to \frac{\sqrt{s^2 - 4\,a\,c} - s}{2\,a}\right\}\right\}$$

```
Result /. {c → 2, s → 2, a → −1}
```

$$\left\{\left\{x \to \frac{1}{2}\left(2+2\sqrt{3}\right)\right\}, \left\{x \to \frac{1}{2}\left(2-2\sqrt{3}\right)\right\}\right\}$$

PM. Another way to find the turning point is that we can choose c such that the square roots are zero. Thus we shift the parabola vertically so that there is only one solution. The neat thing is that we do not have to calculate that shift, we merely set the roots to zero.

11.4.3.4 Solving for tangent lines

To determine the tangent line at a point on the parabola, we first determine the co-ordinates, then the slope, and then we substitute these findings into the definition of a line.

- Suppose that $c = 2$, $s = 2$ and $a = -1$, and the point is at $x = 4$.

```
yval = c + s x + a x^2 /. {x → 4, c → 2, s → 2, a → −1}
```

-6

```
slope = D[c + s x + a x^2, x] /. {x → 4, c → 2, s → 2, a → −1}
```

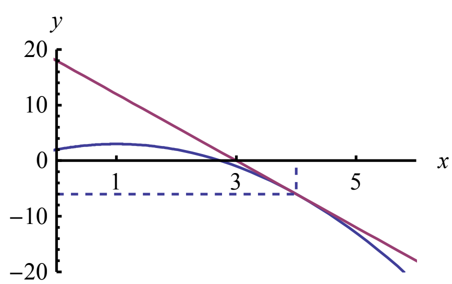
-6

```
yval == slope x + constant /. {x −> 4} // Simplify
```

constant $= 18$

```
y == slope x + 18
```

$y = 18 - 6x$



11.5 Product, quotient and chain rules

11.5.1 Product rule

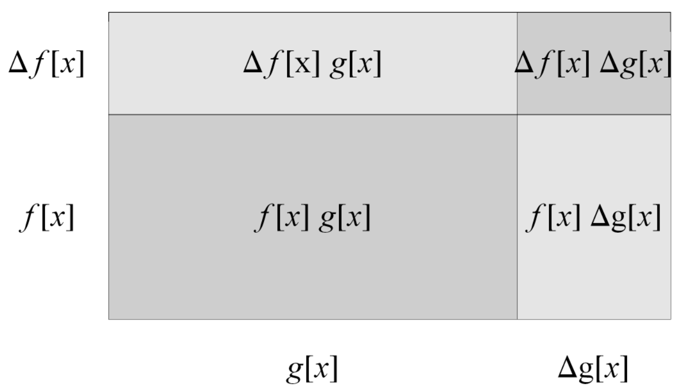
Consider $h[x] = (2\,x + 1)\,(x^2 + 4x + 1234)$. When we want to find the derivative then we can expand the expression, find the separate terms, and then differentiate them. Sometimes it is quicker to use the following "product rule". Define the separate terms as $f[x] = 2x + 1$ and $g[x] = x^2 + 4x + 1234$, thus $h[x] = f[x]\, g[x]$. Then:

When $h = f g$ and f' and g' exist then $h' = (f g)' = f' g + f g'$.

With f' and g': $h'[x] = \{\mathbf{2}\ (x^2 + 4x + 1234)\} + \{(2x + 1)\ (\mathbf{2x + 4})\} = 6x^2 + 18x + 2472$.

The rule can be proven with the following decomposition of terms.

- $h[x + \Delta x] = f[x + \Delta x] \times g[x + \Delta x]$ while $f[x + \Delta x] = f[x] + \Delta f[x]$.



Hence:

$\Delta h(x) = h(\Delta x + x) - h(x) = f(\Delta x + x)\, g(\Delta x + x) - f(x)\, g(x)$

$\Delta h(x) = (f(x) + \Delta f(x))\, (g(x) + \Delta g(x)) - f(x)\, g(x)$

$\Delta h(x) = \Delta g(x)\, f(\Delta x + x) + g(x)\, \Delta f(x)$

$\frac{\Delta h(x)}{\Delta x} = \frac{\Delta g(x)\, f(\Delta x + x) + g(x)\, \Delta f(x)}{\Delta x}$

$\frac{\Delta h(x)}{\Delta x} = (\Delta g[x]\ //\ \Delta x)\ f[x + \Delta x] + (\Delta f[x]\ //\ \Delta x)\ g[x]$

$\frac{\Delta h(x)}{\Delta x} = g'[x]\ f[x + \Delta x] + f'[x]\ g[x]$

Extend domain and set $\Delta x = 0 \Rightarrow h' = g f' + f g'$

The quotient $\frac{\Delta h(x)}{\Delta x}$ is dynamic. $\Delta g[x]\ //\ \Delta x$ and $\Delta f[x]\ //\ \Delta x$ simplify to expressions without Δx in the denominator. When Δx is set to 0 then $f[x + \Delta x] = f[x]$.

11.5.2 Quotient rule

We can state the rule directly:

When $h = f / g$ and f' and g' and h' exist then $h' = (f' g - f g') / g^2$.

When $h = f / g$ then $f = g h$ and we use the product rule. This is also why we require that h' exists. We give all equations and eliminate h.

```
Solve[{h == f / g, f' == g' h + g h'}, h', h]
```

$$\left\{\left\{h' \to -\frac{f\,g' - g\,f'}{g^2}\right\}\right\}$$

11.5.3 Chain rule

The sales revenue of an icecream vendor depends upon the number of ice cream cones sold. This sales volume depends upon the temperature. Thus revenue[sales[temperature]] is a function of a function. A unit change in sales has a particular effect *A* on revenue. A unit change in temperature has a particular effect *B* on sales. To find the total change in revenue the effect of a unit change in temperature must be multiplied by the effect of a unit change in sales volume.

This is called the chain rule and it can be expressed in two ways:

When $y = f[x]$ and $h[x] = \mathrm{g}[y]$ and $f\,'$ and g' exist then $h'[x] = g'[y] * f'[x]$.

Compactly: when $h[x] = g[f[x]]$ and $f\,'$ and g' exist then $h'[x] = g'[f[x]] * f'[x]$.

Even more compact is writing $\mathrm{g}[y = f[x]]$, eliminating a need for *h,* rather using the variables not as mere places but as meaningful in themselves. The proof:

$$\boldsymbol{dg} \,/\, \boldsymbol{dx} = \{\Delta g \,//\, \Delta x, \text{ then set } \Delta x = 0\}$$

$$= \{\Delta g \,//\, \Delta y * \Delta y \,//\, \Delta x \text{ for } (\Delta x = 0 \Leftrightarrow \Delta y = 0), \text{ then set } \Delta x = 0\}$$

$$= \{\Delta g \,//\, \Delta f, \text{ then set } \Delta f = 0\} * \{\Delta f \,//\, \Delta x, \text{ then set } \Delta x = 0\}$$

$$= \boldsymbol{dg} \,/\, \boldsymbol{df} * \boldsymbol{df} \,/\, \boldsymbol{dx}$$

- For example for an unspecified $z = g[x]$ and a function that contains z^2.

```
D[g[x]^2 + g[x], x]
```

$2\,g(x)\,g'(x) + g'(x)$

- If distance is a function of time, speed is the change of distance, and acceleration the change of speed, this simplifies to the second derivative.

```
v[t] == D[x[t], t];
a[t] == D[v[t], t] == D[D[x[t], t], t]
```

$a(t) = v'(t) = x''(t)$

11.6 Inverse function

If $y = f[x]$ is differentiable and if its inverse $x = g[y]$ exists then we can apply the chain rule to find the derivative of that inverse. To avoid cluttering in our formules with $'$ and -1 we write the inverse as g instead of f^{-1}. Then $g[f[x]] = x$ so that the outcome of differentiation to x must be 1. The chain rule gives $g'[y]\, f'[x] = 1$ that solves neatly.

Thus $g'[y] = 1 / f'[x] = 1 / f'[g[y]]$ where we replace $x = g[y]$ in that denominator.

This property is useful to find the inverses of polynomials, thus functions with fractional powers. For positive x the inverse of $f[x] = x^n$ is $x = g[y] = y^{1/n}$. The derivative of the latter appears to be:

$$g'[y] = \left(y^{1/n}\right)' = \frac{1}{f'[x]} = \frac{1}{(x^n)'} = \frac{1}{n\, x^{n-1}} = \frac{x^{1-n}}{n} = \frac{y^{1/n-1}}{n} = \frac{1}{n}\, y^{1/n-1}$$

We observe the same rule: the exponent drops to the base level and the exponent is reduced by 1.

- The derivative of $\sqrt{x}$ is used a lot.

$$(\sqrt{x})' = \frac{1}{2\sqrt{x}}$$

11.7 Arc length along a curve

After being kicked, a ball drops over at 5 meters but how far has it traversed along its curve ? Distance can be measured not just between two points but also along a winding curve. Let our winding curve be $f[x]$.

When we hold slope s fixed and if x moves by Δx then $y = s\, x$ moves by $\Delta y = s\, \Delta x$. The joint movement or arc ΔA is given by Pythagoras.

$$\Delta A = \sqrt{(\Delta \mathrm{x})^2 + (\Delta \mathrm{y})^2} = \sqrt{(\Delta \mathrm{x})^2 + s^2\, (\Delta \mathrm{x})^2} = \Delta x \sqrt{1 + s^2}$$

We recognize the surface change $\Delta A \approx \Delta x\, a[x]$ for some function a. The above uses $a[x] = \sqrt{1 + s[x]^2}$ with fixed slope $s[x] = s$. Let us now use variable $y = f[x]$. A variable slope $s[x]$ comes from the derivative of f, thus $s[x] = f'[x]$. Hence $a[x] = \sqrt{1 + f'[x]^2}$. Then A is this primitive:

$$A[x] = Sur[\sqrt{1 + f'[x]^2}\,,\, x]$$

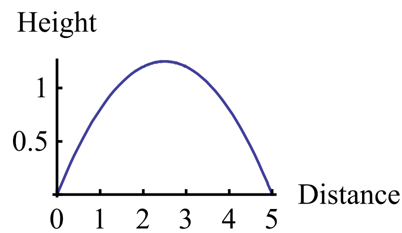
Conditions are that $y = f[x]$ is differentiable and $f[x]$ and $f'[x]$ are continuous (without holes or jumps). There are functions with indeterminate arcs like the

Koch function but that leads too far here.

Return to the kick of the ball. When we express the trajectory of the ball as a function of time then the arc would contain two dimensions, height and time. This is not a proper distance. Distance is in meters and not seconds. For proper distance we need to express the height as a function of the horizontal distance.

- Let the ball be kicked with vertical and horizontal velocities $v = w = 5$, and let gravity be 10. The parabola as a function of horizontal distance is:

$$h = x - \frac{x^2}{5}$$



- The derivative and arc along the curve. From the graph we may guestimate with an isosceles triangle with base 5 and height 1.25, then employ Pythagoras - and check how close that is to the accurate value.

$h' = 1 - \frac{2x}{5}$ gives Integrate$\left[\sqrt{\left(1 - \frac{2x}{5}\right)^2 + 1}, \{x, 0, 5.\}\right] = 5.73897$

See Jon McLoone http://blog.wolfram.com/2010/09/27/do-computers-dumb-down-math-education/

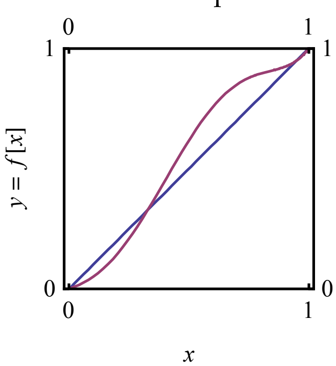
12. For e, $\mathbb{X}$ and $\mathbb{Y}$

12.1 The exponential number e

12.1.1 The notion of a fixed point

Consider function f from x on the interval [0, 1] to y on the interval [0, 1]. A special function is $y = x$, also known as the diagonal. When we consider an arbitrary continuous function that starts at {0, 0} and that ends at {1, 1} then this function must touch or cross the diagonal at least once (not counting the origin). Below graph uses an arbitrary function. This example is smooth but it might have kicks and bounds, and cross more times, as long as it is continuous. A point where it touches or crosses the diagonal is called a *fixed point*: there $f[x] = x$, or the application of the function to that point generates that very point itself again. Repeated application still gives the same point. This is an instance of Brouwer's fixed point theorem.

- Brouwer's fixed point theorem on the unit interval.



We apply this notion now to functions. We collect all the functions that we know into something called "function space" $\mathcal{F}$. Instead of addition we have the addition of derivatives and instead of multiplication we have the product and chain rules. Instead of the diagonal we have the "identity function" - a function is always identical to itself. Now we reach an important conclusion: Given Brouwer's general theorem this function space $\mathcal{F}$ must also have fixed points. In particular: there must exist a function f that is its own derivative, $f' = f$.

This is the theoretical basis of the exponential number e = 2.71828.... We now leave function space and return to good old plane geometry. The notion of a fixed point will stay with us though.

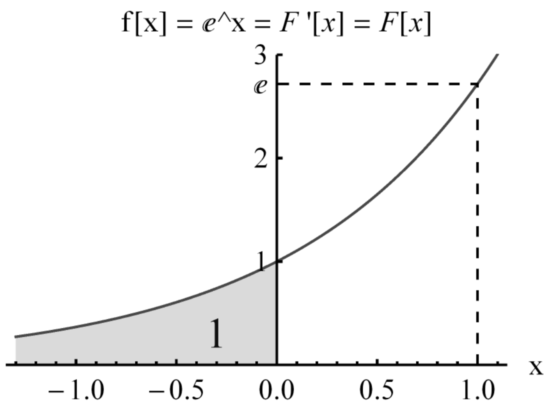
12.1.2 Definition of e

The exponential number e is defined such that the surface under e^x is described by the very function itself. If we measure the surface from 0 then we define $Sur[e^x, x] = e^x - 1$. Since the integration constant is somewhat arbitrary we can also adopt a slight change of measurement of surface: instead of measuring from 0 we now start at -∞. We keep using the same name. Then $Sur[e^x, x] = e^x$ is a fixed point in terms of surface measurement and differentiation. Thus $f[x] = F'[x] = (e^x)' = F[x] = e^x$. Key points are:

- $e^0 = 1$ must be the value of the surface from -∞ to 0.
- $e^1 = e$ must be the value of the surface from -∞ to 1.
- The slope at $x = 0$ must be $e^0 = 1$.

When discussing surfaces and derivatives we are used to two graphs, the surface to the right and the derivative on the left, but now we have one graph only. It is the same function.

- The derivative is the primitive. Numerically e = 2.71828....



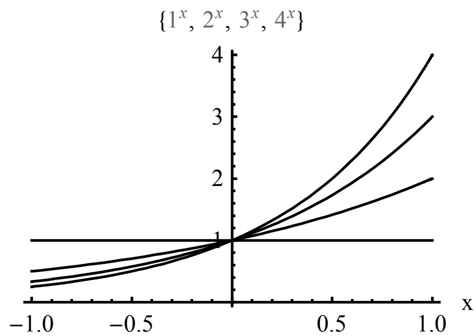
12.1.3 Exponentials and recovered exponentials

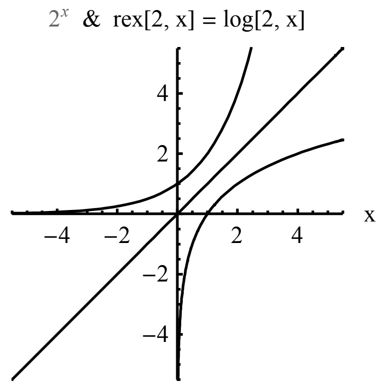
We can understand the meaning of the number e also by looking at graphs of the exponential functions and their inverse functions, the recovered exponentials.

The exponential functions a^x have $a^0 = 1$ and thus all pass through {0, 1}. With respect to the graph to the left and looking at {0, 1} we can reason that there must be a number like e. For $a = 1$ the slope is 0 and for $a = 4$ the slope is very steep. There must be a number E such that the slope must be 1. When we start at {0, 1} and take a slope of 1 then we arrive at {1, 2}. This is the point of 2^x at $x = 1$ as well.

However, 2^x moves below the first rising diagonal so that the number E must be larger than 2. From the graph we see that $2 < E < 3$. Actually $E = e = 2.71828...$

- Left: exponential functions: $f[1] = 1, 2, 3, 4$ and $f[-1] = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}$.
  Right: the mirror image of 2^x over the line $y = x$ gives the inverse rex[2, x].





The inverse to an exponential function $y = a^x$ is the recovered exponent x = rex[a, y] (a.k.a. the logarithm). Both are plotted in the graph to the right for $a = 2$, using the same causal argument x (otherwise flip the graph). All recovered exponents have 0 = rex[a, 1] so pass through the point {1, 0}. Hence:

- For the function e^x and base e we write rex[x] = rex[e, x] or the "natural rex" (a.k.a. the natural logarithm). Then rex[1] = 0 and rex[e] = 1.
- Since e^x passes through {0, 1} with slope 1 then the mirror rex[x] passes through {1, 0} with slope 1 too.

Above graph uses 2^x so is not perfect for a visual check on e. But we can imagine two lines parallel to $y = x$ just tangent to the curves at {0, 1} and {1, 0}.

12.1.4 The derivative of an exponential function

The derivative of an exponential function follows from the chain rule and the presumption that Exp[x] = e^x is the fixed point in differentiation:

$$\frac{d\,a^x}{dx} = \frac{d\,e^{x\,\mathrm{rex}[a]}}{dx} = e^{x\,\mathrm{rex}[a]}\ \mathrm{rex}[a] = a^x\,\mathrm{rex}[a]$$

If we choose $a = e$ then rex[e] = 1. Then indeed $\frac{d\,e^x}{dx} = e^x\,\mathrm{rex}[e] = e^x$.

12.1.5 The derivative of a recovered exponential function

What is the derivative of the rex ? It might be a shortcut to write $x = e^{\mathrm{rex}[x]}$ and use the chain rule but it is better to use the general rule for inverse functions, for practice. With $y = e^x$ and x = rex[y] we get:

$$\text{rex}'[y] = \frac{1}{(e^x)'} = \frac{1}{e^x} = \frac{1}{y}$$

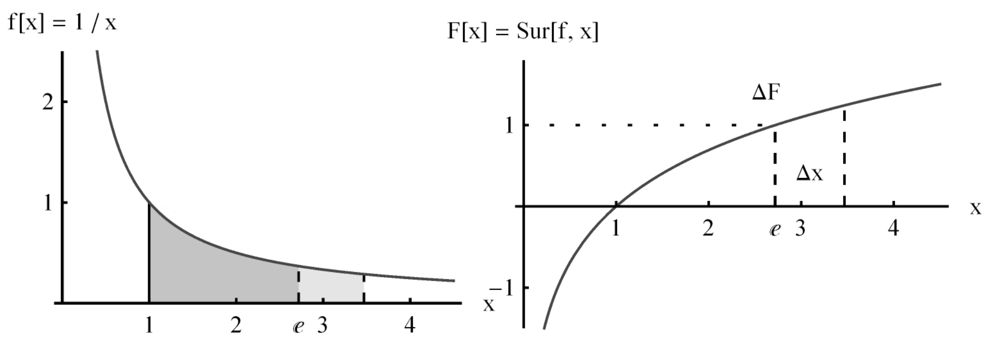
Or rex'[x] = 1 / x.

When there is a different base to recover from then rex[b, x] = rex[x] / rex[b], and thus in general rex'[b, x] = rex'[x] / rex[b] = 1 / (x rex[b]).

12.1.6 The surface under 1 / x

We have been silent on the surface under $f[x] = 1 / x$. We have been taking surfaces from 0 to x but when we look at the graph of this function then this gives us an infinite surface. We might consider function 1 / (x + 1) but there is no law against flexibility so now we take the surface from an arbitrary (judicially chosen) point, namely x = 1. To the right we can plot the integral. We could write Sur2 but stick to Sur (it is integral in general anyway). Thus $F[x]$ = *Sur*[f, x] starts at x = 1 with value 0 and grows to the right. That means that there is also a number such that the surface has precisely value 1. We find: $F[1] = 0$, $F[e] = 1$. Also, $1/x$ at x = 1 has value 1 so the slope of the primitive must be 1 there too, or $F'[1] = 1$.

- Surface size 1 under $1/x$ from x = 1 to e = 2.71828...



For negative values of x the rex is not defined. What works though is taking the absolute value since its derivative is the sign function.

Thus for $f[x] = 1/x$ we have $F[x]$ = *Sur*[f, x] = rex[$|x|$] + C.

12.1.7 A curious aspect of 1 / x

Though we have been discussing exponential functions it is useful to link up with the polynomial power functions for a short moment. For polynomials like x^2, x, 1 and x^{-2} we have nice derivatives and primitives but suddenly the function x^{-1} inbetween seems to behave differently since it links up with a strange rex[x]. It appears to be a matter of selecting a smart integration constant.

A common expression for the integral of the power function x^n causes a small problem when n = -1. The limit for $n \to$ -1 gives infinity. If we include a well chosen constant however then that limit generates the recovered exponent (logarithm).

