# Mathematics in Context

Michael Dreher
Department of Mathematics and Statistics
Heriot–Watt University Edinburgh

Fall 2016

# Preface

## Welcome to the Mathematics Programme

You are now students of Mathematics ! The word *mathematics* comes [4] from the Greek μαθεματικά which is related to the verb μάθεσις which is being translated as *the act of learning, getting to knowledge*, but also as *desire of learning.* So in a sense, if you study mathematics, you also *learn how to learn.*

In this course, you will learn how mathematics is useful in the professional life. One of the first to observe this was Marcus Vitruvius who wrote about $30-15$BC in Book VI of his *Ten Books on Architecture* [6]:

> "It is related of the Socratic philosopher Aristippus that, being shipwrecked and cast ashore on the coast of the Rhodians, he observed geometrical figures drawn thereon, and cried out to his companions: 'Let us be of good cheer, for I see the traces of man.' With that he made for the city of Rhodes, and went straight to the gymnasium. There he fell to discussing philosophical subjects, and presents were bestowed upon him, so that he could not only fit himself out, but could also provide those who accompanied him with clothing and all other necessaries of life. When his companions wished to return to their country, and asked him what message he wished them to carry home, he bade them say this: that children ought to be provided with property and resources of a kind that could swim with them even out of a shipwreck.
>
> These are indeed the true supports of life, and neither Fortune's adverse gale, nor political revolution, nor ravages of war can do them any harm. Developing the same idea, Theophrastus, urging men to acquire learning rather than to put their trust in money, states the case thus: 'The man of learning is the only person in the world who is neither a stranger when in a foreign land, nor friendless when he has lost his intimates and relatives; on the contrary, he is a citizen of every country, and can fearlessly look down upon the troublesome accidents of fortune. But he who thinks himself entrenched in defences not of learning but of luck, moves in slippery paths, struggling through life unsteadily and insecurely.' "

In addition to being the highest form of learning, and to being the origin of culture and of architecture, mathematics appears in most modern professions and scientific disciplines. The Course *Mathematics in Context* will shed some light on these connections. Since everybody already knows that mathematics is relevant to banking, finance, insurance, and taxes, we will not talk about these topics at all; instead we will present applications that are hopefully new to you.

The course *Mathematics in Context* approaches mathematics differently than you have seen it in school, and this is a good thing. At first glance, the *Context* course may look different than the other courses on *Calculus*, *Introduction to University Mathematics*, *Introduction to Statistical Science*, but a second glance (perhaps several weeks later) will convince you that there are many connections to those other courses. The *Context* course deals often with the same topics as the other courses, but looks at them sometimes from a different angle, which helps us to obtain a deeper understanding.

Transitioning into the university mindset takes some time, and therefore it is better to start right now.

# Contents

**1 What is This Course About ? 7**

**2 Foundations 9**

2.1 The Purpose of this Chapter . . . 9

2.2 Common Knowledge . . . 9

2.2.1 Elementary Stuff . . . 9

2.2.2 Functions etc. . . . 12

2.2.3 Exponential Functions and Logarithms . . . 15

2.2.4 Trigonometry . . . 16

2.2.5 Geometry . . . 19

2.3 Some Key Ideas of Calculus and Geometry . . . 24

2.3.1 Taking Derivatives . . . 24

2.3.2 Working with Derivatives . . . 30

2.3.3 Integrals . . . 36

2.3.4 More Geometry . . . 45

2.4 Some Easy Aspects of Physics . . . 47

2.4.1 Motions in $\mathbb{R}^1$ . . . 47

2.4.2 Motions in $\mathbb{R}^3$ . . . 49

2.4.3 Forces . . . 51

2.5 How to do Mathematics . . . 53

2.5.1 The Purpose of Homework at University . . . 53

2.5.2 Survival Tips . . . 54

2.5.3 How to Write Mathematics . . . 54

2.5.4 Recommendations for the Exercise Session . . . 57

2.6 The Greek Alphabet . . . 57

**3 Mathematics in Nature — Rainbows 59**

3.1 General Remarks . . . 59

3.2 Physical Remarks . . . 60

3.3 First Considerations . . . 61

3.4 Where do the Colours Come From ? . . . 69

3.5 Now it is Your Turn . . . 69

**4 Mathematics in Sports 71**

4.1 How to Win a Shot Put . . . 71

4.2 How to Add up Points in Decathlon and Heptathlon . . . . . 75
4.3 How do World Records Evolve Over Time . . . . . 80

**5 Mathematics in Production Planning 83**

5.1 The Objective . . . . . 83
5.2 Matrices . . . . . 83
5.2.1 Cakes and Their Ingredients . . . . . 83
5.2.2 Prices of Ingredients and of Cakes . . . . . 86
5.2.3 What is the Meaning of all this ? . . . . . 87
5.3 Production Planning . . . . . 87

**6 Mathematics in Your SatNav 95**

6.1 Physical Background and Description of the Problem . . . . . 95
6.2 Elementary Ideas of Probability Theory . . . . . 96
6.3 Random Binary Sequences . . . . . 97
6.4 Pseudo–Noise Sequences . . . . . 99
6.5 Advanced Algebraic Methods . . . . . 100
6.6 Navigation in 1D . . . . . 102
6.7 Navigation in 2D . . . . . 108

**7 Mathematics in Your Mobile Phone 113**

7.1 Historical Perspective . . . . . 113
7.2 Mathematical Challenges . . . . . 113
7.3 How to Estimate the Channel . . . . . 114
7.4 How to Reconstruct the Transmitted Signal . . . . . 120

**A Some Program Codes 121**

A.1 The Python Code For Section 3.4 . . . . . 121

# Chapter 1

# What is This Course About ?

This course is *on the one hand* about *mathematical modelling*, which means

> to construct artificial worlds (models) that follow mathematical rules, with the purpose of describing pieces of the real world adequately. The models are as easy as possible and as complex as necessary.

The complexity of the model depends on the desired accuracy. In an example of explaining a rainbow, the rain drops (which produce the rainbow somehow) could be described

- as drops in 3D (with a thin end at the top and a wide end at the bottom),
- or as (perfectly round) balls in 3D,
- or as circles in 2D.

Another example: a mechanical body (such as the earth travelling around the sun) could be modelled as

- an elastic body (which means that it can be deformed)
- or as a rigid body (which can not be deformed)
- or as a point (which has no form).

For an appropriate understanding of the model, a precise formulation of the rules of the model is necessary. Precise means mathematical. We will have to work much more rigorously than at school.

For applying maths in a real world context, you need solid and reliable knowledge in other sciences and in various areas of mathematics — *at the same time !*

In the SatNav example, this means knowledge in

- physics including special relativity and general relativity,
- calculus (taking derivatives of multi-variable functions),
- algebra (polynomials, prime numbers,. . . )
- numerical mathematics (how to solve a collection of equations simultaneously when you have no solution formula),
- probability and statistics and information sciences.

Mathematical modelling is done in a *cycle*:

**Step 1:** start from the real world situation, describe the setting using mathematical concepts, give names to all appearing entities. *Get your thinking straight.*

**Step 2:** translate the question that you wish to be answered into mathematical language, building upon the names and definitions specified in the previous step.

**Step 3:** solve the mathematical problem, obtain a mathematical solution.

**Step 4:** translate back into the real world setting. Check if your answer makes sense. If the answer is not precise enough, go back to Step 1 and refine the model.

*However*, there is more to modelling than just doing some calculations. Let us quote Marcus Felson[1] who worked as professor at the *School of Criminal Justice* at Rutgers University:

> "Still, the mathematical modeller who works closely with the tangible features of crime will do better than the one who looks for deep meaning in crime data. Mathematicians should also be aware that their most advanced models are unlikely to solve basic issues about crime. I submit that the main contribution of mathematics is clear definition and clear thinking. Indeed, mathematical *discipline* and *systematic thinking* may prove much more important than mathematical fanciness in helping us learn more about crime".

This brings us to the second objective of this course — *mathematics as a science*. Because if you wish to get a degree as a Bachelor of *Science*, you have to do *Science*, obviously.

The purpose of science is to expand the knowledge of mankind, which means two things:

- to pass on the scientific knowledge and mindset to the next generation (called teaching and learning),
- to shift the border between the known and the unknown (called research).

At University, you will get into contact with both of them, so let us think about how to do research because the teaching will follow from that naturally. Recall that according to the Learning and Teaching Strategy of this University, teaching should be *research-informed*, so what does a researcher do all day long ? Struggle with the unknown.