- The integral of x^n in common form or with a smart constant.

$$\frac{x^{n+1}}{n+1} \quad \text{versus} \quad \frac{x^{n+1}}{n+1} - \frac{1}{n+1}$$

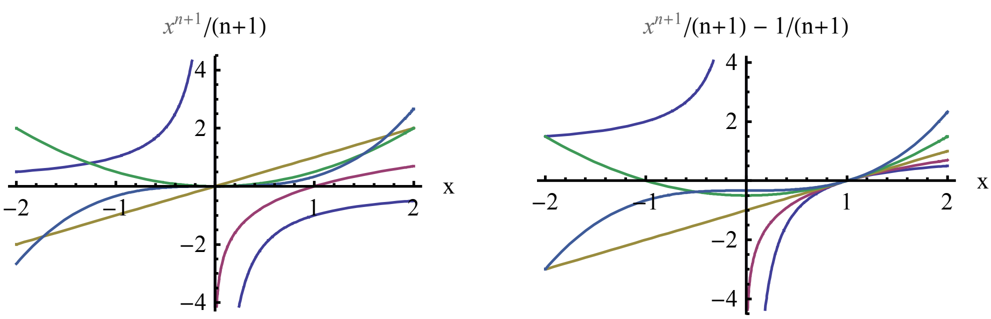
```
Limit[Result[[1]], n → −1] /. Log → $Rex
```

$(\infty \quad \text{versus} \quad \text{rex}[x])$

- When n is a number then *Mathematica* generates the correct result for -1. The integral for x^n for symbolic n can use a constant though. Check $x = 1$.

Power	-2	-1	0	1	2
Function	$\frac{1}{x^2}$	$\frac{1}{x}$	1	x	x^2
$\frac{x^{n+1}}{n+1}$	$-\frac{1}{x}$	$\text{rex}[x]$	x	$\frac{x^2}{2}$	$\frac{x^3}{3}$
$\frac{-1+x^{n+1}}{n+1}$	$1-\frac{1}{x}$	$\text{rex}[x]$	$x-1$	$\frac{1}{2}\left(x^2-1\right)$	$\frac{x^3}{3}-\frac{1}{3}$

- Left: standard, row 3. Right: adjusted, row 4. Check $x = 1$.



There is always the integration constant so the issue is primarily something to be aware of for $n = -1$ or for when graphs look cluttered. But in another way it is nice to see that the recovered exponent is no real deviant in the list of functions.

12.1.8 The algebra of e

12.1.8.1 Consistency proof in terms of deltas

In the section on the core of calculus §11.3.2 we took $\Delta F \approx y\, \Delta x$ and then used the dynamic quotient to find the unknown y. For e^x we know y by definition. Thus for $y = f[x] = F\,'[x] = F[x] = (e^x)' = e^x = \text{Exp}[x]$ we can directly work with the deltas and we do not need to take the quotient. When our definition is inappropriate we should find a contradiction:

$$\Delta F = \text{Exp}[x + \Delta x] - \text{Exp}[x] \approx e^x\, \Delta x$$

$$e^{x+\Delta x} - e^x \approx e^x \Delta x$$

$$e^x (e^{\Delta x} - 1) \approx e^x \Delta x$$

$$(e^{\Delta x} - 1) \approx \Delta x$$

Setting $\Delta x = 0$ gives $0 = 0$.

The definition $f[x] = F'[x] = F[x] = (e^x)' = e^x$ does not result into a contradiction and appears a consistent construct. The definition works since e^x itself drops out and we retain an expression that vanishes when it should. The exponential format turns the $x + \Delta x$ expression into a multiplicative term such that proportional elimination is possible. This is a property of exponents that is true by necessity, and hence in hindsight it should not come as a surprise that we can find such an exponential base for the fixed point in differentiation.

A contradiction seems to arise in the point that the deduction above is still valid for values $e = 6$ and 50, or whatever. Thus when $\Delta x = 0$ then these both satisfy $(e^{\Delta x} - 1) = \Delta x$. Any value does. This question however means that we turn from the issue of surfaces and differentiation to the issue of finding a good numerical estimate for e.

PM. When we do the same steps with a^x and use the chain rule on the derivative then we find $(a^{\Delta x} - 1) \approx \text{rex}[a]\, \Delta x$. Above is a special case for $a = e$. Write $\text{rex}[a]\, \Delta x = b$ and then $(a^{\Delta x} - 1) = (e^{\Delta x\, \text{rex}[a]} - 1) = (e^b - 1) \approx \text{rex}[a]\, \Delta x = b$ shows that this case reduces to the original problem for e.

12.1.8.2 Numerical approximation of e

Above deduction suggests a program to approximate e. Let $h = (\text{est}[h]^h - 1)$ imply the estimate which then solves into est[h] = (1 + h)^(1/ h). This is clearer for $n = 1/h > 1$ so that estn[n] = (1 + 1/n)^n while taking a power is easier than taking a root. Taking $n \to \infty$ gives an ever better estimate:

$$\begin{pmatrix} h & 0.1 & 0.01 & 0.001 & 0.0001 & 0.00001 & 1.\times 10^{-6} \\ \text{Estimate} & 2.59374 & 2.70481 & 2.71692 & 2.71815 & 2.71827 & 2.71828 \end{pmatrix}$$

We know that this program must converge on e since we have proven above that defining $f[x] = F'[x] = F[x] = (e^x)' = e^x$ is viable, so that if we let $\Delta x \to 0$ then $\Delta F \to 0$ and est[Δx] is constrained too. As with Θ though there may be better programs to estimate e too.

This is a key insight: the expression $(e^{\Delta x} - 1) \approx \Delta x$ should not be read as $(\text{est}[\Delta x]^{\Delta x} - 1) \approx \Delta x$ in the *only* sense of approximation. That confuses the approximation of e itself with the issue of calculus on surfaces. Of course it is always possible for any value of Δx to calculate some est[Δx] but that is not the portent of $\Delta F \approx y\, \Delta x$. The definition $f[x] = F'[x] = F[x] = (e^x)' = e^x$ has been given for some algebraically relevant number and not for some approximation.

For Θ we relied on the continuity of space and the existence of the ratio between circumference and radius; and, OK, we had a program to calculate it. For e we rely on the continuity of space between 2 and 3; and, OK, we also have a program.

With the dynamic quotient we can avoid limits in the determination of the derivatives. We cannot avoid the use of limit-methods in the numerical expansions of Θ and e. Those are points in space that require special methods if we try to catch them in a system of arithmetic. We can use space itself to measure space, like using a rod called "meter" to measure length and diagonals. If we want to map this into arithmetic then such algorithms are required.

12.1.8.3 Using the dynamic quotient

The dynamic quotient worked great for unknown derivative y and the polynomials. For the exponential functions it can lead us up the wrong alley. When we forget that $(e^x)' = e^x$ by definition then we create a quotient that does not easily resolve. Let us see what happens:

For any x:

$$e^x \equiv \boldsymbol{d}e^x \,/\, \boldsymbol{d}x = \{(\mathrm{Exp}[x + \Delta x] - \mathrm{Exp}[x]) \,//\, \Delta x, \text{ set } \Delta x = 0\}$$

$$e^x = \{(e^{x + \Delta x} - e^x) \,//\, \Delta x, \text{ set } \Delta x = 0\}$$

$$e^x = e^x \ * \ \{(e^{\Delta x} - 1) \,//\, \Delta x, \text{ set } \Delta x = 0\}$$

$$1 = \{(e^{\Delta x} - 1) \,//\, \Delta x, \text{ set } \Delta x = 0\}$$

Again e^x itself drops out, and we retain an expression that must be unity. Up to now we have been blessed with Δx dropping out. This does not seem to happen at this instance. The method at least suggests the approximation algorithm. But it can lead us astray if we do not keep the issue straight.

As said: we must beware of quotients like $(x - x) \,//\, (x - x)$ since the parts look like variables but are in fact constants. Now $(e^{\Delta x} - 1) \,//\, \Delta x$ looks dangerously similar. There are two views:

- The give-away is that the dynamic quotient must be equal to a number, in this case 1. We are tempted to multiply and create $\{\Delta x = (e^{\Delta x} - 1), \text{ set } \Delta x = 0\}$ but we have not formalized this step so we better be careful about that. Instead, we rather drop the dynamic quotient. It was introduced in a situation with unknown derivative, with $\Delta F \approx y\ \Delta x$ defined for surfaces and unknown y. We should return to this original situation and then it appears that the case is resolved. Because y is not unknown. The conclusion is that e is consistent (and we have an algorithm).

- A formal approach is to take it that 1 = {($e^{\Delta x}$ − 1) // Δx, set Δx = 0} is an acceptable conclusion that allows us to algebraically replace that particular dynamic quotient and program by the number 1. This author has actually followed this course for a few years before he realized the give-away and the first approach. This formal approach is less satisfactory though, since the Δx does not actually disappear even though the dynamic quotient is intended to let us accomplish that. Looking at the issue after some years afresh the give-away was seen. Having the first approach of course provides a base for this formal approach ... which is always nice to have if not actually needed.

12.2 Angles and arcs

12.2.1 Elementary deduction

We have been using x as the independent variable. Now the independent variable will be the angle and x will be the dependent variable.

For angle α we have co-ordinates $\mathbb{X}_\alpha$ and $\mathbb{Y}_\alpha$ on the unit circle. They are functionally dependent. When we take the derivative of one then the chain rule ends in the derivative of the other.

$$\mathbb{X}[\alpha] = \sqrt{1 - \mathbb{Y}(\alpha)^2}$$

$$d\mathbb{X}[\alpha]/d\alpha = \mathbb{X}'[\alpha] = -\frac{\mathbb{Y}(\alpha)\,\mathbb{Y}'(\alpha)}{\sqrt{1 - \mathbb{Y}(\alpha)^2}} = -\frac{\mathbb{Y}(\alpha)\,\mathbb{Y}'(\alpha)}{\mathbb{X}[\alpha]}$$

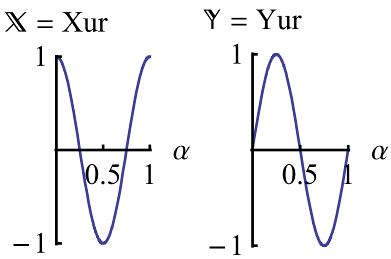
Dividing by the derivative of $\mathbb{Y}_\alpha$ so that all derivatives are on the left, gives the slope or tangent $s[\alpha]$:

$$\frac{\mathbb{X}'[\alpha]}{\mathbb{Y}'[\alpha]} = -\frac{\mathbb{Y}[\alpha]}{\mathbb{X}(\alpha)}$$

This must hold for all α. When we look at the plots of the original functions and check out the slopes then those must be constrained and must be zero at particular points. There are only two functions that satisfy these demands: $\mathbb{X}_\alpha$ and $\mathbb{Y}_\alpha$ themselves, up to a proportional constant. Thus the trigonometric functions have themselves as their derivatives.

Since it later appears that this proportional constant is Θ we can directly state:

$$d\,\mathbb{X}[\alpha]\ /\ d\alpha\ =\ \mathbb{X}'[\alpha]\ =\ -\,\Theta\,\mathbb{Y}[\alpha] \qquad d\,\mathbb{Y}[\alpha]\ /\ d\alpha\ =\ \mathbb{Y}'[\alpha]\ =\ \Theta\,\mathbb{X}[\alpha]$$



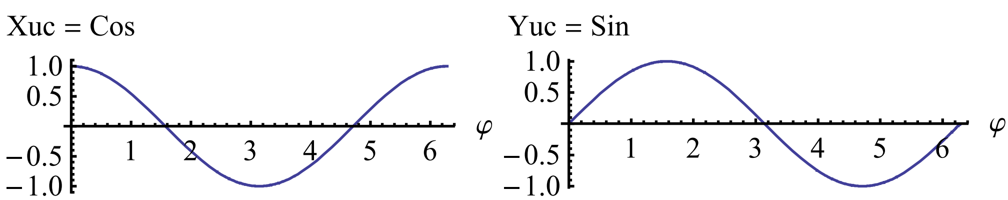
The slope of y_{ur} at $\alpha = 0$ must be Θ $\mathbb{X}[0] = \Theta$. The slopes of the waves are between values $-\Theta$ and Θ. (If you plot for a small range around 0 you might include a line x $\Theta + 0.001$ to show that they are parallel.)

12.2.2 For Cos and Sin

For Cos and Sin we get a proportional constant 1:

$d\mathrm{Cos}[\varphi] / d\varphi = \mathrm{Cos}'[\varphi] = -\mathrm{Sin}[\varphi]$ $\qquad$ $d\mathrm{Sin}[\varphi] / d\varphi = \mathrm{Sin}'[\varphi] = \mathrm{Cos}[\varphi]$

Looking at their graphs we might want to normalize by saying that a surface must be 1 but there is no particular reason yet why this would be so. However, the slope of Sin at $\varphi = 0$ must be 1 (angle 1/8 or 45°), as can be shown below, and that allows us to calibrate the proportional constants in the deductions. Since Cos[0] = 1, the proportional factor in the derivatives for Cos and Sin is 1.



Using the cosine rule on $\boldsymbol{d}\mathrm{Cos}[\varphi] / \boldsymbol{d}\varphi$ gives:

$$\Delta\mathrm{Cos}[\varphi] = \mathrm{Cos}[\varphi + \Delta\varphi] - \mathrm{Cos}[\varphi] = \{\mathrm{Cos}[\varphi]\,\mathrm{Cos}[\Delta\varphi] - \mathrm{Sin}[\varphi]\,\mathrm{Sin}[\Delta\varphi]\} - \mathrm{Cos}[\varphi]$$

$$\Delta\mathrm{Cos} // \Delta\varphi = \mathrm{Cos}[\varphi]\ \frac{\mathrm{Cos}[\Delta\varphi] - 1}{\Delta\varphi} - \mathrm{Sin}[\varphi]\ \frac{\mathrm{Sin}[\Delta\varphi]}{\Delta\varphi} = -\mathrm{Sin}[\varphi]$$

Look at the slopes at $\varphi = 0$. For Cos at $\varphi = 0$ we find that the slope must be zero, so that with the dynamic quotient $d\mathrm{Cos}[\varphi]/d\varphi\ [0] = \left\{ \frac{\mathrm{Cos}[0 + \Delta\varphi] - \mathrm{Cos}[0]}{\Delta\varphi} = \frac{\mathrm{Cos}[\Delta\varphi] - 1}{\Delta\varphi}\right.$, then set $\Delta\varphi = 0\} = 0$. See below for the more complex $\frac{\mathrm{Sin}[\Delta\varphi]}{\Delta\varphi} = 1$. Substituting these outcomes in $\Delta\mathrm{Cos} // \Delta\varphi$ gives that $\boldsymbol{d}\mathrm{Cos}[\varphi] / \boldsymbol{d}\varphi = -\mathrm{Sin}[\varphi]$.

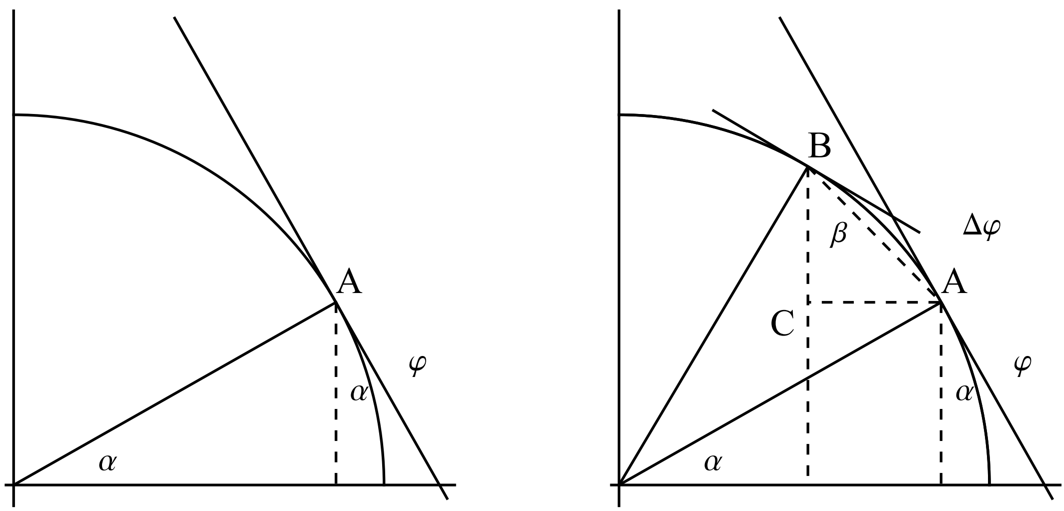
PM. We now also find the proportional constant for the functions depending upon angles: $\mathbb{Y}'[\alpha] = \Theta\ \mathrm{Sin}'[\alpha\ \Theta] = \Theta\ \mathrm{Cos}[\alpha\ \Theta] = \Theta\ \mathbb{X}[\alpha]$.

For the trigonometric functions (1) the derivatives are analytically easier than the surfaces but they translate into each other and (2) the arcs are easier than the angles. Since Cos and Sin do not get Θ in their derivative they and the measurement in radians can be very practical when derivatives are involved - even though the scale (radians) needs getting used to.

12.2.3 Geometric interpretation

12.2.3.1 The basic layout

The graph give the first quadrant of the unit circle. First look at the left. With angle α there is arc $\varphi = \alpha\ \Theta$. The co-ordinates of A are $\{X_A, Y_A\} = \{x_{\mathrm{ur}}[\alpha], y_{\mathrm{ur}}[\alpha]\} = \{\mathrm{Cos}[\varphi], \mathrm{Sin}[\varphi]\}$. At A we take the tangent, which is perpendicular to the radius. Thus angle α can also be found in the triangle created by projecting A onto the horizontal axis.



On the right hand side an increment $\Delta\varphi$ creates point B. In triangle ABC angle $\beta = \angle$ABC lies between α and $\alpha + \Delta\alpha$. Thus $\beta = \alpha + f[\Delta\alpha]$ for a fractional function. For its arc we find $\psi = \varphi + g[\Delta\varphi] \leq \varphi + \Delta\varphi$. Thus g is zero for $\Delta\varphi = 0$.

We have $\Delta\mathrm{Sin}[\varphi] = \mathrm{Sin}[\varphi + \Delta\varphi] - \mathrm{Sin}[\varphi] = Y_B - Y_A = BC$.

12.2.3.2 The proof by approximation

A traditional approach is to take $AB \approx \Delta\varphi$:

$$\text{Then } \Delta\mathrm{Sin} \,//\, \Delta\varphi = \frac{BC}{\Delta\varphi} = \frac{\mathrm{Cos}[\psi]\, AB}{\Delta\varphi} \approx \frac{\mathrm{Cos}[\psi]\, \Delta\varphi}{\Delta\varphi} = \mathrm{Cos}[\varphi + g[\Delta\varphi]].$$

Setting $\Delta\varphi = 0$ gives $\boldsymbol{d}\mathrm{Sin}[\varphi] \,/\, \boldsymbol{d}\varphi = \mathrm{Cos}[\varphi]$.

We may not be satisfied with this entirely since $AB \approx \Delta\varphi$ requires a theory on approximation that would much complicate calculus.

12.2.3.3 The proof by algebra

It is better to show that the slope of Sin at $\varphi = 0$ must be 1, since then the proof continues as above. The slope at zero is {ΔSin[φ] // $\Delta\varphi$ for $\varphi = 0$, set $\Delta\varphi = 0$} with (Sin[0 + $\Delta\varphi$] - Sin[0]) // $\Delta\varphi$ = Sin[$\Delta\varphi$] // $\Delta\varphi$.

We can simplify by looking at Sin[φ] // φ and set $\varphi = 0$.

- On the left hand side, take point $H = \overrightarrow{X} = \{1, 0\}$.
- Triangle OAH has height h = Sin[φ]. Its area is $\Sigma = h\, w / 2$ = Sin[φ] / 2.
- A sector of a circle is a part of the circle enclosed between two radii and their arc. The area of the α sector is a part φ / Θ of the whole ½ r^2 Θ, thus $\Phi = \varphi / 2$.
- The values of Tan[φ] are always on the line $x = 1$. Regard the right triangle OHD with D on extended OA (draw this yourself): it has area T = Tan[φ] / 2.
- Given the area locations we have $\Sigma \le \Phi \le$ T. Thus Sin[φ] $\le \varphi \le$ Tan[φ].
- Dynamic quotient by Sin[φ] and allow potentially $\varphi = 0$: $1 \le \varphi$ // Sin[φ] $\le$ 1/ Cos[φ].
- Set $\varphi = 0$: then 1 ≤ {φ // Sin[φ], set $\varphi = 0$} ≤ 1 squeezes a result that can be set to 1.
- If φ // Sin[φ] = 1 at $\varphi = 0$ then also Sin[φ] // φ = 1 at $\varphi = 0$.
- There is no approximation but logical deduction: Sin[φ] // φ = 1 at $\varphi = 0$.
- It is actually the slope at $\varphi = 0$: {ΔSin[φ] // $\Delta\varphi$ for $\varphi = 0$, set $\Delta\varphi = 0$} = 1.

In summary, the derivatives for the trigonometric functions follow from these steps (that do not depend upon approximation): (1) using the chain rule we find that the ratio of the derivatives is minus the slope, or tangent, i.e. the ratio of the primitives, (2) this holds for all angles and arcs, so the functions differentiate into each other with unknown proportional constant, (3) the slope 1 for Sin at 0 identifies the proportional constant.

See http://ocw.mit.edu/courses/mathematics/18-01sc-single-variable-calculus-fall-2010/part-a-definition-and-basic-rules/session-8-limits-of-sine-and-cosine/

12.2.3.4 A potential trap

The following is a deduction with a tricky step that is useful to be aware of.

$$\boldsymbol{d}\text{Sin}[\varphi] \,/\, \boldsymbol{d}\varphi = \{\Delta\text{Sin}[\varphi] \,//\, \Delta\varphi, \text{ then set } \Delta\varphi = 0\}.$$

Consider BC = Cos[ψ] AB = Cos[$\psi + h[\Delta\varphi]$] $\Delta\varphi$, for $\psi = \varphi + g[\Delta\varphi]$ and function h.

$$h = \begin{cases} \text{For } \Delta\varphi \neq 0 & h[\Delta\varphi] = \text{ArcCos}[\text{Cos}[\psi]\, AB \;/ \Delta\varphi] - \psi \\ \text{For } \Delta\varphi = 0 & \text{set } h[0] = 0 \end{cases}$$

$$\text{Then } \Delta\text{Sin} \;//\; \Delta\varphi = \frac{BC}{\Delta\varphi} = \frac{\text{Cos}[\psi]\, AB}{\Delta\varphi} = \frac{\text{Cos}[\psi + h[\Delta\varphi]]\, \Delta\varphi}{\Delta\varphi} = \text{Cos}[\varphi + g[\Delta\varphi] + h[\Delta\varphi]].$$

Setting $\Delta\varphi = 0$ gives $\boldsymbol{d}\text{Sin}[\varphi] \,/\, \boldsymbol{d}\varphi = \text{Cos}[\varphi]$.

The tricky thing is: In this way we can also prove that the derivative is Sqrt[φ]. Namely, substitute Cos[ψ] AB = Sqrt[ψ + $H[\Delta\varphi]$] $\Delta\varphi$, with $H[\Delta\varphi]$ = (Cos[ψ] AB / $\Delta\varphi)^2$ - ψ and for $\Delta\varphi = 0$ set $H[0] = 0$. A possible reason to reject this H is that it is not continuous at 0 but this leads too far into the realm of numerical continuity while it should suffice to use continuity in algebra and formulas.

12.2.4 The basis for Euler's form

On the unit circle z = Cos[φ] + i Sin[φ] so that $\boldsymbol{d}\, z \,/\, \boldsymbol{d}\, \varphi$ = - Sin[φ] + i Cos[φ] = $i\, z$. Differentiation gives the same number though with coefficient i. Since we know that e^x has itself as the derivative, we can express the one into the other.

While the derivative for $v[\varphi] = r\, e^{i\,\varphi}$ is $v'[\varphi] = i\, r\, e^{i\,\varphi} = i\, v[\varphi]$ then for angles we get a scaling factor just like happens for rex[b, x] for a base other than e. Thus $w'[\alpha] = i\, \Theta\, w[\alpha]$.

- The derivative gives a scale factor $i\, \Theta$ (setting $r = 1$)

$$w'(\alpha) = i\, \Theta\, e^{i\, \alpha\, \Theta} = \mathbb{X}'(\alpha) + i\, \mathbb{Y}'(\alpha) = -\Theta \sin(\alpha\, \Theta) + i\, \Theta \cos(\alpha\, \Theta) = i\, \Theta\, w(\alpha)$$

12.2.5 Arc length reconsidered

In §11.7 we derived how to calculate the length of a trajectory along a winding curve. We can apply that method also to the unit circle. The y can be expressed as a function of the x, we determine the derivative and apply the integral expression for the arc.

$$Y = \sqrt{1 - X^2}$$

$$\text{Y'} = -\frac{X}{\sqrt{1 - X^2}}$$

Integrate$\left[\sqrt{\mathbf{1 + Y'^{\wedge}2}}\,,\ \mathbf{X}\right]$ **// PowerExpand**

$$\sin^{-1}(X)$$

The result is ArcSin, indeed an arc function on the circle. It is a bit tricky though since ArcSin normally applies to Y and not to X. The solution to Y = Sin[φ] is φ = ArcSin[Y]. The arc measured by Sin is from {1, 0} counterclockwise. The integral above is for X going clockwise, and the arc traversed is from {0, 1} to {X, Y}. Thus we expect outcome Θ/4 - ArcSin[Y] = Θ/4 - ArcSin[$\sqrt{1 - X^2}$]. It appears to be an equivalent expression. A plot (not shown) may provide encouragement and in *Mathematica* with FullSimplify:

**FullSimplify[Pi/2 – ArcSin[Sqrt[1 – X ^ 2]] == ArcSin[X] ,
Assumptions → X ∈ Reals]**

$\sin^{-1}(X) = 0 \bigvee X \geq 0$

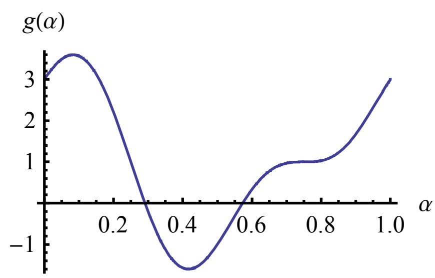
The calculation of arc length thus gives a consistent result. Above integration does not result into an explicit expression in terms of X without the Sin function. A development explicitly in terms of X and powers of X might be nice for a change. However, Sin^{-1}= ArcY exactly expresses the arc.

12.2.6 A typical exercise

To determine a minimum or maximum we can use that the slope must be zero. However, a point of inflection also has a zero slope (e.g. a function rises, goes flat, and rises again). In that case we can check on the value of the second derivative. For teaching it is a shortcut to allow a graphical inspection. (In the following, consider what happens at α = 3/4.)

12.2.6.1 In angles

Consider the function $g[\alpha]$ = 2 𝕏[α] + 𝕐[2α] + 1 in the domain [0, 1]. Calculate the global minimum and maximum.



A global maximum or minimum must have a corner solution or a zero slope so that the derivative must be zero. Given the graph the latter is the case.

$g'[\alpha]$ = -2 Θ 𝕐[α] + 2 Θ 𝕏[2α] = 0

𝕐[α] = 𝕏[2α]

𝕐[α] = 𝕐[1/4 - 2α]

$\alpha = (1/4 - 2\alpha) + k \quad \vee \quad 1/2 - \alpha = (1/4 - 2\alpha) + k \quad$ for k = 0, 1, 2, ...

$3\alpha = 1/4 + k \quad \vee \quad \alpha = - 1/4 + k$

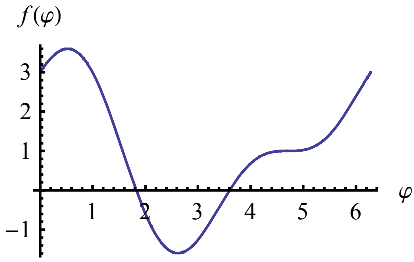
$\alpha = 1/12 + k / 3 \quad \vee \quad \alpha = 3/4 + k$

Requiring $0 \leq \alpha \leq 1$ gives $\alpha \in \{1/12, 5/12, 3/4\}$.

W.r.t. the graph, the global maximum is at $\alpha = 1/12$ and the global minimum at $\alpha = 5/12$. (And $\alpha = 3/4$ gives a point of inflection.)

12.2.6.2 In arcs

Consider the function $f[\varphi] = 2\cos[\varphi] + \sin[2\varphi] + 1$ in the domain $[0, 2\pi]$. Calculate the global minimum and maximum.



A maximum or minimum must have zero slope so that the derivative must be zero, or we have a corner solution. Given the graph the first is the case.

$f'[\varphi] = -2\sin[\varphi] + 2\cos[2\varphi] = 0$

$\sin[\varphi] = \cos[2\varphi]$

$\sin[\varphi] = \sin[\pi/2 - 2\varphi]$

$\varphi = (\pi/2 - 2\varphi) + 2\pi k \quad \vee \quad \pi - \varphi = (\pi/2 - 2\varphi) + 2\pi k$ for $k = 0, 1, 2, \ldots$

$3\varphi = \pi/2 + 2\pi k \quad \vee \quad \varphi = -\pi/2 + 2\pi k$

$\varphi = \pi/6 + 2/3\,\pi\,k \quad \vee \quad \varphi = 3/2\,\pi + 2\pi k$

Requiring $0 \leq \varphi \leq 2\pi$ gives $\varphi \in \{\pi/6, 5/6\,\pi, 3/2\,\pi\}$.

W.r.t. the graph, the global maximum is at $\varphi = \pi/6$ and the global minimum at $\varphi = 5/6\,\pi$. (And $\varphi = 3/2\,\pi$ gives a point of inflection.)

13. Partial derivatives

13.1 A function depending upon two variables

13.1.1 Principle

Function $y = f[x]$ is one-dimensional and associates with a two-dimensional point $\{x, y\}$. While this book is about the plane that is 2D it behoves to be aware of extensions to the 3D case and higher. Linear algebra subsequently allows us to imagine a more general vector space $\{x_1, ..., x_n\}$. We have seen this already with the linear regression and all the error points. These were only glances however. Closing this book on 2D, it remains a good question what message we may take home from these glances at multidimensionality.

It will be a matter of judgement but the best answer seems to be: the notion of partial derivatives.

Regard an example from the market place. When our expenditure consists of apples and bananas $Z = p_a\, q_a + p_b\, q_b$ then the inclusion of another apple will raise expenditure by p_a. This is that partial derivative. The inclusion of another banana will raise expenditure by p_b which is the partial derivative on the other variable and which gives a different outcome since generally the prices differ.

Living in a one-dimensional world with a one-dimensional mind, eating one-dimensional porridge with one-dimensional spoons, the latter outcome is strange. The change in Z has two possible outcomes instead of only one ? Both p_a and p_b ! What is happening here ? Well, the paradox is quickly resolved. The partial derivative gives the effect on the total when only one variable changes. When another or more variables change then the total may change differently.

The message to take home from multidimensionality thus is: that there is a key difference between the notions of total and partial derivatives. For the one-dimensional world the differential has to be the total derivative. If we presume that this would also be the case for the more-dimensional world then this would cause serieus errors of judgement.

This expenditure example is in linear space where the partial derivatives are given by prices that do not change. In general the partial derivatives change themselves and they change all at the same time too. The way to find the proper partial derivative is to keep the other factors constant and vary only that what we are interested in. This gives the standard Latin expression in economic analysis: *ceteris paribus,* meaning: keeping the other aspects to be the same.

Note how easily we are guided into error. Above we concluded that the derivative of $y = c + s\ x$ would be $y' = s$. What, however, if those constants are not really constants, but change too, perhaps not as a function of x but as a function of some other variable ? Did we then really capture the proper change in y ?

13.1.2 Notation

Let $z = f[x, y]$. The partial derivative of f with respect to one of its variables while keeping the other(s) constant is denoted with round deltas. An application is:

$$\frac{\partial}{\partial x}(x^2 + 3\,y^3) = 2\,x \qquad\qquad \frac{\partial}{\partial y}(x^2 + 3\,y^3) = 9\,y^2$$

We collect these partial derivatives into the total derivative:

- The total differential of f is $df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}\,dy$ or the weighted sum of the differentials of separate variables weighted by their partial effects.

  For the expenditure on apples and bananas: $dZ = p_a\,dq_a + p_b\,dq_b$.
- The total derivative of f with respect to a single variable uses the differential on all variables. For example, with time t: $df / dt = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}$.
- With $x = x[t]$ and $y = y[t]$ then $f'[t] = \frac{\partial f}{\partial x}\,x'[t] + \frac{\partial f}{\partial y}\,y'[t]$.

  For expenditure on apples and bananas and constant prices we have 2D space $\{Z, t\}$ that works via other 2D spaces $\{q_a, t\}$ and $\{q_b, t\}$.

13.1.3 We already saw this

We have already seen an application of this in the product rule.

$$d\,(f\,g) = g\,df + f\,dg$$

Namely, take $F = f\,g$, then $F'[x] = \frac{\partial F}{\partial f}f'[x] + \frac{\partial F}{\partial g}\,g'[x]$ and $\frac{\partial F}{\partial f} = g$ and $\frac{\partial F}{\partial g} = f$.

It actually is used also in addition: $\boldsymbol{d}\,(f + g) = \frac{\partial(f+g)}{\partial f}\,f'[x] + \frac{\partial(f+g)}{\partial g}g'[x]$ where the partial derivatives only happen to be 1.

13.2 Applications

Karel Drenth at the Technical University Delft gave a course in transport science for students in operations management. When you are an operations manager and do not know some elementary physics and do not know what the engineers are talking about then you can end up in a great mess. Karel allowed me to translate his course into English and it is available on the web (TSOM). I find it a wise and engaging selection of real world issues that shows the power of analysis. Many of the *Mathematica* programs used in this present book I originally wrote to support Karel's course and textbook. It was a shock to hear that Karel suddenly died in 2010. He was a great person and wonderful teacher. TSOM does not explicitly employ partial derivatives so I will not use it here but if the topic interests you then it is a good exercise to spot where that method is actually relied upon.

Drenth, K.F. & Th. Colignatus (2000), "Transport science for operations management. Understanding elementary physics and mechanical engineering", http://thomascool.eu/Papers/TSOM/Index.html

13.2.1 Application to estimation

An instructive application of partial derivatives concerns linear regression. Remember the example in §10.9.3: we had a lot of observations in the $\{x, y\}$ plane, each observation represented by a dot, and we fitted a line through the cloud, selecting the line with minimal sum of squared errors: SSE = $\Sigma\, e_i^2$. Now, how exactly does this work ? To get to the details we need partial derivatives.

We can do this for a long list of observations but it suffices to use three, for example {2, 3}, {5, 6} and {7, 6}. Let us try to find the line $y = c + s\, x$ that fits those three points best.

The errors are $3 = c + s\, 2 + e_1$ and $6 = c + s\, 5 + e_2$ and $6 = c + s\, 7 + e_3$. Squaring and summing the errors gives:

$$\text{SSE} = \Sigma\, e_i^2 = (3 - c - 2\, s)^2 + (6 - c - 5\, s)^2 + (6 - c - 7\, s)^2.$$

A (sum of) square(s) is minimal when its slope is zero. So we take the derivative and set it to zero. But there are two variables c and s, so there arise two partial derivatives. Each partial derivative gives the contribution of an estimated coefficient to the total error. The total derivative will be zero when each partial derivative is zero by itself.

The two partial derivatives are:

$$\text{eq}(c) = \frac{\partial\, \text{SSE}}{\partial\, \text{c}} = -2\,(-c - 7\, s + 6) - 2\,(-c - 5\, s + 6) - 2\,(-c - 2\, s + 3)$$

$$\text{eq}(s) = \frac{\partial\, \text{SSE}}{\partial\, \text{s}} = -14\,(-c - 7\, s + 6) - 10\,(-c - 5\, s + 6) - 4\,(-c - 2\, s + 3)$$

Setting these to zero gives the so-called "normal equations" (also simplifying):

$$\mathrm{eq}(c) = 6\,c + 28\,s - 30 = 0$$

$$\mathrm{eq}(s) = 4\,(7\,c + 39\,(s - 1)) = 0$$

- The two parameters give us two equations and when we solve them then we find our estimates with minimal SSE.

```
Solve[{6 c + 28 s – 30 == 0,  7 c + 39 (s – 1) == 0}, {c, s}]
```

$$\left\{\left\{c \to \frac{39}{19},\, s \to \frac{12}{19}\right\}\right\}$$

- The result is the same when using the linear fit routine in *Mathematica*.

```
data = {{2, 3}, {5, 6}, {7, 6}};
model = LinearModelFit[ data, x, x] // Rationalize
```

$$\text{FittedModel}\left[\boxed{\frac{12\,x}{19} + \frac{39}{19}}\right]$$

$$\{\text{R}^{\wedge}2 \to 0.842105,\ \text{Correlation} \to 0.917663\}$$

In Chapter 11 we might have given this explanation for finding the best estimate. For example we could have fitted the proportional line $y = s\,x$ that has only one coefficient. But the focus in that chapter is on finding out what calculus actually means while regression is rather an application. Also, optimizing over one coefficient does not optimally convey how the method of least squares works. Instead, regression is a great showcase to convey and understand what partial derivatives are.

13.2.2 Application to taxation

13.2.2.1 Principle

In taxation there is a discussion what tax schedule the government should choose. A lump sum per capita or let it depend upon income, and how so ? Should high incomes pay a bit more (strong shoulders) or should they stop paying once they have contributed their share - and then how to determine that share ? One argument is that the rich benefit from social stability at least in proportion to their wealth or perhaps a bit more: if police and judge do not function then a rich person may lose more. A financial gain depends upon individual initiative but also upon social circumstance, as for example inventing an automobile and getting it sold on the market requires a state of technology, the existence of suppliers, consumers and market institutions. Such arguments about "should" and "deserve" have moral components. A somewhat more objective argument is based upon incentives and total revenue as a measure of the collective load. People take tax as a disincentive. A tax punishes earnings and a tax on income causes people to reduce work effort. The impact of this disincentive is larger for higher incomes because those incomes are higher. If the government wants to maximize total tax revenue then it might have to lower taxes for higher incomes to incite them to produce more taxable income. The government choice then depends upon behaviour and not morals. Behaviour again depends upon the utility that people derive from their net income and the leisure from not working.

See Seth Chandler's http://demonstrations.wolfram.com/TaxRatesAndTaxRevenue/

Let the taxpayer's complex personality be reduced to a utility function with work h and leisure 1 - h and net income depending upon the wage as $w\ h$ - T[$w\ h$]. We use the Cobb-Douglas form with one parameter a.

```
utility = (1 - h)^a (w h - T[w h])^(1-a);
```

The taypayer maximizes utility. The utility function is first rising in hours worked but after a while declines, so the first order condition for a maximum applies and we set the slope with respect to hours to zero. This generates a rather complex equation. Because of the chain rule the equation has a first derivative on the tax function, in other words the marginal tax rate. A simpler expression follows from solving for that marginal tax rate.

```
eq = D[utility, h] == 0;

Solve[eq, T'[w h]] // FullSimplify
```

$$\left\{\left\{T'(h\,w) \to \frac{a\,T(h\,w) - w\,(a+h-1)}{(a-1)\,(h-1)\,w}\right\}\right\}$$

In other words, the behaviour targetted at maximizing utility puts a restriction on solution space, and the taxpayer chooses h such that the marginal tax rate plays a role. When formulating tax policy the King of the Realm should pay attention to that rate.

When we assume proportional tax T[x] = $r\ x$ with constant marginal rate r then the hours worked reduce to the parameter in the utility function:

```
D[utility /. {T[x_] :> r x}, h] == 0;

FullSimplify[Result, Assumptions → {w > 0, 0 < a < 1, 0 < r < 1}]
```

$$(1-h)^{a-1}\,(a+h-1)\,(-h\,(r-1)\,w)^{-a} = 0$$

```
Solve[Result, h]
```

$$\left\{\left\{h \to 1 - 0^{\frac{1}{a-1}}\right\}, \{h \to 1 - a\}\right\}$$

For this utility function the marginal tax rate of a proportional tax drops out again. When all taxpayers have this utility then the government has little problem in optimizing revenue. The situation in reality is more complicated, notably with the wage depending upon hours worked. The example of the Cobb-Douglas utility function is useful to show the special case when the marginal tax rate drops out again but in practice that rate will play a role. The question that concerns us now is whether the tax schedule can be chosen with some sophistication.

13.2.2.2 A formal tax schedule

Let tax[y] be the tax schedule depending upon individual income y. The average is tax[y] / y and the tax on the marginal dollar is tax′[y]. Concepts are:

- the schedule is proportional if the average is constant: $\frac{d}{dy}(\text{tax}[y] \ / \ y) = 0$, thus tax[$y$] = $s\ y$ for a constant s
- the schedule is regressive if the average reduces with income: $\frac{d}{dy}(\text{tax}[y] / y) < 0$
- the schedule is progressive if the average increases with it: $\frac{d}{dy}(\text{tax}[y] / y) > 0$.

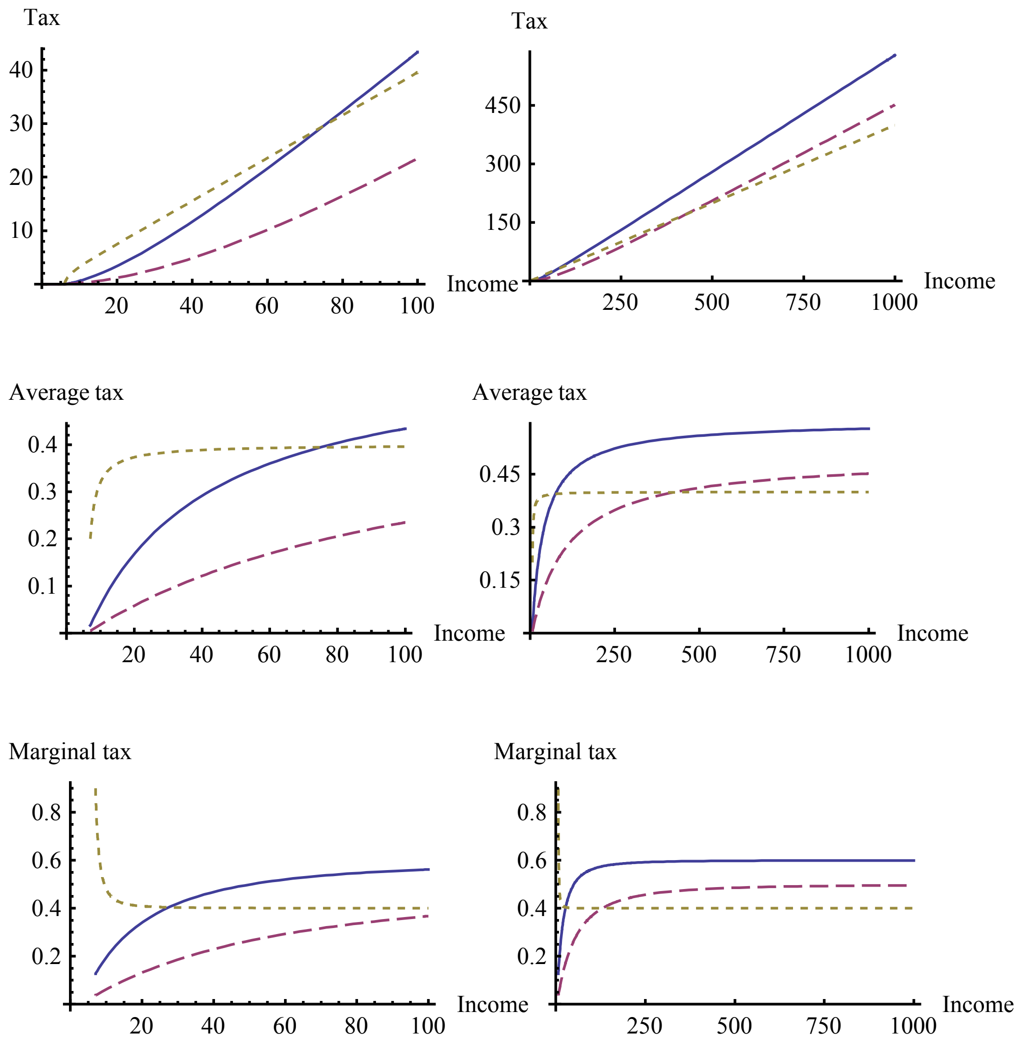
These concepts have only limited meaning since the rates may differ per income bracket. A tax schedule however tends to be implemented as a system that is uniform. Let us consider some common schedules.

The Bentham tax schedule (1) allows for an exemption level x, based upon the consideration that people first have to earn their own subsistence just to survive, and (2) there-after has a rate r. Thus Bentham[y, x, r] = $r\ (y - x)$ for $y \geq x$ and zero below it. This can also be formulated as a proportional tax minus a tax credit that then also could be given to those under subsistence: CreditTax[y, x, r] = $r\ y - r\ x$.

Governments tend to employ piecewise linear tax functions with brackets that build upon the Bentham schedule. This causes additional political debate about where to put those brackets. This kind of debate can be avoided in a general curved tax schedule like CurvedTax[y, x, r, c] = $r\ (y - x) * y \ / \ (y + c)$ with $y \geq x$ and parameters $\{x, r, c\}$. In this function exemption x follows from subsistence, r from the marginal rate for the highest incomes, and curve parameter c from the requirement on total tax revenue.

The graphs below may look a bit different for the domains [0, 100] and [0, 1000], expressed in \$1000 per annum. We consider three cases: curved tax with parameters {6, 0.60, 30} versus {6, 0.50, 100} versus {6, 0.40, -5}. With $c < 0$ there is a recovery from exemption towards a proportional rate.

- Legend: check the limit marginal rates in the bottom right graph. Solid line for {6, 0.60, 30}, course dashing for {6, 0.50, 100} fine dashing for{6, 0.40, -5}.



13.2.2.3 Dynamics and a constant average rate

What is generally lacking in this discussion about optimal taxation, and a reason why it is a good introductory example for partial derivatives, is an awareness about what happens to taxes when time passes. The discussion tends to assume that the schedule remains the same over time but it doesn't. The tax schedule generally gets adapted to inflation and economic growth, and thus the general rise of welfare over time.

Properties are: (1) Taxes tend to be progressive given the Bentham argument. (2) If the tax schedule is not adapted to inflation then this causes "inflation creep": people get taxed at ever higher rates since their income will rise with inflation. (3)

When national income rises also due to economic growth then the government would get an ever larger share of it and after a while a political party comes into power to reduce taxes to a normal level again. In sum, we can understand the process by looking at the total derivative. Use the curved tax form:

$$d\text{tax} = \frac{\partial \text{tax}}{\partial y}\, dy + \frac{\partial \text{tax}}{\partial x} dx + \frac{\partial \text{tax}}{\partial r}\, dr + \frac{\partial \text{tax}}{\partial c}\, dc$$

The change in the tax thus depends upon the change in the tax parameters too. There are some general properties here. In practice the parameter r will be kept constant over time since it reflects views on the limit marginal rate that is socially desirable for the rich. Exemption x rises with the general rise in welfare. Nowadays also the poor need a computer to participate in society. Parameter c rises with the need for tax revenue. In a simulation we would distinguish the general change of income from individual changes and generate a distribution of effects. For now it suffices to consider a representative case. The parameters x and c then tend to rise with the same factor as y.

Let us put these properties into formulas. When we have income y and parameters x and c in one year and next year the incomes rise to $p\ y$ due to inflation and economic growth, then a wise tax function of that new year will have parameters $p\ x$ and $p\ c$. Looking at the average tax we see that it does not change:

$$\text{avtax}(1) = \frac{r\,(y - x)}{c + y} = \text{avtax}(p) = \frac{r\,(p\,y - p\,x)}{c\,p + p\,y} = \frac{r\,(y - x)}{c + y}$$

What does this mean for the marginal rate, the disincentive to earn income ?

13.2.2.4 Dynamic marginal tax rate

To find the marginal rate we take the total derivative to time with r dropping out since we keep it constant:

$$d\text{tax}/dt = \frac{\partial \text{tax}}{\partial y}\, dy/dt + \frac{\partial \text{tax}}{\partial x} dx/dt + \frac{\partial \text{tax}}{\partial c}\, dc/dt$$

It is informative to divide all by dy/dt again:

$$\frac{\text{dtax/dt}}{dy/dt} = \frac{\partial \text{tax}}{\partial y} + \frac{\partial \text{tax}}{\partial x}\frac{dx/dt}{dy/dt} + \frac{\partial \text{tax}}{\partial c}\,\frac{dc/\text{dt}}{dy/dt}$$

and then dt drops out too. We could have done that directly but then we would have missed the element of time. Conclusion: when parameters are indexed on income then the dependence can be modelled to be upon y and not upon time. The total differential consists of a direct effect and the effect of the changes in the parameters.

$$\frac{\text{dtax}}{dy} = \frac{\partial \text{tax}}{\partial y} + \frac{\partial \text{tax}}{\partial x}\frac{dx}{dy} + \frac{\partial \text{tax}}{\partial c}\,\frac{dc}{dy}$$

With the proportional adjustment of x to y we have $\frac{dx}{dy} = \frac{p\,x - x}{p\,y - y} = \frac{x}{y}$ or that the marginal is the average. This expression can also be read as $\frac{dx}{x} = \frac{dy}{y}$ or $d$rex[x] = $d$rex[y] or that the rates of growth are the same. The same with c and hence:

$$\frac{\text{dtax}}{dy} = \frac{\partial\text{tax}}{\partial y} + \frac{\partial\text{tax}}{\partial x}\frac{x}{y} + \frac{\partial\text{tax}}{\partial c}\frac{c}{y}$$

This is the dynamic marginal tax rate under proportional adjustment. It gives the marginal tax rate but adjusted for the change in parameters when these are indexed on the rise of the overall level of income.

It should come as no great surprise that the dynamic marginal under proportional adjustment appears to be equal to the average marginal tax rate.

- The dynamic marginal can be equal to the average tax.

$$\textbf{tax} = \left(\mathbf{r}\,(y - \mathbf{x})\,\frac{\mathbf{y}}{\mathbf{y} + \mathbf{c}}\right);\ \frac{d\text{tax}}{dy} == \mathbf{D[tax,\ y] + D[tax\ ,\ x]} * \frac{\mathbf{x}}{\mathbf{y}} + \mathbf{D[tax\ ,\ c]} * \frac{\mathbf{c}}{\mathbf{y}}\ \mathbf{//\ Simplify}$$

$$\frac{\text{dtax}}{\text{dy}} = \frac{r\,(y - x)}{c + y}$$

13.2.2.5 Concluding

The following is a small example of how a dynamic marginal rate can equal a normal average. Let exemption be \$17000, and let the statutory marginal rate thereafter be 60%. Someone earning \$51000 pays a tax of \$20400, on average 40%. Let all incomes grow 5%, and exemption be indexed on national income. Then exemption becomes \$17850, income \$53550, and tax \$21420, again 40%. Thus on the (dynamic) "marginal dollar" this person doesn't pay 60% but 40%.

Consider national income Y and national tax revenue T. If the average tax level is to be constant then $T / Y = (T + \Delta T) / (Y + \Delta Y)$. A little algebra shows that $\Delta T / T = \Delta Y / Y$, or that tax revenue grows at the same rate as national income, and that $\Delta T / \Delta Y = T / Y$, or that the dynamic marginal tax rate equals the average tax. In balanced growth this would also hold at the individual level - and not just for a representative or average taxpayer since balanced growth means that all taxpayers participate in the growth. So you may be rich and have a high statutory marginal tax rate of 60% but if your average is 40% then, with general economic growth and subsequent wise adjustment of tax rates, your dynamic marginal will also be 40%. The impact of tax incentives will also depend upon this dynamic marginal or the average rate rather than only upon the statutory marginal.

Part V. Meta

Parts I - IV had a textbook layout and have been completed with the above. We now take the upper ground and look back at it.

1. Chapter 14: Euclid's *Elements* and axiomatic development are at Van Hiele's Level 3. The subject matter of points and lines is simple so it may well be that many 12-year olds can handle this level. In time there will be a computer game and we will see whether this indeed is true at what scale. However, to appreciate it as an axiomatization of space requires special attention at a wiser age. And how does it relate to non-Euclidean geometry ?
2. Chapter 15: The above has been composed with particular didactics in mind. At each point there might have been a footnote discussing the didactic turn, but this would clutter the text and block the free flow of the textbook format, while actually most didactic issues require an integrated discussion that is longer than a footnote.
3. Chapter 16: It will be clarifying to succinctly list the news in this book compared to other texts. You might check that first, to hone your mind to the points to be aware of. But it will be an advantage to first read the body of the text as if you are a novice and only later compare your own notes on new discoveries.

Obviously, the above only has a textbook *format* as this book is not quite a proper textbook. Exercises are lacking. Subjects could be dosed better along the OSAEP/I steps of Van Dormolen (see below). A textbook should also allow for the multi-dimensional intelligences of our students. The above has not been discussed in a school team nor has it been tested in class. We have not discussed hours. For 200 pages and 4 pages per hour, and 3 hours per week, the above might take 17 weeks. But all this is premature. As this book is a primer, we now set out for the real business of the didactics in the education of mathematics.

14. Axioms and reality

14.1 The axiomatic method

14.1.1 Laws

The Codex Hammurabi and the Bible's Ten Commandments created lists of *do and don'ts*. Euclid was a lawyer of space and his *Elements* applied that legal notion to geometry, refined the laws of space and presented the first axiomatic system in mathematics. It is up for discussion (1) whether axiomatics is a good teaching method (for geometry), (2) whether geometry still is the best subject for teaching about axiomatics and the deductive method. Propositional logic seems simpler and is directly concerned with deduction itself. Perhaps though arithmetic and algebra are better since they do not directly claim to be about deduction itself. Preferably all are used, to cater to the multidimensional intelligence on both language, symbol, number and space. We mention only a few points now.

14.1.2 An abstract system

By abstraction from reality we get a formal system, that differs from its (intended) interpretation in that no longer the semantics apply but only the syntax. The advantage of a formal system is that we are no longer distracted by hidden assumptions from our understanding of the problem area. All that is relevant to make something work is put in schemes that anyone can operate, even someone who does not understand the issue (like a computer). Let us take the subject of geometry with all its semantics as discussed above and let us try to create a formal axiomatic system for it. We then get an empty formal structure that we might interprete in various other ways too. In the axiomatic method we not only provide axioms and rules for deduction but we also must state a list of symbols and formation rules for which the axioms must hold. The relations just discussed are depicted in the following diagram. The situation actually is slightly more complex, since what isn't drawn is that we discuss these relations in a meta-language.

$$\text{(Intended) interpretation} \;\; \underset{\textit{Interpretation}}{\overset{\textit{Abstraction}}{\rightleftarrows}} \;\; \text{Formal system}$$

Once we have created this formal system there naturally arise a number of questions such as whether it really is a "good" system. Axiomatic theory has developed a number of criteria to judge on that "goodness".

The traditional method to "prove" the adequacy of an axiomatic system is to provide an existing example in the real world that forms a model for the system. Since the world is assumed to be consistent (there is only one reality at a time), a good fit would show that we have found a good formal model. It appears to be enlightening to analyze what we actually mean by "a good fit", since that generates all kinds of properties of systems that we may not have been aware of before.

DeLong (1971): "Our aim at formalization will be achieved if the informal theory (...) is an interpretation of the formal system." (p 106) and properties of a formal system are that it is "consistent, correct, independent, expressively and deductively complete, and decidable. (... and ...) may be made categorical" (p141).

14.1.3 Axiomatics versus enumeration

The axiomatic method differs from the enumeration method. The first uses only rules of substitution, expansion and contraction, that can be applied at liberty and that can deduce individual statements. The enumeration method lists all possibilities and checks them all, trying to design efficient algorithms. Given the infinity of possibilities in space Euclid's axiomatic method was magical in that it actually worked in dealing with those. Of course the magic is in choosing limited objectives that indeed can be attained. Nowadays with computers the enumeration gains in force, for example when the computer creates a graph so that we check its form. Analysis can be seen as already a step towards enumeration as we distinguish in kinds of functions (polynomial, exponential) and kinds of properties (intersection, extreme value). There is a (hidden) structural identity between these two methods, notably where the algorithm uses the same kinds of rules. There may be a difference though with respect to "finding new truths".

14.1.4 Axiomatics versus deduction in general

Given the (hidden) structural identity of the axiomatic method and the method of enumeration, it becomes a valid question why mentioning the axiomatic method at all. The point is that the axiomatic method still is the standard in mathematics for a proper definition of a system. This choice of standard clearly has its origin in *The Elements*.

That being said, this book takes a relaxed attitude towards axiomatics. It appears that the difference between the axiomatic method and a perhaps less formal but still deductive system becomes somewhat fuzzy. If we see the axiomatic method as merely substitution of truths in truths according to a truth-conserving rule then we are right to criticize this for neglecting solution strategies that reduce the time for a proof. Mathematical formalism as a goal in itself has little value as well. The objective of a proof is to convince a critical person and it may suffice that he or she recognizes the proof, as long as the method remains valid. In the methodology of science it appears that a surprising number of issues are not fully defined. Axiomatization of those issues seems overdone, though a bit more formalism sometimes helps. A useful deductive system, even not fully axiomized, still has the main properties of an axiomatic system, in that its terms and transformation rules must be defined somehow.

14.2 Non-Euclidean geometry

An important aspect for a science and its methodology is the relation between its definitions and the reality that those definitions (should) reflect. Creative minds coin definitions that maximize explanatory power. Yet, there are always these two aspects to be aware of.

PM. This present section only selects some passages from my discussion of an article on 100 years of Einstein, in *The Economist* January 1 2005. I am not a physicist, and after studying economics now for some 40 years I am close to understanding a tiny fraction of that field, so the following is with restraint. The connection is that geometry deals with measurement. Non-Euclidean geometry was discovered by Bolyai, Lobachevski and Gauss but we can link up to Einstein's application to physical reality.

Colignatus (2005b) http://129.3.20.41/eps/get/papers/0501/0501003.html

PM. My major book on economics Colignatus (2005a) has the very words "Definition & Reality" in its title.

A key issue in the theory of science is the issue of measurement. Physics before Newton suffered huge losses in intelligence, time and energy to discussions on unobservables and metaphysics. This in fact lasted partly into the 19th century with discussions on the 'ether'. Their solution was to put a stop to fruitless discussion and concentrate on what can be measured. You don't know what it is, but it moves this way, at that speed, and if you hit it here, then it moves there. This technical approach worked wonders, though it still seems that some theorists assume some 'whats' to derive their theories on the 'hows' (as Bohr's atom model).

Jan Tinbergen copied this more technical approach of measurement to economics, creating with Frisch and others the approach of 'econometrics'.

A key notion below will be that physics might 'overshoot' by concentrating on measurement and by neglecting definitions and logic. Econometrics is open to that same risk.

(PM 2011. Well, the world economy still suffers from the economic crash of 2007+.)

One way to understand what modern physicists often do, is that they, apparently within their philosophy of measurement, directly associate particles or waves with mathematical terms. This differs from the approach in economics where one starts with a theory and then develops hypotheses about measurable phenomena. Of course, many parts of physics may follow the latter approach too but apparently many other parts of physics follow that first approach that can generate confusion.

The Economist reports: "(…) Maxwell showed that it [light] consists of oscillating electric and magnetic fields. This immediately raised the question of what the fields were oscillating in. At that time, no one could conceive of waves which were not vibrating in some medium. The ocean has waves in water, and sound waves travel through air; it seemed nonsense to imagine that waves could just "be". (…) Lorentz (… derived …) that there was a contraction in the direction of the Earth's movement (…) Einstein realised (…) that there was no seem about it. Space was really contracting, and time was slowing down."

This, you will note, is a *non-sequitur*. It doesn't make logical sense. What Einsteins model does is to stop imagining what those waves oscillate in. Instead he focusses on the measurement results and makes these the absolute source of wisdom. This is not necessarily the best answer to the question what those waves oscillate in, since you might also develop a theory and deduce testable hypotheses. Einstein does deduce testable hypotheses but without a theory about what those waves "be". How can they exist without being something ?

Einsteins model subsequently seems to confuse the definition of space, given by the definitions of Euclid, and empirical space as measured by the instruments of physicists. (PM 2011. This links up with our discussion of geometry. Demonstrations of non-Euclidean geometry tend to be presented within our 3D world otherwise we have a hard time imagining them. Einstein apparently uses a model to get rid of measurement errors but this is something else than determining what space "really" is. When our mind forms an image of space, it rather depends upon the Euclidean definitions (Pythagoras), otherwise we don't know what we are speaking about. Those non-Euclidean models are imagined within this space.)

Modern physicists shy away from the possibility that space and time have independent definitions within the mathematical modelling of the world. They regard space and time as what they measure. However, they don't seem to see that they can be hopelessly confused when they measure speed in meters / second while those meters and seconds change under measurement. My impression is that it is better to accept measurement error and try to explain that error.

When physicists get weird readings, then there can be measurement errors and something may happen in interaction of their instruments with what they try to measure. If all instruments, and all the best of them, show the same measurement error, then there can still be such an interaction. Physicists, apparently within their philosophy of measurement, tend to conclude that reality is weird, with "space contracting and time slowing down". The proper approach would rather be to stick to the Euclidean definition of space (and time) as independent concepts that likely form part of the mind and the ability to think itself, and subsequently judge observations in those terms.

While Euclid's definition of space creates emptiness, it may well be that empirical space is filled with 'something' that allows oscillations. Presumably, electro-magnetism is the proof that such 'something' exists. There are reports that the 'void' would be able to produce particles. Also, there can be phenomena in that 'void' that appear to us as 'contracting' or whatever. All that is OK. But if you want to understand what space is, you would rather turn to Euclid where contracting is out of the question by definition.

It may also be that I simply don't understand what Einstein did. But then this article of *The Economist* really hasn't been clear enough.

There is a caveat for both sciences with respect to mathematics. There is a danger with mathematicians that they lose track of reality and the very aim of their research. Paradoxes like the liar paradox, the Russell set theory paradox, Gödel on his epi-phenomenon on the liar paradox, and the like generate confusion. Some solutions proposed by mathematicians are no deep mathematical results though many think so. Kenneth Arrow with his theorem on voting caused much havoc, since, though the math is right, his interpretation wasn't. Thus, it is difficult to strike a balance between mathematics and reality, and more awareness of this problem would help research. It might be wise to include more statistics in your programme of research.

14.3 Synthesis and analysis

Alongside model theory and an application to non-Euclidean geometry we should distinguish clearly as well: (1) the technical use of the terms *analytic* and *synthetic* for geometry (2) didactics, i.e. the study how people gain knowledge, (3) the branch of philosophy called epistemology that studies how man acquires knowlegde, with questions such as whether there might be a mathematical /

platonic world as apart from reality that can be found by intuition.

14.3.1 A technical term

Technically in solving a geometric question, the best approach is to assume that the problem already has been solved, to write down all that we know about it ('analysis'), sort it out, and the answer pops up so that the proof can be constructed leaving out inessentials ('synthesis'). Euclidean analysis is to reduce to earlier axioms and proofs. René Descartes did not create analysis but helped create a new *way* of analysis. Henceforth a main line in Western philosophy links up with the distinction between the ('synthetic') geometry of Euclid's *Elements* and 'analytic geometry' by Descartes (and Oresme and Fermat).

An excellent summary is given by professor Beaney: "According to Descartes, it is analysis rather than synthesis that is of the greater value, since it shows "how the thing in question was discovered", and he accuses the ancient geometers of keeping the techniques of analysis to themselves "like a sacred mystery"."

Michael Beaney: http://plato.stanford.edu/entries/analysis/s4.html

And: "The philosophical significance was no less momentous. For in reducing geometrical problems to arithmetical and algebraic problems, the need to appeal to geometrical 'intuition' was removed. Indeed, as Descartes himself makes clear in 'Rule Sixteen', representing everything algebraically - abstracting from specific numerical magnitudes as well as from geometrical figures - allows us to appreciate just what is essential (...) The aim is not just to solve a problem, or to come out with the right answer, but to gain an insight into *how* the problem is solved, or *why* it is the right answer. What algebraic representation reveals is the structure of the solution in its appropriate generality. (...) Of course, 'intuition' is still required, according to Descartes, to attain the 'clear and distinct' ideas of the fundamental truths and relations that lie at the base of what we are doing (...)"

The situation is a bit problematic. Intuitions are opposed to new techniques. In both cases there is analysis, the old geometers using their (trained) intuitions to decompose new propositions to basic propositions in *The Elements*, and Descartes decomposing with his new techniques. It is also hard to see how there can be synthesis without analysis, and analysis without synthesis. This fits Beaney's statement on Leibniz: "Furthermore, we can see how, on Leibniz's view, analysis and synthesis are strictly complementary (...) For since we are concerned only with identities, all steps are reversible. As long as the right notation and appropriate definitions and principles are provided, one can move with equal facility in either an 'analytic' or a 'synthetic' direction (...)". (And this fits the way how we set up integrals and derivatives.) The situation can be understood however as a mere pragmatic distinction between the use of the new techniques by Descartes, called anaytic geometry, and the old ways, called synthetic geometry, merely since synthesis is opposed to analysis. The new baby needs to have a name, after all. A bad name has the advantage of the need of explanation.

Hence, as a major conclusion: Descartes used a term in a particular sense that started a life of its own in Western philosophy; and we should not confuse any of this, neither Descartes nor philosophy, with the kinds of geometric methods.

14.3.2 Gain of knowledge

Pragmatically: adding to knowledge is a learning process. A great distinction then is between the Van Hiele levels 0 and 3 (see §15). Things can be experienced and intuited yet it requires some work to arrive at a deductive whole. Euclid and his time saw *The Elements* as a top result in the deductive method, properly so. Only the Kantian suggestion of 'synthesis' ≈ 'intuition' deviated from this, dubiously so.

The process of learning provides some cause to wonder where knowledge comes from. Van Hiele (1973) discusses the classic Piaget case of a child learning about height and volume, with glasses of different sizes where the same volume of water causes different heights. It is not clear whether the child uses its terms in the same meaning as we do. It is clear though that the child has intuitions about water, glasses, height and volume, whatever those intuitions might be, but they seem somewhat aligned with our own intuitions (and those of mine actually still are quite vague). When the child 'learns', the terms get more aligned (and possibly also the intuitions). All this makes sense in terms of the Van Hiele levels, and we should be very careful about philosophical interpretations.

As another major conclusion: it suffices to say that we *define* space with Euclid and his distance measure. We can be happy with that definition as a most economic summary of our notions and experience (applying economy to the handling of information). Thus, our mind entertains conceptions alongside reality, and those conceptions are only adopted because they are so powerful in dealing with reality.

14.3.3 Epistemology

The distinction between analysis and synthesis got a life of its own in the realm of philosophy and epistemology, i.e. the study of how man gains knowledge. Cartesius's method in *Discours de la Méthode* (of which *La Géométrie* was an appendix) uses four main rules, one of which is fundamental doubt, and here he concludes: "But I immediately became aware that while I was thus disposed to think that all was false, it was absolutely necessary that I who thus thought should be something; and noting that this truth *I think, therefore I am,* was so steadfast and so assured (...) that I might without scruple accept it as being the first principle of the philosophy that I was seeking." (p118-119). Cartesius judges that from pure thought follows existence: "Taking, for instance, a triangle, while I saw that its three angles must be equal to two right angles, I did not on this account see anything which could assure me that anywhere in the world a triangle existed. On the other hand, on reverting to the examination of the idea of a Perfect Being, I found that existence is comprised in the idea of a triangle that its three angles are equal to two right angles (...)" (p122). For us however there are two meanings of the word "existence": (1) proper existence in reality, (2) definitions created by the mind (to deal with that reality). Cartesius is at risk confusing those two. *Cogito ergo sum* is OK, but not everything cooked up exists in reality.

Subsequently a more problematic role is played by Immanuel Kant. Beaney: "Whatever criterion we might offer to capture Kant's notion of analyticity, the fundamental point of contrast between 'analytic' and 'synthetic' judgements (...) lies in the former merely 'clarifying' and the latter 'extending' our knowledge. (...) Kant writes that "To *construct* a concept means to exhibit *a priori* the intuition corresponding to it" (...) According to Kant, then, the whole process is one of *synthesis*. But the two activities mentioned here are both part of what the ancient geometers called *analysis* (...) What is remarkable about Kant's conception is the way that it has inverted the original conception of analysis in ancient Greek geometry (...) 'Analysis' is left with such a small role to play that it is not surprising that it is condemned as useless."

Kant suggested that our sense of space would be a "synthetic a priori", meaning that we arrive at knowledge about space merely because of the concepts involved and our ability to grasp those concepts. This is a difficult point. The Piaget / Van Hiele example shows that this 'grasping' does not happen by itself. It takes a lot of effort in fact, when we see how hard children are working each day. That being said, at one point most people do indeed seem to develop a notion of space. Above, my suggestion was that our definition of space actually is Euclidean, since when we imagine non-Euclidean space then this is within the confines of an Euclidean environment. Our definition arises in interaction with real space. But, once it has been formed, it is only proper to distinguish (1) the empirical experiences from (2) the mental construct that has arisen. The mental construct compactly summarizes earlier notions and experience but is not those themselves. Hence it is very much that "synthetic a priori" - but it is doubtful whether this is in the same sense as Kant had (and I have read only summaries and I do not have the time to read the apparently terse original). It is difficult to say that it 'extends' knowledge, since we have not defined how we measure knowledge. Merely saying that we have acquired a new notion is not sufficient since we already concluded that it summarizes and compacts earlier notions and experiences.

As a major conclusion: It is useful that we have (1) technical terms for geometry, (2) didactics, (3) model theory; and it is useful that epistemology helps us verbalize the wonder of it all; but it does not seem that epistemology has more to contribute than just that verbalization of our sense of wonder (other than confusion with the first 3 aspects).

15. Didactics

15.1 Didactics are relevant

Mathematics has developed over the centuries into a great building. The emphasis was not on didactics though. There are nooks and crannies, crooks and nannies, weird turns and windy windows. In teaching we must also teach to the standard so that students can read the relevant texts. However, teaching comes with a responsibility of its own. Sometimes re-engineering is in order.

15.2 The general approach

15.2.1 Aims of this book

This paragraph is only to remind you of the aims set out on the first page of the book and its abstract. This is a primer in the format of a textbook.

A common format for textbooks is that they state accepted knowledge. News is relegated to the journals, and it may percolate there a few decennia before it sinks down into the textbooks. This textbook contains basic accepted knowledge and then includes the news. Logically there is no harm in this approach since the news follows from those basics. The whole is presented as a textbook to show what it means, that it can be done, and that it has a great result.

15.2.2 The didactic approach

Learning goals are generally knowledge, skills and attitude. In this book there are hardly exercises while attitude hardly comes from reading a book. Hence we are focussed on knowledge. The didactics then are guided by the Van Hiele levels:

- Level 0: visualization and intuition
- Level 1: description, sorting, classification
- Level 2: informal deduction
- Level 3: formal deduction

Van Hiele (1973:177) gives the following example: (0) An isosceles triangle is recognized like an oak or mouse are recognized. (1) It is recognized that the triangle has the property of at least two equal sides or angles. (2) Relations between properties are recognized: at least two equal sides iff at least two equal sides. (3) The logical reasons for these relations are considered: why *if*, and what does it mean to reverse an implication ? Van Hiele (1973:179): "At each level we are explicitly busy with internally arranging the former level." (my translation)

Van Hiele (1973:179) on geometry: "At the base level we consider space like it appears to us; we can call this *spatial sense* (like *common sense*). At the first level we have the geometric spatial sense. (E.g. measuring degrees of an angle / TC.) At the second level we have mathematical geometric sense; there we study what geometric sense involves. At the next level we study the mathematical logical sense; it then concerns the question why geometric manners of thought belong to mathematics." (my translation) Importantly, at each level the same words may be used but with different intentions, complicating mutual understanding.

The levels do not provide information about the boundaries of topics, and they are not strong when it comes to finalizing a topic and switching to a next one (that builds upon the earlier). In this book we spend most time in Level 1 and 2, and there are some patches that peek into possibilities for Level 3 - but then for various subjects that some teachers might rather see as subjects of themselves. For us, geometry is a Level 0 for moving towards Level 1 and higher in analytic geometry. It may be that a reader picks up some properties on isosceles triangles that move the reader up to level 2 in geometry on isosceles triangles, but for us that would be a happy circumstance and not our prime target. Chapter 1 on geometry does contain some deductions but these merely rekindle what is known, wet the appetite and set the stage.

In moving from one level to the next, Van Hiele (1973:149+) identifies phases: (1) intake of information (examples), (2) bounded orientation (direct instructions), (3) explicitation (making explicit, verbalization in own words of what is known), (4) free orientation (extending the relationships in the network), (5) integration (summarizing and compacting what has been learned, often old fashioned learning). Van Dormolen recognizes similar stages: Orientation, Sorting, Abstraction, Explicitation, Processing & Internalisation (OSAEP/I). A teacher using this book would have to dose these.

We reject Freudenthal's "realistic math" in its more extreme interpretation. This is best discussed in separate paragraphs.

15.2.3 It hinges on what counts as experience

Van Hiele and Freudenthal overlap in the starting point in experience. The question remains what kind of experience we choose. Working in the plane itself is seen by Freudenthal as too abstract while Van Hiele in principle allows the notion that it might be experience too. Mental thought is an abstract process by nature and we can have experience in that.

Modern research on the brain clarifies many aspects of mental processes. Operational definitions of thinking and consciousness however cannot replace the definition of thinking as experienced by the conscious self. When we look for a definition of what thought is, in that experience of being conscious, then we quickly arrive at a Platonic version of ideas. In the mind's eye a triangle has a purity about it that is not caught in any drawing. Also mudd becomes perfect mudd. There is no difference between an image of a triangle and an image of mudd, or even an image of a sunset, in the sense that they are constructed out of the same mental elements that can only be pure. It are these mental ideas that education deals with, and experience in reality is only a tool to reach them. This does not mean that we have to be full Platonists in assigning an indestructible and immortal quality to these ideas. Thought and thinking, consciousness and awareness, are primitive notions for the thinking intellect itself, and up to this day and age of human history they do not generate any additional information for more conclusions than their very experience.

15.2.4 What we can assume and build upon

Students have sufficient experience with the plane since making drawings in kindergarten. When they think about a triangle it is as abstract as it can get because such thought is abstract by nature. We can draw many triangles on paper but the notion of a triangle in the mind is an entirely different matter, and when the student thinks about a triangle then it is that notion that is in the mind and not the drawing on the piece of paper. What counts are the lingering notions in their abstract imagination that have to be activated. When we put labels to angles on paper and draw supporting lines then we use paper images to enter new concepts into the mind. It remains an essentially abstract activity, with pen and paper only tools for communication. It distracts and confuses when mental clarification is mixed with the application to reality. Application to reality is relevant but should be dosed wisely.

15.2.5 Finding the proper dose and perspective

My book *A Logic of Exceptions* maintains that the force of logic derives from reality. If a truck approaches and if you do not jump aside then it will hit you. Mimicking this, *A Logic of Exceptions* starts with electrical switches to clarify the constants of propositional logic. In this case we do not need to explain these constants since we presume that students already know them. We only help making them explicit. The empirical examples are only intended to highlight the properties and to pave the road towards formalization. Here the electrical switches do not distract since the case is not presented as an exercise in building electrical circuits. The examples help to focus on the logical properties. Electrical switches are as good an example as language, and in a way a better example since the focus in logic is already so much on language that it helps to provide another angle.

For analytic geometry it may be argued that a bucket and a faucet that adds a liter per minute would be a similar good starting point. This is dubious however. If the objective is to distinguish linear processes from other processes then indeed examples in reality are the stepping stones, but that is another issue than linking up with geometry. The example distracts from the very abstract notion that we want to establish. "Realistic math" might require a student to spend a sizeable part of the lesson time on realistic examples trying to figure out what is the point. When supporters of "realistic math" argue that students of geometry do not understand a linear process without such examples as the bucket, then the reply is that those teachers have not spent sufficient effort in providing the abstract tools to perform the mental process.

It are different mental processes: imagining a bucket and faucet and imaging a graph of a linear function. The bucket and faucet have been learned in kindergarten. The graph and its geometric interpretation first have to be learned before they can be imagined and linked up to the bucket and faucet. Once we have the graph then it is OK to say, and indeed we ought to say, as this book does, that the bucket and faucet are an interpretation and application, and only then there can be that flash of understanding that shows that the link has been achieved. Once an aspect of the plane has been conquered then abstract understanding can be easier related to those other cases from reality, which means that those other examples are relevant for the Van Dormolen Processing & Internalisation stages. But first we must develop the geometry of that graph, using the mental images of geometry itself.

In the same vein this book is hesitant with respect to angles and trigonometry. We first get a firm grounding in the horizontal and vertical axes before we introduce trigonometry. Perhaps this has to do with my own experience that economic research relies on lags to generate a business cycle so that angles are nowhere to be seen unless you start deriving them. Students in electrical engineering and music might want to start out with angles right away. In the current set-up angles and trigonometry appear when they start making sense in the story as it develops, and hopefully that is a convincing argument for the little wait.

15.2.6 The challenge

There is a challenge though. Eudlid's *Elements* and his axiomatics have been the standard for more than two millenia. They are at Van Hiele's highest level. Perhaps 12-year olds can deal with those abstractions, as they actually are rather simple. But it becomes a bit different when we try to incorporate the advances in analytic geometry and calculus. Here are concepts that better be developed at a lower level and Van Hiele then wins from Euclid. Here Freudenthal steps in and resorts to the richness of reality, and at first that seems like a golden solution. Indeed, axiomatic geometry is at Level 3 and not at Level 0 ! However, as explained Freudenthal's approach is not convincing since it neglects that thought is abstract by nature. Rather than going sideways into reality we should focus more on the processes of thought and thinking itself.

15.2.7 A missing link

We should provide for an abundance of words and concepts in the abstract plane, so that the student has enough to hold on to for visualization and intuition. A missing link in geometry appears to be that those anchors are rather absent.

When you visit a new city then you tend to like it when the streets already have names. Suppose that you would be forced to invent your own labels, like "that crooked street with the blue shop" and then hope that other people understand you. Current textbooks on geometry send out students to conquer the plane but present it as a verbal desert, without conceptual guidance other than the x and y co-ordinates. The Van Hiele Level 0 requires them to visualize and to activate their intuition, yet that also requires a richness of words and concepts - that currently are lacking. Euclidean geometry has a poverty of points and lines that can intersect, be parallel or overlap: and though it is a great exercise in logic it must be admitted that Freudenthal has a point that Euclid's approach is not so appealing to the average student over the last two millenia. Conventional analytic geometry is an improvement since drawings are supported with formulas, and vice versa, yet again, its richness is only developed over time, and at the Level 0 and 1 there still isn't much to visualize and intuite and verbalize.

In particular, it will be useful to extend the plane with a nomenclature of "named lines". Chapter 4 opens with them and then builds up - see there to check what this means. A quick reply will be that we already have names, such as $x = 1$, $x = 2$, ... for vertical lines for example. Those names derive from a formal development however. Instead we rather first create standard names that fit the experience with the plane. This will provide the fertile ground, where the coin can drop when experience is morphed into abstract understanding.

It may be argued that it is fairly simple to draw a line and determine the starting value on the vertical axis and its slope. Exercises and realistic examples then provide for learning. However, experience shows that students later have difficulty with the horizontal and vertical lines. Why a line works as it does tends to remain elusive. A conclusion is that it is better to start with named horizontals en verticals and then awaken the motivation that a general formula will be useful.

Thus the didactic suggestion here is that the notion of "named lines" can be the missing link that resolves the issues in the choice between dropping Euclid and moving towards analytic geometry and calculus (and not just Descartes but along the lines of Van Hiele). The notion of these named lines caused the very layout of Chapter 4 on lines and subsequently from there the layout of the whole *Conquest of the Plane.*

Pierre Marie van Hiele argued most of his life (May 4 1909 - November 1 2010) in favour of the use of vectors already in elementary school. Though he has been greatly valued for his ideas on the didactics of mathematics, he never succeeded in overcoming the opposing views. Vectors even appear late in highschools. The missing link suggested here is hopefully helpful. Logically, if this is indeed the missing link that has been provided only now, then teachers seem to have been right in resisting Van Hiele's suggestion, since the picture is complete only now. Alternatively, the suggestion of named lines is not really a missing link and only one of the possible bridges, and we are underestimating the capabilities of pupils and students all over the board.

Clearly, the proof of the pudding is in the eating, and only empirical testing will show whether students indeed learn faster following the approach presented here. If this book would be mistaken, and "realistic math" would still be needed to propel the more practically minded students, then, the lame argument becomes, it would suffice to include it in this book as well, and the advantage of this book would remain to be its logical order and novel concepts.

15.2.8 A longer discussion

See *Elegance with Substance* for a longer discussion on the didactics of mathematics.

A Logic of Exceptions has much more historical discussion than this present *Conquest of the Plane.* This has the - historical - explanation that logic has been besettled with paradoxes while planar geometry has been relatively calm since the issue of non-Euclidean geometry could be settled a little more than 100 years ago. It is a point of attention for a possible next edition that good stories may be included here too at one point of time.

A final point of note is that I do not have clear ideas about what would motivate a 12 or 14 year old kid to be interested in analytic geometry and calculus. Van Hiele (1973) rightly remarks that students and pupils hardly can be motivated for what they learn since they do not know yet what they will learn. A common ground is that man is a curious ape and cherishes the flash of insight. Mathematics is a language and it can be fun to learn a new language and a new world. Paul Goodman (1962, 1973) *Compulsory miseducation* remains sobering though. While *Conquest of the Plane* concentrates on knowledge the didactic setting naturally is a complex whole, in which motivation plays a key role, and it is mandatory to keep that in focus.

15.3 Proportions

15.3.1 The issue

Proportion space is presented in *Elegance with Substance* (2009:21). A proportion (ratio) is a point in proportion space. A proportion (ratio) 1 : 2 associates with the point {2, 1} in proportion space. The fraction 1 / 2 is the slope of the ray through {2, 1} and the origin. These concepts are well-defined. Proportion (ratio) is two-dimensional while fraction is a number on the real number line (also found at x = 1 intersecting with a ray). EwS is a bit short and implicit but in the body of this present book explicit definitions and clarification on proportion space are given.

H. Pot (2009abcd) calls attention to the confusion in education on proportions and fractions. He lists examples in teaching material – even for the training of elementary school teachers – where it is suggested that proportions and fractions would be the same, while they are not.

In the Dutch TAL project, see KNAW (2003) and Freudenthal Instituut (2009), it is stated for an audience of students who want to become primary school teachers: "Fractions, percentages, decimal numbers, and proportions are different descriptions of something that we can regard *in some respect* as the same." (my translation from Pot (2009d) and my italics). The italized qualifier makes this alright (though vague) but the surprise is that Pot reports that the authors did not want to go into detail "because of the targetted readership". This is surprising since we definitely would want primary school teachers to understand the distinction between proportion and fraction.

The issue adds support to the conclusion on the need for re-engineering math education. This concerns not just improved education for elementary school teachers (at least in the Netherlands), but we have to deal with the Euclidean legacy in our culture and language, and we have to deal with the two-dimensionality (the Van Hiele argument on vectors).

15.3.2 History

Fractions are important. The Egyptians had only fractions ½, 1/3, 2/3 and ¾, and their mathematics stagnated for 3000 years on this.

The distinction between proportions and fractions is influenced heavily by history. Pythagoras assumed that all phenomena could be measured in ratio's of natural numbers and was horrified when one of his students showed that this was impossible for $\sqrt{2}$ (the student apparently didn't survive this). Henceforth, the Greeks solved their problem by focussing on geometry where such quantities can be dealt with, with some ingenuity. By consequence, the Greeks developed a theory of geometric proportions and did not develop the theory of arithmetic and subsequently algebra.

Another consequence is that the Greek theory of proportions entered the textbooks on mathematics, likely to stay there forever, even though there is now a better theory of arithmetic. Instead of just referring to slope and fraction, textbooks on mathematics also employ proportion and if proper explain that Euclid intended the ray itself rather than just the slope.

To understand the distinction between proportions and fractions we thus need this historical overview, where (1) the theory of proportions is a more primitive but intellectually more complex theory while (2) arithmetic is more powerful and develops (a) number theory and (b) division (giving fractions) as the inverse of multiplication. When these two theories (frames of reference) are not distinguished then confusion enfolds.

It is a moot question whether it is wise to teach these two theories and their history in primary school indeed. But proportions are endemic in our language and culture, and relevant when we scale items up or down, and thus it seems that the issue cannot be avoided.

PM. With the burden of history, we cannot avoid tradition and convention. The mathematical symbol π is defined on a circle as the ratio of the circumference to the diameter. Radians are the arc divided by the radius. This follows the convention of using the ratio instead of concentrating on the number on the number line. This is no problem once it is obvious that a ratio can be projected into a number. But it may be confusing when that is not understood. There is a discussion of a dimensionless number or ratio versus the idea that when we measure we always use a unit, see §15.4 below.

15.3.3 Theory

In itself the distinction between one and two-dimensional might not quite hold. Babylonian numbers (degrees, minutes, seconds) might be seen as a somewhat multidimensional phenomena as well. But in standard theory we start with natural numbers, and then include zero, negatives, fractions (rational numbers) and irrationals to create the number line.

There is a problem with the concept of "same". Since the point {5, 3} is not the same point as {10, 6} we may wonder whether 3 : 5 and 6 : 10 are the "same" proportions. For fractions it is standard that 3 / 5 = 6 / 10 are identically the same number, but for the two-dimensional ratio's we need the expression 3 : 5 :: 6 : 10 to show equivalence, or "equality other than pure identity".

Thus, the primitive but complex theory of proportions (frame of reference) comes with a notion of equivalence or equality that differs from identity. Two line sections of the same length need not be identically the same line sections but their lenghts would be identically the same. In the more advanced theory of arithmetic this is replaced by a notion of equality that is identity.

These (theory-dependent) definitions for proportion, fraction, identity and equality (other than identity) are pure theoretical developments and suffice for theory. Next there is education, and education may require additional terms for the communication between teacher and student.

15.3.4 Education

Pupils at elementary school tend to learn about proportions and fractions from cutting up pies and cakes. Adding up fractions can become an intricate matter in that manner. Their understanding might be helped by having access to the graphs in proportion space. Van Hiele has explained that pupils at elementary school can already deal with vectors and co-ordinates. Who sees the display in proportion space for dealing with fractions will tend to agree that we should not withhold it. Thus much of our discussion on analytic geometry would in the future be done in elementary school. One supposes that "evidence based education" can clarify what kids can handle. However, such research needs to be subtle. We need research that does clearly distinguishes the different aspects.

Of course, the addition of slopes is different from the addition of vectors, and thus the pupils better grow aware that it matters what labels are on the axes. Thus it is not wise to add {1, 2} and {1, 3} to {2, 5} in proportion space. If it is introduced then it seems necessary to introduce it at the same time with a vector diagram, to prevent future confusion.

It is useful to remind here that Van Hiele (1973:196-204) is rather convincing on the suggestion that elementary school spends too much time on fractions (and too little on vectors).

15.3.5 Terms

The words we use ought to be well-defined. The body of the text has been extended with good definitions for proportion and fraction. These definitions are also important for a good understanding of the dynamic quotient and its consequences for calculus.

Thus the educational terms must be used so that they do not create new confusions. Pupils and students must be made aware of these aspects: form, action, result.

There are aspects that some regard as calculation but that are algebra. Possibly calculation is algebra anyway. For algebra, it is useful to manipulate y / x or y // x as a form. If it is thought of as a number only then it might be hard to see how it can be manipulated. For example $(2\pi) / \pi$ is a fraction of two irrationals and thus would seem to be irrational too, but algebra shows that it is rational. Such manipulations are important. Thus it remains useful to distinguish both form and number. But only as aspects to consider the same expression y / x.

Some pupils or students will regard the decimal expansion 0.5 as the number so that the fraction 1 / 2 is only a form, or instruction for calculation. This issue has been discussed in the body of the text above. *Mathematica* does the same, with N[ ] the operator for the change to the decimal system (the line at denominator = 1 in proportions space). Indeed, educational simplicity for primary school arises when this approach is adopted. At that level of educational simplicity we might indeed forget about the distinction between proportions and numbers and only use proportion or ratio 1 / 2 and number 0.5. We then only have numerical equality and forget about identity. The only problem might be that textbooks in primary education use different definitions than textbooks in secondary education and up.

PM. Its seems that (1) "ratio" is a verb in the Greek theory of Proportions, and (2) "dynamic quotient" is a verb in the theory of arithmetic. Perhaps there is didactic advantage in translating or projecting the historical distinction into a distinction into verb and noun. This is not a question that we can resolve just here. Perhaps this book should have adopted the term "ratio" instead of "dynamic quotient" but that would have occluded the historical meanings with the new interpretation with respect to seeming division by zero and its link to calculus.

15.3.6 Choosing axioms in proper primitives

Pot (2009abcd) suggests to take proportionality as a basic concept, primitively understood by the human mind, and to subsequently develop the notion of number from there. This is converse to the now standard approach to take number as basic and subsequently develop proportionality. However, when the confusion has been resolved by better understanding the historical development of the rational and irrational numbers, and the distinction between the theory of proportions and the theory of arithmetic, then there is no need to change the standard approach.

15.3.7 Conclusion

Proportion and fraction (number) are well defined, and have their related but own theories. A highschool graduate and by implication a primary school teacher should be able to understand and reproduce these theories and definitions. It remains a choice what is taught at what phase in education. Given good didactics and following Pierre van Hiele, it should be possible for a large section of the population in primary school to learn both vectors and the dimensional distinction between proportion and fraction (number).

15.4 Trigonometry

15.4.1 Trig rerigged

The presentation of trigonometry in this book is based upon my paper *Trig Rerigged* (2008). The advantage of this book is that trig is discussed in the context of analytic geometry, while the few years that have passed have allowed a fresh look and some angles to be polished up. Personally I found it rewarding to now reformulate the cosine rule as the key theorem of analytic geometry. Putting that at the center, after Pythagoras and the introduction of the co-ordinates, clarifies to the student much better that a sound formulation for trigonometry is key in the conquest of the plane.

Another key step in this present book was to select the capitals $\{X, Y\}$ for the unit circle as opposed to $\{x, y\}$ on the plane in general. This small step only built on the earlier choice of Xur and Yur but it did simplify notation and enhanced clarity overall, for example for the expression of the derivatives. I found it still enlightening myself to present x_{ur} and y_{ur} from the bottom up, as know-how still differs from do-how.

The paper *Trig rerigged* combines both formulas and didactic evaluation. In this section we will only do the latter since the formulas are in the body of the book. Sometimes the text here may come across as unsubstantiated since there is no direct support in formulas. However that substance has been provided above and you are referred to the specific elaboration in the body of the text.

To repeat: Independently, Bob Palais also judged the selection of π over Θ to be a historical error.
See Palais, R. (2001a), "π Is wrong!", The mathematical intelligencer, Vol 23, no 3 p7-8.

15.4.2 Abstract of Trig Rerigged

Didactic issues in trigonometry concern the opaque names of sine and cosine and the cluttering of questions with π or 360 whereas a simple 1 suffices. The solution is to use the 'unit turn' or 'unit of measurement (meter) around' (UMA) as the yardstick for angles. This gives the x_{ur} and y_{ur} functions for the $\{X, Y\}$ co-ordinates on the circle with unit radius. The relevant mathematical constant is $\Theta = 2\pi$ (capital theta, reminiscent of a circle) rather than π and it comes into use much less when we use UMAs instead of radians. The sine and cosine remain relevant for the derivative. The common term 'dimensionless' appears to confuse 'no unit of measurement specified' (with a metric, in planimetry and trigonometry) with 'no dimension' (a pure number, in number theory).

15.4.3 Establishing the relevance of trigonometry

The trigonometric functions and their properties are presented in this book only when the need arises. They are not a subject that lives by itself and for which some day the relevance might suddenly appear. In this book, the cosine rule is used, and only implicitly, when we establish the key theorem of analytic geometry. Only then it gains a lease on life. The proof that Sin at 0 has slope 1 is only used when it appears relevant to calibrate the derivatives. As trigonometry can create typical formulas and concepts, this gradual approach appears very attractive. The general idea is the conquest of the plane, not how to handle waves.

15.4.4 Angles and arcs

Using radians to measure angles is economic in terms of concepts but appears a setback in terms of didactics. It is useful to speak about angles and arcs as separate aspects. Angles are the pointy bits and turns around, and arcs are those curves. The problem of measuring the pointy bits is solved by using turns, also expressed in an arc measure, but the latter does not obliterate the terms and concepts involved. Hence, this book uses angles α and β and arcs φ and ψ.

15.4.5 The order of presentation

1. The traditional approach takes angle ≈ arc φ as primitive, introduces degrees, Cos and Sin, then the co-ordinates x and y, and as a third step generates the inverses φ = ArcCos[x / r] and φ = ArcSin[y / r].
2. We start with geometric angle α and analytic geometric co-ordinate X_v and our second step is φ = ArcX[X_v], so that X_v = Cos[φ] is the inverse. Since vectors v differ from arcs φ there will be no confusion in writing $X_v = X_\varphi$ for Cos. There may be confusion between arc φ and angle α so $X_v = \mathbb{X}_\alpha = \mathbb{X}[\alpha] = x_{\text{ur}}[\alpha]$ = Xur[α] uses another symbol. This abundance of symbols helps emphasizing the aspect that is relevant at a point of discussion: dependence of X_v upon the vector, of $\mathbb{X}_\alpha$ upon the angle, range of $x_{\text{ur}}[\alpha]$ upon the unit circle, and the procedure of calculation Xur[α] e.g. within *Mathematica*.

15.4.6 Cause and effect

Functions are written as $y = f[x]$ and for definitions $f[x]$:= *procedure with x.* Then:

1. The traditional method is to define Cos = x / r. Later extended into Cos[φ] = x / r. There however is no φ on the right hand side. This turns cause and effect around. This is not a proper definition but an equation to solve.
2. The proper relation is $x = r$ Cos[φ]. Given an arc we calculate the co-ordinate.

Perhaps historically tables for Cos and Sin were constructed on the measurement of angles and sides of the triangle yet this is not a valid argument for reversing causality or reason.

15.4.7 Sloppiness is never good. Do not underestimate students

Traditional books and presentations on the internet are often a bit vague about the meaning of Cos and Sin. They may say it but do not hammer it down or forget about it. They even write Cos[x] and Sin[x] so that when you write x / r = Cos[x] as you thought to have learned then you get stuck. The proper expressions are $x = r$ Cos[φ] and $y = r$ Sin[φ] but some books think that you can learn about Cos and Sin but are unable to learn the Greek alphabet.

15.4.8 Information and algorithm

1. A definition of Cos[φ] = x / r is also uninformative as to what to do. OK, now I have defined a "Cos". What, in all goodness, is it ? Students tend to get sums with x and r to make the division x / r which they learn to call Cos, which is essentially an exercise in arithmetic. In the unit circle $r = 1$ and x = Cos[φ] so there is nothing to divide. If you have such an x then saying that this is Cos is only another way of saying that you have that x. That $r = 1$ cannot be seen as an exception since the unit circle is the very place where the action is, the circle of all ratios. It is a defining element in understanding what you are doing.
2. In this book, if you have x and r then you use φ = ArcX[x / r]. Cos is not needed, you can directly use the arc function on the co-ordinate. The division x / r is a short intermediate step of normalization, which step you understand because you have the proper definitions; and the renormalization disappears directly in the function input call. Possibly, as an intermediate step, you write the equation $x = r$ Cos[φ] and solve for φ. There is no logical need for this intermediate step, but you may do this to relate what you learned in this book to what you read on the internet.

To recognize that x is, or requires, Cos is only relevant in current practice since you must find the proper inverse (since that name is ArcCos that has Cos in it). But: you only need to do that because Cos differs in name from the x that you already have. There is no X in the name of Cos. And ArcCos had no X in its name either. To shorten the path of all that coding and decoding, X_φ and ArcX are definitely faster.

15.4.9 Directionless Cos and Sin

The prime didactic question with respect to Cos and Sin concerns their relevance for any angle, with whatever direction, versus their role in the triangles fixed by the unit circle. If students learn to associate them with any direction, then they must unlearn and readjust towards co-ordinates. If they learn to associate them with co-ordinates then they must unlearn and adjust to arbitrary direction.

1. Traditionally, Cos and Sin are introduced without a system of co-ordinates, for right triangles in any direction. We can use that ratio to expand from a triangle and given a larger radius we can calculate the other sides.
2. This book avoids the introduction of Cos and Sin without a system of co-ordinates. For triangles with arbitrary direction we have not sought to calculate angles and sides. We wait with doing so till the proper tools have been developed. With a system of co-ordinates we feel better equiped to introduce specific names since now we can orient the triangle in the proper position towards axes that have names.

The choice is difficult to decide on. What would be best might transpire in a randomized controlled trial, perhaps running over several generations. We would also need to define what is 'best' and how to balance the spatial sense with the handling of the co-ordinates.

However, it is the impression of this author that the spatial sense is basic, so that the tradional order of presentation of triangles without co-ordinates is best. This approach is followed in this book by having the first chapter on geometry in directionless space. This is the Van Hiele Level 0. The first step after that however are the angles counted by turns (UMA) and in the second step there are x_{ur} and y_{ur}. Starting with triangles in any direction does not mean that Cos and Sin have to be introduced at that stage. For proportional expansion there is no need for the specific names of Cos and Sin, since proportions will do by themselves. Only the mix of angles and sides causes their use, but, that is a very specific learning target that must be evaluated within the context of the whole, and then it drops in importance; even becomes counterproductive.

There is no strong objection, though, to the traditional method of using Cos and Sin for right triangles of arbitrary direction. For that matter the Hypotenuse, Adjacent and Opposite sides can be used, including the SOHCAHTOA rule (sine equals opposite over hypotenuse, cosine equals adjacent over hypotenuse, tangent equals opposite over adjacent).

Note though that it should be AHCOHSOAT because of the order of both cause and effect and x and y. Preferable also AHXOHYOAT for x_{ur} and y_{ur} alongside. I would prefer to use Radius and Slope too so that we arrive at ARCORSOAs or, finally, proper ARXORYOAS. (Adjacent divided by radius equals Xur, opposite divided by radius equals Yur, opposite divided by adjacent equals slope.)

It needs to be considered though whether learning a rule like that differs essentially from reorientating the triangle so that it fits in a system of co-ordinates.

In a textbook for 3rd year advanced highschool co-ordinates are used already in chapter 1 while trig is discussed much later. There seems little loss and only gain to include orientation.

Both methods could be good practice in understanding what the orientation of a triangle actually is. Triangles would have to be rotated correctly anyway. There is only so much that you can do at a certain age. For example, the plot on proportionality uses a two-dimensional function: if you would want to explain proportionality in elementary school then you likely don't have those concepts. Nevertheless we should beware of thinking that deep exercises in geometric trigonometry are required before making the step towards analytic geometry.

Still, once the co-ordinates have been introduced, the names of the cosine and sine functions do not link up to the already known expressions for the horizontal and vertical axes, i.e. the x and y values. Students have to calculate these x and y values but when the instructions use the names Cos and Sin then they are not explicitly told to calculate these co-ordinates. They are asked to calculate some weird sounding names that seem as something completely different. Surely, textbooks have a line that explains that Cos and Sin are the values on the unit circle. Subsequently though all work is done in those names. Students then observe the Morning Star and the Evening Star, thinking that those are different, out of the blue that it is the same old Venus. We can blame students for forgetting about that single line of clarification. It will be much and much better to use functions Xur and Yur defined on the unit circumference circle and that range on the X and Y values of the unit radius circle.

Hence, if you intend to use Cos and Sin at a lower educational level then it is a good investment for later courses to use Xur and Yur instead (even when you do not use a system of co-ordinates).

15.4.10 Quadrants

In the traditional approach the awareness of the quadrants is relatively weak. It may not matter much since the whole exercise is already burdened by squeezing a 2D problem into 1D functions (or having it 4D, in the mapping from the Euclidean plane to the polar plane). The procedure in this book is to stick to the standard plane as long as possible and this really drives home that those curious Cos and Sin only render co-ordinates *and* that you have to check the quadrants.

15.4.11 The xur or cosine rule

1. The traditional approach first derives the cosine rule for the addition of arcs and uses this to define the multiplication of vectors. This approach is conceptually centered on the polar co-ordinates. Arcs are seen as important

while it is neglected that the student is still struggling with co-ordinates.

2. Here we take the standard system of co-ordinates as basic. We only investigate the role of addition of arcs once the issue of vector multiplication arises. Trigonometry gets only introduced once the need for it arises and that need can be understood.

In terms of results it does not not matter where you start but it is thought here that working with X_v better relates to the system of co-ordinates than Cos[φ] (even though Cos[φ] is only an opaque way of writing X_v).

There seems to be an argument that angles, Cos and Sin are analytically more basic than the vector notation, see the generation of the complex numbers. This is however a matter of preference on how to build the argumentation on conquering the plane. Defining something and then giving an existence proof, as done here, may be judged to be more elegant and transparant than following the way how everything was historically discovered and put together.

15.4.12 Measuring angles and arcs

The units of measurement of angles are degrees (max 360) or radians (max 2π) or grad (max 400) instead of a clear 1 (unit unspecified) or 1 unit of measurement (unit specified).

The conventional measures are ratios and obscure the point that the angle (pointy section of the plane) is de facto measured by length of arc in some system of measurement. For length there has already been defined a standard, namely the meter, so why not use it again for the circumference ? An *angular circle* with a circumference of 1 meter better clarifies that we are measuring length. The unit of measurement then is 'unit meter around' (UMA). This can be made dimensionless as a 'turn' (as a fraction of that maximal unit length around) or as 'unit of measurement around', where a turn is one unit.

The traditional approach makes mathematical courses more tedious than necessary for understanding angles. The π needlessly clutters the argument in two ways. Students struggle to find the values $2k * 3.14...$ on their ruler while it would be more convenient to use 1 for the full circle around and with $k = 0, 1, 2, ...$ there is nothing to multiply. Secondly, if a fraction or multiple of π is to be used at all, it is more convenient to use $\Theta = 2\pi$.

See §9.2 on the measurement of Θ for the discussion on the unit of measurement and dimension (-lessness). Here and now is a good place to extend somewhat on the notion of dimensionlessness.

The discussion is a bit complex since notions in number theory tend to derive their names from the theory of space, so that it may be hard to keep distinguishing the two. We should distinguish the distance in space, that provides a metric in space, from the 'metric' that may be defined on a set of pure numbers (0, the next 0', the next 0'', ...). The 'metric' for pure numbers can be based upon a calculation scheme $||z_1 - z_2||$ and the metric in space follows from our experience and conceptualization of space. It is 'analytic geometry' to associate the two. Note though that 'association' implies that there are two different realms and not necessarily only one. It is a bit amazing that the fundamentals of analytic geometry still clog up the didactics of geometry and trigonometry but hopefully we achieved clarity.

The continuity in measures derives from their spatial extension. Using a ratio is a mathematical simplification, eliminating the need to construct a unit circle, but does not affect the notion that lengths are involved. When we divide a meter by a meter, or 1 m / (1 m), the dimension drops out, but if we look at the meter that we have just measured it still is that 1 meter.

These 'dimensionless' numbers or ratios cause an epistemological question. We distinguish reality from the human mind. It might be that reality is only granular and that continuity is an illusion created in the mind, in the same way as, concerning time, the 'now' is a construct of the mind for the ephemeral border, or actually only logical border, between 'past' and hypothetical 'future'. It is more conventional however to assume that space in reality is continuous and that we create measures in number theory to mimic this property of space. Thus we distinguish crude figures and lengths on paper from the pure figures and 'dimensionless' numbers in the mind. On paper we may take a unit of measurement (say, the rod in Paris) but in the mind there is no place for such a physical object. In sum, in planimetry, 'dimensionless' stands for 'no unit of measurement specified'.

The notion of 'dimensionless' number interpreted as 'no dimension' remains epistemologically dubious when we relate this to the measurement of length, basically on paper and subsequently in the human mind. For, how could it be that these 'no dimension' numbers are nicely ordered and apparently have a distance metric such that e.g. halves are twice as distant as quarters ? Where does the notion of continuity come from ? In practice we assume that space is continuous. Apparently, there is a subtle distinction between 'no dimension' and 'unspecified dimension (unit)'. Apparently, the mind thinks about space with unspecified dimensions and not quite without dimensions. This is similar as drawing a line on paper and arbitrarily affixing 0, 1, 2, ... numbers along it, with the numbers at (approximately) the same distance, and writing down that these are meters while in fact they will be something else, with the true metric defined on the spot. Imagining triangles, circles and line sections in the mind, we must admit that they all have some apparant 'size', albeit 'size in the mind', all in (some) proportion to the other things that we may imagine for comparison. Thus the 'dimensionless' numbers in trigonometry still reflect length, with a space metric, albeit with unspecified unit.

The latter is an important didactical conclusion. Some mathematicians tend to think that trigonometry deals with 'dimensionless' ratios (apparently meaning 'no dimension' as in number theory) and not with length, and they seem to deny that the notion of 'ratio' implicitly has a metric and that this metric is related to the notion of length itself. This present discussion suggests to bring the implicit relation out into the open by explicitly referring to length and the 'unit meter around' or 'unit of measurement around'. Students in trigonometry then learn to switch between actual length with a specified unit and length without a specified unit (ratios). This would be an advance in clarity compared to the current practice where ratios are defined and where it is suggested that we are not measuring length but merely calculating 'no dimension' numbers just like in number theory.

The basic point is that our topic of interest here is space, with its figures and angles. In trigonometry the space metric is a priori, and abstract numbers without dimension (and the number 'metric') support the analysis, but cannot replace that notion of a metric contained in the notion of space. Admittedly, number theory can have its own origins. Possibly we start counting on our fingers and toes and then apply the same technique to spatial distance. But the experience that walking 50 kilometers is more tiring than walking 10 meters, and other experiences with space, need not depend upon counting. Indeed, in number theory we can define a set of numbers without dimension, and there we can define a 'metric' $||z_1 - z_2||$ on those numbers, but this 'metric' is not a metric as in space (real or in the mind). Perhaps it works the other way around. We can take ratios, i.e. express lengths as multiples of a standard unit length, and we can abstract from the space metric to also create such a set of pure numbers, and then what works for space can also be reflected in those numbers.

A reader objected to the use of the UMA, categorizing it as part of "realistic education in mathematics" as advocated by Freudenthal, and arguing that this kind of education is damaging to the development of mathematical skills and abstract thought. Only the abstract 'no dimension' interpretation was considered proper. This reader then does not see the limitations to the 'no dimension' view, as explained above. The objection also came as a surprise to me, since I had no intention of introducing Freudenthalian 'realism' (indeed with its current excesses). If it is possible to see the present suggestion as belonging to that Freudenthal approach then it is a mere coincidence and I actually cannot vouch for that. And curiously, another reader tends to see some value in the Freudenthal approach. The point however is that this book only wants to clarify how angles can be best dealt with.

Thus, a basic notion of analytic geometry is that there is a distinction between 'no dimension' (number theory) and 'unspecified unit' (space). When this is accepted then there can be no objection when a 'turn' is translated as 'unit of measurement around'. There exists a circle with unit circumference, and its distances have the exact values as turns around. The question then centers on whether we should specify what the unit of measurement is. The SI unit is the meter. My suggestion is not to hide that the unit of measurement might well be that SI unit. What this book proposes is that students become capable in translating specific measurements into a bit more abstract mathematical constructs and vice versa. Since there is length on an arc involved, it is required that UMA is mentioned. This will help the student to understand what trigonometry is about. Not mentioning UMA, not explaining what an angle is, withholding the evidence, will hinder the development of abstract thought.

PM. See *A Logic of Exceptions* for a rejection of Cantor's theorem on the power set because of its nonsensical application of selfreference. In addition to that: the diagonal argument on the decimal numbers also conflicts with the very manner how they are created. Thus $\mathbb{N}$ and $\mathbb{R}$ may be as large, and we should not express this argument on continuity as a distinction between $\mathbb{N}$ and $\mathbb{R}$.

15.4.13 Derivatives

We find that $\frac{d}{d\alpha} x_{ur}[\alpha] = - \Theta\ y_{ur}[\alpha]$ and $\frac{d}{d\alpha} y_{ur}[\alpha] = \Theta\ x_{ur}[\alpha]$ because of the scale factor. When the expression becomes more involved then the additional coefficient does not matter much in added complexity. Nevertheless, where we first saw a reduction in mention of Θ the derivative appears to be a place where it will show itself more often. This property implies that radians and Cos and Sin defined on the radians will remain in use, especially for heavy users of calculus.

If one presents Cos and Sin first then it must be explained to students that it would have been feasible to first present x_{ur} and y_{ur} and then present $\text{Cos}[\varphi] = x_{ur}[\varphi / \Theta]$ and $\text{Sin}[\varphi] = y_{ur}[\varphi / \Theta]$ as the scaled versions with the sometimes more attractive derivatives - but that this order of presentation has not been chosen, for whatever reason Cos and Sin are presented first.

It was an option for this book to systematically employ Xuc and Yuc instead of Cos and Sin. But there is a risk of confusion with Xur and Yur, while Cos and Sin are the overwhelming standard in books, calculators and internet. The present approach has an optimal ring to it.

15.4.14 In sum

In the traditional approach we see needless arithmetic, an overcomplexity in relabeling, a reversion of cause and effect, and a misdirection as to where the place of the action is. We shouldn't forget though that mathematics still is a crowning jewel of human achievement.

Current trigonometry is needlessly torturing our students. The torture derives mainly from conventional thinking and not from the math itself. So students arguing for a more transparent trigonometry have math on their side.

15.5 Calculus

15.5.1 The derivative is algebra

Improving the logical base of calculus on the issue of 'division by zero'.

The history of this text is as follows. *A Logic of Exceptions* considers the logical paradoxes. Retyping it in 2007 caused me to consider the paradoxes of division by zero too, out of a sense of completeness. There still was some lingering doubt with respect to the lectures in Analysis that I attended as a student back in 1974 and the Weierstraß construction for derivatives and continuity. In logic there is the difference between implication and inference, and inspired by the difference in economics between statics and dynamics as ways of analysis I had already in 1980 classified logic into static implication and dynamic inference. Hence in 2007 the dynamic quotient was born. The paradoxes of the derivative and the approach discussed here already got a section in *A Logic of Exceptions* in 2007. A longer paper with the present title *The derivative is algebra* of July 2007 is on my website. (Later this was linked to the difference between verb and noun in general - and in 2007/8 in a course on the didactics of mathematics I discovered that Gray & Tall had developed the term "procept".) It has been polished up and appeared as chapter XI in *Elegance with Substance* 2009. The chapter can be reproduced here with little additional comment except for the points that the main body of this *Conquest of the Plane* (1) improves on the derivative of the exponential function, (2) extends with the derivatives of Cos and Sin, (3) contains § 2.3.3 and 2.3.4 with extensive definitions for the process of division, (4) does not discuss the relative error that is crucial and is discussed below. You miss some references to pages in *Elegance with Substance* but the relevant concepts like the distinction between verb and noun are also in this present book.

The text here is intended for mathematicians, since the creation of the dynamic quotient and its application to calculus are a novel contribution to mathematics. The text is also intended for teachers as it clarifies the difficulties in teaching calculus. The text is not intended as an introduction to calculus for students since that is presented in the body of this book. While the text below develops the mathematical theory it has been a challenge indeed to compose an introduction for students from the bottom up. It is satisfactory to see that it indeed can be done and that calculus in this manner finds a natural place with analytic geometry. The cost is that the introduction above does not discuss the notion of the relative error yet, which is explained in the text below. It plays a role in judging on algebraic form. Reviewing the whole I am again impressed by the contributions of our great mathematicians who allow us to take this journey.

15.5.2 Abstract

Calculus can be developed with algebra and without the use of limits and infinitesimals. Define y / x as the 'outcome' of division and $y // x$ as the 'procedure' of division. Using $y // x$ with x possibly becoming zero will not be paradoxical when the paradoxical part has first been eliminated by algebraic simplication. The Weierstraß $\varepsilon > 0$ and $\delta > 0$ and its Cauchy shorthand for the derivative $\lim_{\Delta x \to 0} \Delta f / \Delta x$ are paradoxical since those exclude the zero values that are precisely the values of interest at the point where the limit is taken. Instead, using $\Delta f // \Delta x$ and then setting $\Delta x = 0$ is not paradoxical at all. Much of calculus might well do without the limit idea and it could be advantageous to see calculus as part of algebra rather than a separate subject. This is not just a didactic observation but an essential refoundation of calculus. E.g. the derivative of $|x|$ traditionally is undefined at $x = 0$ but would algebraically be sign[x], and so on.

15.5.3 Introduction

Since its invention the zero has been giving trouble. Mathematicians solved the paradoxes by forbidding the division by zero. But the problem persisted in calculus, where the differential quotient relies on infinitesimals that magically are both non-zero before division but zero after it. Karl Weierstraß (1815-1897) is credited with formulating the strict concept of the limit to deal with the differential quotient. However, he did not resolve the paradoxical aspects.

Regard these expressions, three well-known and the fourth a new design.

1. The difference quotient $\Delta f / \Delta x = (f[x + \Delta x] - f[x]) / \Delta x$ for $\Delta x \neq 0$. Note that one would see this as a result and not as a procedure.
2. The differential quotient or derivative $f'[x] = \boldsymbol{df} / \boldsymbol{dx} = \lim_{\Delta x \to 0} \Delta f / \Delta x$.
3. The current "theoretical true meaning of the derivative" with outcome value L: $\forall\, \varepsilon > 0\ \exists\, \delta > 0$ so that for $0 < |\Delta x| < \delta$ we have $|\Delta f / \Delta x - L| < \varepsilon$.
4. The new suggestion: $f'[x] = \boldsymbol{df} / \boldsymbol{dx} = \{\Delta f // \Delta x$, then set $\Delta x = 0\}$. This means first algebraically simplifying the difference quotient, expanding the domain with 0, and then setting Δx to zero. NB. "//" is defined in §2.3.4.

Let us consider the various properties.

15.5.4 The old approaches

The theory of limits is problematic. The limit of e.g. x / x for $x \to 0$ is said to be defined for the value $x = 0$ on the horizontal axis yet not defined for actually setting $x = 0$ but only for x getting close to it, which is paradoxical since $x = 0$ would be the value we are interested in. Mathematicians get around this by defining a special function $f[x] = x / x$ with split domain but this requires a separate $f[0] = 1$ statement, while it is faster to write $x // $x.

Also, the interpretation given by Weierstraß can be rejected since that definition of the limit still excludes the value (at) $\Delta x = 0$ which actually is precisely the value of interest at the point where the limit is taken. This is a conceptual inconsistency.

While the Weierstraß approach uses predicate logic to identify the limit values, the new alternative approach uses algebra, the logic of formula manipulation.

Fermat, Leibniz, Newton, Cauchy and Weierstraß were trained to regard y / x as sacrosanct such that it indeed doesn't have a value for $x = 0$. They worked around that, so that algebraically y / x could be simplified before x got its value. While doing so, they created a new math that appeared useful for other realms. These new results gave them confidence that they were on the right track. Yet, they also created something overly complex and essentially inconsistent. Infinitesimals are curious constructs with no coherent meaning. Bishop Berkeley criticized the use of infinitesimals, that were both quantities and zero: who could accept all that, need, according to him, "not be squeamish about any point in divinity". The standard story is that Weierstraß set the record straight. However, Weierstraß's limit is undefined at precisely the relevant point of interest. 'Arbitrary close' is a curious notion for results that seem perfectly exact. When we look at the issue from this new algebraic angle, the problem in calculus has not been caused by the "infinitesimals" but by the confusion between " / " and " // ".

The present discussion can be seen as reviving the Cauchy approach but providing another algebraic interpretation that avoids the use of 'infinitesimals'. The impetus comes from the notion of the dynamic quotient in algebra. We cannot change properties of functions but we can change some interpretations. Undoubtedly, the notion of the limit and Weierstraß's implementation remain useful for specific purposes. That said, the discussion can be simplified and pruned from paradoxes.

Struik (1977) incidently states that Lagrange saw the derivative as algebraic. See there for details and why contemporaries thought his method unconvincing.

15.5.5 The algebraic approach

In a way, the new algebraic definition is nothing new since it merely codifies what people have been doing since Leibniz and Newton. In another respect, the approach is a bit different since the discussion of 'infinitesimals', i.e. the 'quantities vanishing to zero', is avoided.

The derivative deals with formulas too, and not just numbers (as conventionally). It uses both that $\Delta f \,//\, \Delta x$ extends the domain to $\Delta x = 0$ and that the instruction "set $\Delta x = 0$" subsequently restricts the result to that point.

Since we have been taught not to divide without writing down that the denominator ought to be nonzero, the following explanation will help for the proper interpretation of the derivative: first the expression is simplified for $\Delta x \neq 0$, then the result is declared valid also for the domain $\Delta x = 0$, and then Δx is set to the value 0. The reason for this declaration of validity resides in the algebraic nature of the elimination of a symbol, as in x // x = 1, and the algebraic considerations on 'form'.

The true problem is to show why this new definition of ***df* / *dx*** makes sense.

15.5.6 Stepwise explanation of the algebraic approach

Let us create calculus without depending upon infinitesimals or limits or division by zero.

1. We distinguish cases $\Delta x \neq 0$ and $\Delta x = 0$, and the (*) implicit or (**) explicit definition of relative error $r[\Delta x]$.
2. Let $F[x]$ be the surface under $y = f[x]$ to the horizontal axis from 0 till x, for known F and unknown f that is to be determined (note this order). For example $F[x] = x^2$ gives a surface under some f and we want to know that f.
3. Then the change in surface is $\Delta F = F[x + \Delta x] - F[x]$. When $\Delta x = 0$ then $\Delta F = 0$.
4. The surface change can be approximated in various ways. Of these $\Delta F \approx y\, \Delta x = f[x]\, \Delta x$ is the simplest expression with explicit y. (Alternatives are e.g. $\Delta F \approx f[x + \Delta x]\, \Delta x$, or inbetween with $\Delta y = f[x + \Delta x] - f[x]$, $\Delta F \approx \Delta x\, (y + \Delta y /2)$.)
5. The error will be a function of Δx again. We can write ΔF in terms of $y = f[x]$ (to be found) and a general error term $\varepsilon[\Delta x]$, where the latter can also be written as $\varepsilon[\Delta x] = \Delta x\; r[\Delta x]$ where $r[\Delta x]$ is the relative error. When $\Delta x = 0$ and thus $\varepsilon[\Delta x] = 0$ then the relative error can be seen as undefined and it can be set to zero by definition.
6. We have these relations where we multiply by zero and nowhere divide by zero or infinitesimals.

	(*) *Implicit definition of r*	(**) *Explicit definition of r*
$\Delta x \neq 0$	$\Delta F = y\, \Delta x + \varepsilon[\Delta x]$	$r[\Delta x] \equiv \Delta F \,/\, \Delta x - y$
$\Delta x = 0$	$\Delta F = 0 = u\, \Delta x + \varepsilon[\Delta x]$ for any u; select $u = y$	$r[\Delta x] \equiv 0 = u - y$ for $u = y$

7. Simplify $\Delta F / \Delta x$ algebraically for $\Delta x \neq 0$ and determine whether setting $\Delta x = 0$ gives a defined outcome. When the latter is the case, take u as that outcome.
8. Thus $u = \{\Delta F // \Delta x$, then set $\Delta x = 0\}$. (Setting a to a value b is denoted as $a := b$.)
9. We then find $u = y = f[x]$ which can be denoted as $F'[x]$ as well.

For example, the derivative for $F[x] = x^2$ gives $\boldsymbol{d}F$ / $\boldsymbol{d}x$ = {(x + Δx)^2 – x^2) // Δx , then Δx := 0} = {2x + Δx, then Δx := 0} = 2x. This contains a seeming 'division by zero' while actually there is no such division.

The selection of $u = y$ is based upon 'formal identity'. This is a sense of consistency or 'continuity', not in the sense of limits but in the sense of 'same formula', in that (*) and (**) have the same form (each seen per column) irrespective of the value of Δx. By this choice the form is not affected by the value of Δx.

The deeper reason (or 'trick') why this construction works is that (*) evades the question what the outcome of ε[Δx] // Δx would be but (**) provides a definition when the error is seen as a formula. Thus, (*) and (**) give exactly what we need for both a good expression of the error and subsequently the 'derivative' at Δx = 0. The deepest reason (or 'magic') why this works is that we have defined $F[x]$ as the surface (or integral), with both (a) an approximation and (b) an error for any approximation that still is accurate for Δx = 0. When the error is zero then we know that $F[x]$ gives the surface under the $u = y = f[x] = F'[x]$ which is the function that we found. There is no approximation but exactness.

In summary: The program is $F'[x]$ = $\boldsymbol{d}F$ / $\boldsymbol{d}x$ = {ΔF // Δx , then set Δx = 0}. The definitions (*) and (**) give the rationale for extending the domain with Δx = 0, namely form.

(PM 2011: Select e.g. Δ$F \approx y^2$ Δx as the approximation. Then (*) suggests form $u = y^2$. But (**) has form ΔF / Δx - y^2 and in ΔF / Δx there is no suggestion of a square so the choice $u = y^2$ is problematic. The relative error features as a criterion because it allows an identification of ΔF / Δx as a separate form, and an identification of its outcome as the y that we are looking for.)

15.5.7 Implications

The proper introduction to calculus is to start with a function that describes a surface and then find the derivative. Since we only use equivalences, this also establishes that the reverse operation on the derivative gives a function for the surface.

The relation to the slope only arises in point (4) above. Traditionally the derivative is created from the question to find the slope at some point of a function. This tradition also suggests a separate development for the integral, e.g. with Riemann sums. This traditional approach tries to be as 'simple' as possible. However, it makes things more complex. Instead, here we find that the slope comes as a fast corollary – seeing that ΔF // Δx would be the tangent if it is defined.

Let us look closer into the difference between starting from slopes or from surfaces.

The derivative of $|x|$ is traditionally undefined at $x = 0$ but would algebraically become sign[x]. For $x \neq 0$, we can consider the various combinations and find the normal result, sign[x]. For $x = 0$ the dynamic quotient gives $(|0 + \Delta x| - |0|) \,//\, \Delta x = |\Delta x| \,//\, \Delta x =$ sign[Δx]. Setting $\Delta x = 0$ gives 0. Hence in general $|x|' =$ sign[x].

The traditional approach to $|x|$ is a bit complicated. Cauchy naturally gives 0 at 0 too. Traditionally the derivative is used for finding slopes and then the amendment on Cauchy was to hold that the right derivative differs from the left derivative, hence traditionally there is no general derivative. However, there is a multitude of 'tangent' lines at 0, that is, when tangency is not defined as having the same slope as the function (which slope seems undefined at 0) but as having a point in common that is no intersection.

In our approach, when we are interested in slopes, then it remains proper to consider these left and right derivatives. We do not need to speak about limits but merely can point to the different values of the derivative sign[x] in the intervals $(-\infty, 0)$, $[0]$, $(0, +\infty)$. Depending upon the definition of 'tangent': (a) "Tangent" lines that have the point {0, 0} in common without intersection then can have slopes from –1 to 1. (b) "Tangent" lines that have the same slope as the function however have only the three slopes –1, 0, 1.

The dynamic quotient is the leading impetus here and the issue starts with algebra so that slopes come in only second. $|x|$ is the surface under some function f. Any approximation of changes in the surface, when the surface value is $|0| = 0$, finds a perfect answer with zero relative error by requiring $f[0] = 0$. The general function appears to be sign[x]. The choice to extend the domain of Δx with value 0 at $x = 0$ derives from a notion of consistency of the form of the relative error in the approximation. This is sufficient though not necessary. One could argue that the relative error is not defined when $\Delta x = 0$ but this runs counter to our choice to define it as 0. This choice again relates to the form of the relations in step (6).

15.5.8 Students

Generations of students have been suffering. Teachers of math seem to have overcome their own difficulties (mainly by stopping to think) and thereafter don't seem to notice the inherent vagueness.

Students not only suffer from the vagueness but also from the notation . Many forget to write "lim($\Delta x \to 0$)" as the first part of each differential quotient, each separate line again and again for each step of the deduction, assuming that stating it once should be sufficient to express that they are taking the limit. Some 'take the limit' so that for them Δx has become 0, and then, just to be sure, they still mention "… + Δx" arguing that it should not matter when you add 0. Those 'official mathematical errors' will be past.

Conversely, if the new notation of dynamic division is adopted also for general purposes, then the algebraic origin of the derivative will be sooner recognized, strengthening the insights in logic and algebra. Time can be won for more relevant issues.

Teachers may be less tempted to distinguish between 'those who know the truth' (Deep Calculus, the ε and δ) (who thus actually are wrongfooted) and 'those who only learn the tricks' (Superficial Calculus).

Didactics remain an issue. Above nine steps are somewhat elaborate while the short program {ΔF // Δx , then set $\Delta x = 0$} sums it up and suffices. Possibly some randomized controlled trials in education would bring more light in the question what explanation works where.

15.5.9 The chain rule

The chain rule is an important result and can found directly as follows.

$$\boldsymbol{d} f \,/\, \boldsymbol{d}x = \{\Delta f \,//\, \Delta x, \text{ then set } \Delta x = 0\}$$

$$= \{\Delta f \,//\, \Delta g * \Delta g \,//\, \Delta x \text{ for } (\Delta x = 0 \Leftrightarrow \Delta g = 0), \text{ then set } \Delta x = 0\ \}$$

$$= \{\Delta f \,//\, \Delta g, \text{ then set } \Delta g = 0\} * \{\Delta g \,//\, \Delta x, \text{ then set } \Delta x = 0\}$$

$$= \boldsymbol{d}f \,/\, \boldsymbol{d}g \; * \; \boldsymbol{d}g \,/\, \boldsymbol{d}x$$

15.5.10 The derivative of an exponential function

[NB. Added comment in February 2011: The key deduction is improved upon in the main body of the book in §12.1.8.3, notably by moving from the dynamic quotient to the surface identity. The text of this paragraph can remain here for the didactic aspects.]

The derivative of an exponential function follows from the chain rule and the presumption that $\exp[x] = e^x$ is the fixed point in differentiation:

$$\frac{\partial\, a^x}{\partial x} = \frac{\partial\, e^{x\,\mathrm{rex}[a]}}{\partial x} = e^{x\,\mathrm{rex}[a]}\ \mathrm{rex}[a] = a^x\,\mathrm{rex}[a]$$

The reasoning thus is:

(i) All functions can be expressed as an exponential function for any nonnegative base number b, as $f[x] = \exp[b\ , \mathrm{rex}[b\ , f\,[x]]]$.

(ii) We presume that in this class of all possible bases there is a fixed point in differentiation. Call this base the number e. Thus by definition $(\,e^x\,)' = e^x$.

(iii) We can calculate e from the property $e^x \equiv \boldsymbol{d}e^x \,/\, \boldsymbol{d}x = \{e^x\,(e^h - 1) \,//\, h, \text{ set } h = 0\}$.

This gives $1 = \{(e^h - 1) \,//\, h$, set $h = 0\}$. By setting $(e^h - 1) = h$ and solving $e = (1 + h)^\wedge(1/h)$ we find the approximate value of e by taking h close to zero.

(iv) That there is an actual number e with 'infinite accuracy' follows from (iii) and from notions of continuity ('there are no holes between 2 and 3').

(v) From the chain rule we find in general rex[a] = $\{(a^h - 1) \,//\, h$, set $h = 0\}$.

Thus, the dynamic quotient $(a^h - 1) \,//\, h = (e^{h\,\mathrm{rex}[a]} - 1) \,//\, h$ does not simplify easily. However, when we use the chain rule then we can avoid using this explicit expression and actually find its value by implication.

Some meta-comments are:

a. The number e remains an algebraic concept like the number π .

b. The procedure to first presume e and its property, and only then calculate / approximate it, and thus prove its existence by calculation, summarizes an intricate historical development, but does not invalidate the existence proof.

c. In this case approximate values for e are found as we would normally take a limit. But the limit is not applied for the derivative.

d. The notion of a limit by itself still has its advantages, e.g. for the limit to infinity, and thus for 1 // 0 again. It would not be right not to mention limits in education.

e. There remains a distinction however between algebraic simplification and extension of the domain on the one hand and the traditional concept of a limit on the other hand. This distinction causes the insight that the derivative is an algebraic notion rather than dependent upon infinitisemals.

f. Given that limits can be defined in acceptable manner suggests that calculus can be developed by using limits. Indeed, complex ways can be used for what is simple.

15.5.11 Conclusion

History is a big subject and we should be careful about drawing big historical lines. But the following seems an acceptable summary of the situation where we currently find us after the historical introduction of the zero.

The introduction of the zero in Europe around AD 1200 gave so many problems that once those were getting solved, those solutions, such as that one cannot divide by zero, were codified in stone, and pupils in the schools of Europe would meet with bad grades, severe punishment and infamy if they would sin against those sacrosanct rules. Tragically, a bit later on the historical timeline, division by zero seemed to be important for the differential quotient. Rather than reconsidering what 'division' actually meant, and slightly modifying our concept of division, Leibniz, Newton, Cauchy and Weierstraß decided to work around this, creating the concepts of infinitesimals or the limit. In this way they actually complicated the issue and created paradoxes of their own.

The Weierstraß $\varepsilon > 0$ and $\delta > 0$ and the derivative's shorthand $\lim_{\Delta x \to 0} \Delta f \,/\, \Delta x$ are paradoxical since those exclude the zero values that are precisely the values of interest at the point where the limit is taken.

Logical clarity and soundness can be restored by distinguishing between the (formal) act of division and the (numerical) result of division. Using $\Delta f \,//\, \Delta x$ and then enlarging the domain and setting $\Delta x = 0$ is not paradoxical at all.

The distinction between static and dynamic division suggests that the Weierstraß purity may be overly pedantic for the main body of calculus. The exact definition of the limit is of great value but not necessarily for all of calculus. Indeed, 'most' derivatives can be found without the Weierstraß technical purity and 'many' courses already teach calculus without developing that purity. Thus there is ample cause to bring theory and practice more in line.

[NB. Added comment in February 2011: There is a paradox that I may refer to but have not developed further. In the Weierstraß definition of continuity around some x_0 it may be that there is some begging of the question, as the $\varepsilon > 0$ and $\delta > 0$ that are used may require their own infinitesimals.]

16. The news

16.1 Introduction

My hero, original physicist and later economist Jan Tinbergen, once noted that so much was published that every author had the duty to succinctly state what was new. These are points for *Conquest of the Plane,* including points refined here that were already stated earlier in *A Logic of Exceptions* and *Elegance with Substance*, assuming that you have not read all works by this author.

Tinbergen: http://nobelprize.org/nobel_prizes/economics/laureates/1969/

16.2 Major

16.2.1 Calculus

Definition of the dynamic quotient and redefinition of calculus using the dynamic quotient. Solution of the paradoxes of division by zero precisely at the point of interest (bishop Berkeley). Clarification of pitfalls when using the dynamic quotient (new paradoxes ?). Historical explanation.

Deduction with the dynamic quotient of the rules like chain rule and also the more complex derivatives of e^x and the trigonometric functions. Clarification of the distinction between the algebraic meaning of e and the manner how it is numerically approximated. Subsidiary: Didactic presentation at this level of education of e^x as a fixed point for derivatives in function space.

Demonstration that it is possible and better to start with surfaces instead of slopes. Via the change in surface there is a direct connection to the slope, which otherwise must be established separately again. Subsidiary: Consistent joint presentation of primitive and derivative.

Relatively minor but important for didactics: Surfaces under constant and linear functions can be introduced using only elementary tools, and thus allow the introduction of the notions of primitive and derivative without complexity. These are stepping stones towards complexity, and not examples (given after a complex introduction). Once these concepts are understood then the idea to generalize makes it acceptable to introduce some more complexity.

PM. David Tall now suggests for didactics to use a computer to zoom in on a function and show graphically that it can be stretched to become linear - whence it would be possible to create a tangent line. This is a serious mistake to be avoided. That mistake uses slope instead of surface, it uses numbers instead of algebra, it still neglects the very point where the derivative is taken, it neglects the clarity of the dynamic quotient, it does not educate students to the proper use of a computer language and decimal expansion. Strikingly, that proposal is offered by David Tall himself, who as one of the conceivers of the "procept" ought to be sensitive to the notion of the dynamic quotient.

David Tall, draft at http://www.warwick.ac.uk/staff/David.Tall/pdfs/chapter11_calculus.pdf

16.2.2 Trigonometry

Definition of the unit circumference circle or angular circle and functions x_{ur} and y_{ur} on it, that range on the unit radius circle.

Angle first measured as plane section. Unit 1 stands for the whole plane. Subsequently refined into Turn or Unit of Measurement Around (UMA), measured along the circle with unit circumference.

Consistent development of trigonometry from the Euclidean co-ordinates, with proper X and Y names, and demonstration that Cos and Sin are inverse functions.

Definition of $\Theta = 2\,\pi$.

Using radians to measure angles is economic in terms of concepts but appears a setback in terms of didactics as it appears useful to speak about angles and arcs as separate aspects. Angles are created by the pointy bits and turns, and arcs are those round curves. The problem of measuring the pointy bits is solved by using turns also expressed in an arc measure but this does not obliterate the terms and concepts involved. Hence, this book uses angles and arcs.

Clear discussion of the didactics of trigonometry with respect to these new findings.

Clarification for students that Cos and Sin are important only because of their derivatives. Presentation of an optimal compromise with respect to the use of angles and their unit measure.

16.2.3 Didactics in general

Explanation and clarification that western culture and language, and certainly the teaching in mathematics, still is infused with Euclids view on proportions, and that we have not yet adapted fully to the development of arithmetic and algebra.

Clarification of the distinction between no dimension and no specified dimension.

Named lines as a crucial step in didactics as a missing link between teaching *The Elements* and teaching analytic geometry. As explained in the didactics section this relatively innocuous idea formed the core around which the other ideas on calculus and trigonometry coalesced into the layout of this book, allowing also other suggestions from *Elegance with Substance* to find a place in the logical order.

Presentation of the key theorem of analytic geometry. The issue is already known as the addition rule for the cosine, and its key role between Euclidean co-ordinates and the polar plane and trigonometry is known too. Yet this role does not clearly transpire in textbooks. It is now put in place. PM. I constructed the proof in §6.2 myself, consider it very clear and am not aware of a similar format elsewhere.

16.3 Minor

Didactic emphasis on the verb and noun use in mathematics. (Parallel development with the Gray & Tall “procept” - a less accessible term.)

2 + ½ is the proper form (noun) instead of 2½.

Definition that 0.25 = 1/4, instead of that the decimal would be an approximation. Subsidiary: Explanation how this relates to how computer scientists program decimals on a computer.

The use of a tilde for rounding down to $0.\tilde{5}$ or up to $0.\underset{\sim}{5}$.

DoSqrt as a stepping stone between Sqrt (noun) and solving (verb) a quadratic equation.

Recovered exponent (rex) as a better name than logarithm.

Proportion space. Within a set of clear definitions for proportion, ratio, fraction, division, number.

Inclusion in a textbook of the place of the derivative of rex[x] between the polynomials (with thanks to Richard Fateman for making me aware of this).

16.4 Matter of taste

A fast track from geometry and arithmetic and algebra to analytic geometry and calculus. Discussion of geometry in succinct manner to allow spatial sense to awaken with some proofs but also relying on mere seeing and paper cutting, leaving the more serious treatment to analytic geometry.

The mix and steps in this book from co-ordinates to vectors to complex plane to linear algebra. While a traditional treatment develops each area separately and then proves the same things but in a different language so they only appear different, confusing students, the present approach emphasizes that these are mere different languages, so that there is little lost in moving quickly from one version to the other, but in fact gained in allowing each language to do the proof that it does best. This approach provides a better base for subsequent specialization.

Discussion of the role of the Pythagorean Theorem in all its forms met throughout this book. Explanation that the book is essentially a contemplation about the meaning of this theorem - that can also be a matter of definition.

Clarification to students how algebra is developed from arithmetic.

Using the general form of a line rather than only the functional form. Introduction of the notion of correspondence at this level of teaching.

Clarification also in an introductory textbook of the duality in solving two equations either for the points or the coefficients.

An elegant way to introduce linear and matrix algebra. First a demonstration what results can be obtained purely as a matter of logic, and only at a late stage look in detail how matrix multiplication would be defined.

Clearer focus on the role of the four quadrants for the trigonometric functions.

The derivatives of the trigonometric functions are found by analysis with only a very limited role for geometry - though the geometric interpretation is shown.

Presentation of linear regression and partial derivatives as key applications that belong in an introduction to analytic geometry and calculus. Link up with the determinant of a matrix as a measure of association. Link up with the geometry of correlation and the cosine as the correlation coefficient.

The role of the parabola as something of relatively small interest. The relevant concepts like intersection, vertex and slope can also be shown by more elementary forms like line and circle. When the complex plane is quickly introduced then the Quadratic Formula is less of an issue. The parabola reduces to an example application, at a late stage. It mainly supports vector analysis in the decomposition of vertical and horizontal movement when calculating the length of the arc.

Calculation of the arc of a circle using the general formula for calculating an arc.

Development of the textbook within the environment of *Mathematica,* and attractive application and own new additions.

Conclusion

One of the main points of *Elegance with Substance* is rather sober: "Didactics require a mindset sensitive to empirical observation which is not what mathematicians are trained for." EwS contains suggestions on how to re-engineer the industry. This will require involvement from various parts of society. That being said, it still are the teachers of mathematics who currently are given the responsibility to judge what is mathematically sound to teach the students. Being locked in tradition it is difficult to get out. While *Elegance with Substance* contains a shopping list for improvements and should be enough if a reader only lets the imagination roam freely, it is not a textbook, and a reader and especially a teacher may judge that the suggestions are dispersed and do not add up to a useful whole. *Conquest of the Plane* then provides this textbook format. My fellow teachers in mathematics now have an example of what it may become. It still is a primer only, and it does not cover all material required for a decent education, but it does fulfill the promise: it allows students to conquer the plane and it is a didactic existence proof for teachers.

For students: The aims of this book are modest. A student completing this book may still not know how to "construct an equilateral triangle on a line section using only a ruler and a compass", as simple as a geometry question can get. This book does not *train* in geometry. It does not train in any of the other subjects discussed either. The book aims at *understanding*. It aims at removing the traditional clutter in mathematical textbooks that block understanding. What students are capable of once completing this book - what is stated on the opening page - remains vague and would need to be established in practice. What is clear however is that students would be able to select their area of specialisation. Or know that different methods are valuable for their own strengths and that more training is required to grow competent. The aims set out at the beginning still are proper, and ought to have been achieved.

For teachers: Given the sorry state of mathematical education it is not likely that my fellow teachers of mathematics will pick up this book quickly. An addict to smoking is not cured easily. Teaching math does not only affect the teacher but has an impact on others. An addict with a social impact needs feedback and restraint from outside. It is not wise to let mathematicians be the only ones in charge to decide what proper math is and how it should be taught. They are not empirical scientists. If mathematicians are to live up to the ethic of mathematics and science in general to respect the evidence then they apparently must be reminded of it. Fortunately there are many professions where mathematics is used, in physics, engineering, biology, economics, evidence based medicine, psychological research, while also language teachers can have a say for example on verbs and nouns. How this could fit together is discussed in *Elegance with Substance.* The present book *Conquest of the Plane* can play a role in that process. It allows the other professions to compare traditional math textbooks with the present layout and hence better see the responsibility of the sciences to control the mathematical addiction to tradition for tradition only.

Literature

Colignatus is the name of Thomas Cool in science.

I thank Robert Palais, Peter Kop and Abraham Roth for comments on *Trig rerigged.* I thank Bob Parks for the EWP project at Washington University at St. Louis. The body of the book refers to various demonstrations at the WRI site and also other websites, all not mentioned here again, all thanked.

Alders, C.J. (1961), “Goniometrie”, Noordhoff

Alders, C.J. (1962), “Planimetrie”, Noordhoff

Aronson, E. (1972, 1992a), “The social animal”, Freeman, sixth edition

Aronson, E. (ed) (1973, 1992b), Readings about The Social Animal”, Freeman, sixth edition

Beaney, M. (2009), “Early Modern Conceptions of Analysis. Supplement to Analysis”, Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/analysis/s4.html

Cool (1999, 2001), “The Economics Pack, Applications for *Mathematica*”, Scheveningen, ISBN 90-804774-1-9, JEL-99-0820

Colignatus (2005a), “Definition and Reality in the General Theory of Political Economy”, 2nd edition, Dutch University Press

Colignatus (2005b), “The Economist on 100 years of Einstein”, http://ideas.repec.org/p/wpa/wuwpgt/0501003.html

Colignatus (2007a), “A Logic of Exceptions”, Thomas Cool Consultancy & Econometrics, ISBN 978-90-804774-4-5, http://thomascool.eu/Papers/ALOE/Index.html

Colignatus (2007b), “Voting Theory for Democracy”, 2nd edition, Thomas Cool Consultancy & Econometrics, ISBN 978-90-804774-5-2

Colignatus (2007j), “Refutation of “the” proof for Cantor’s Theorem and restoration of the “set of all sets””, http://thomascool.eu/Papers/ALOE/2007-07-29-OnCantorsTheorem.pdf

Colignatus (2007k), “Improving the logical base of calculus on the issue of “division by zero””, http://thomascool.eu/Papers/ALOE/2007-07-31-ImprovingTheLogicOfCalculus.pdf

Colignatus (2007m), “A difficulty in proof theory”, http://thomascool.eu/Papers/ALOE/2007-08-02-ADifficultyInProofTheory.pdf

Colignatus (2009), “Elegance with Substance. Mathematics and its education designed for Ladies and Gentlemen”, Dutch University Press, http://thomascool.eu/Papers/Math/Index.html (PDF via MPRA)

Damasio, A. (2003), “Looking for Spinoza”, Harcourt

Davidse, A. (2009), “Stevin, Noten”, http://www.xs4all.nl/~adcs/stevin/noten.html

DeLong, H. (1971), “A profile of mathematical logic”, Addison Wesley

Descartes, R. (1637), “Discours de la Méthode”, reprinted in (1958) “Descartes, Philosophical writings, Selected and translated by Norman Kemp Smith”, A modern library book, 43, Random House

Dormolen, J. van (1976), “Didactiek van de wiskunde”, tweede herziene druk, Bohn, Scheltema & Holkema

Drenth, K. & Th. Colignatus (2000), “Transport science for operations management. Understanding elementary physics and mechanical engineering”, Thomas Cool Consultancy & Econometrics, http://thomascool.eu/Papers/TSOM/Index.html

Dijksterhuis, E.J. (1990), “Clio’s stiefkind. Een keuze uit zijn werk door K. van Berkel”, Bert Bakker

Elgersma, S. and W. Verdenius (1974), “Voortgezette analyse”, syllabus Mathematical Institute, Rijksuniversiteit Groningen

Eijk, E. van der, c.s. (2005), “Moderne wiskunde. VWO 3a&b”, editie 8, Wolters-Noordhoff

KNAW (2006), “Teaching and Learning Trajectories”, http://www.onderzoekinformatie.nl/nl/oi/nod/onderzoek/OND1315405/

Freudenthal Instituut (2009), “TAL, tussendoelen en leerlijnen”, http://www.fi.uu.nl/talbovenbouw/

Goodman, P. (1962, 1973) “Compulsory miseducation”, Penguin Education Specials

Gulik-Gulikers, I. van (2005), “Geschiedenis van de niet-Euclidische meetkunde”, Epsilon uitgaven 21

Hiele, P.M. van (1973), “Begrip en inzicht”, Muusses

Hughes-Hallett, D. and A. Gleason et al. (2000), “Calculus. Alternate Version”, 2nd edition, John Wiley & Sons

Nahin, P.J. (1998, 2007), “An imaginary tale: The story of $\sqrt{-1}$ ”, Princeton Science Library

Palais, R. (2001a), “π Is wrong!”, The mathematical intelligencer, Vol 23, no 3 p7-8

Palais, R. (2001b), “The Natural Cosine and Sine Curves”, JOMA, http://mathdl.maa.org/mathDL/4/?pa=content&sa=viewDocument&nodeId=483

Pot, H. (2009a), “Aan 27 onderwijs-mensen, e-mail plus twee bijlagen, 29 juni 2009”, email June 29. Pot, H. (2009b), “Verhoudingen - in soorten. Soms is het woord verhouding synoniem voor positief getal, soms ook niet. Hoe zit dat?”, paper July 6. Pot, H. (2009c), “Literatuur-citaten (soms met commentaar), als achtergrond bij de tekst: “Verhoudingen - in soorten””, paper July 6.

Pot, H. (2009d), “Is een verhouding wel of niet hetzelfde als een breuk. Oproep”, Euclides 84, 8 p308-309

Reichard, L. c.s. (2008), “Getal & Ruimte. Wi VWO B deel 1, 2, 3”, EPN

Scheelbeek, P.A.J. and W. Verdenius (1973), “Analyse”, syllabus Mathematical Institute, Rijksuniversiteit Groningen

Spiegel, M.R. (1963), “Advanced Calculus”, SI (metric edition), Schaum’s Outline Series

Struik, D.J. (1977), “Geschiedenis van de wiskunde” SUA (“A concise history of mathematics”, Dover 1948)

David Tall (about 2010), “Chapter 11. Blending Ideas in the Calculus”, in “How Humans Learn to Think Mathematically”, draft at http://www.warwick.ac.uk/staff/David.Tall/pdfs/chapter11_calculus.pdf

Sullivan, M. (1999), “ College algebra”, Prentice Hall

Wallace, D.F. (2003), “Everything and more. A compact history of ∞”, Norton

Wolfram, S. (1992), “Mathematica”, Addison-Wesley

Wolfram, S. (1996), “Mathematica 3.0”, Cambridge