There are various sciences, and each one has its own methods for its research:

- in philosophy and literature, scholars *discuss* a lot. But this is not discussion in the sense of a random chit-chat, instead an academic discussion that follows rigorous academic standards.
- in biology, scientists make *observations* which are then being *systematized* and *classified*.
- in chemistry and physics, scientists make *experiments*, from which they then draw conclusions. These conclusions then lead to theories. And in order to check whether we can trust these theories, more experiments are being conducted. This cycle between devising a theory and checking it experimentally then repeats indefinitely.
- in mathematics, scientists do *PROOFS*, which are lines of thought that must obey specific rules of logics. The logical correctness is crucial here, for otherwise the proof simply does not exist.

Proofs are actually something really useful, because they give us confidence that our results are correct. The derivative of the sine function is the cosine function, and this is true because we can prove it. And once a mathematical statement is proved, we can rely on it for eternity. Compare us to the nutritionists whose recommendations about what you should and should not eat change every few years.

Let us consider the sine function and its derivative a bit longer. So we write down a theorem with the wording "The derivative of the sine function is the cosine function". The proof must be logically correct, so we must get our thinking straight. And you cannot get your thinking straight if you don't know what the words you are using actually mean. Therefore, we will have to write three definitions somewhere in front of our theorem: one definition that specifies what the sine function is, one definition for the cosine, and one definition for the word "derivative".

Doing proofs is a fundamental technique for the creation of reliable knowledge in mathematics.

[1] What every mathematician should know about modelling crime. *European Journal of Applied Mathematics*, volume 21, pages 275–281, 2010

# Chapter 2

# Foundations

## 2.1 The Purpose of this Chapter

- We list various mathematical facts you are expected to know.
- We explain the meaning of derivatives, integrals, vectors. The goal here is to obtain a deeper understanding, because otherwise you will not be able to answer real-world questions.
- We give some gentle introduction into easy aspects of physics.
- We give some hints on how to do mathematics successfully, and we clarify what we expect from you in this course and in the exercise classes.

## 2.2 Common Knowledge

Please read this section yourself because it will not be presented in class. You are expected to be familiar with what follows, in the sense of "somebody wakes you up at 3am and you must be able to explain it".

### 2.2.1 Elementary Stuff

**Numbers**

The number zero is neither positive nor negative. The set of natural numbers is $\mathbb{N} = \{0, 1, 2, \dots\}$. Sometimes we need to prohibit the zero, and for this purpose we have $\mathbb{N}_+ = \{1, 2, 3, \dots\}$. The set of integers is $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$. The set $\mathbb{Q}$ of all rational numbers contains all the fractions $\frac{m}{n}$ where $m$ and $n$ are from $\mathbb{Z}$, but $n$ is not zero. Later this will be abbreviated as follows:

$$\mathbb{Q} := \left\{ \frac{m}{n} : m \in \mathbb{Z}, \quad n \in \mathbb{Z}, \quad n \neq 0 \right\}$$

and this is being read as "$\mathbb{Q}$ is defined as the set of expressions $\frac{m}{n}$ with the property that $m$ is an element of $\mathbb{Z}$, $n$ is an element of $\mathbb{Z}$, and $n$ does not vanish".

Each rational[1] number has a decimal representation. Examples are

$$\frac{2}{8} = 0.25, \quad \frac{1}{6} = 0.1666666\dots, \quad \frac{7}{22} = 0.31818181818181818\dots \quad \frac{2}{7} = 0.285714\overline{285714}\dots.$$

We see that this decimal representation sometimes stops (as in the case of $\frac{2}{8}$), or it never stops *but eventually becomes periodic*. In some exceptional cases, this decimal representation of a rational number is not unique, which means that one rational number possesses two representations, such as

$$\frac{1}{2} = 0.5 \quad \text{but also} \quad \frac{1}{2} = 0.49999999\dots.$$

[1] The word *rational* comes from the Latin noun *ratio* which means *reason* or *proportion*. This means that a rational number is the proportional of two numbers.

The reason is that 0.5 and 0.49999… are actually the same number. Because if they were different, then there would be a third number between them (and this third number would be different from the two), but which decimal representation would then the third number have ?

And then there are other numbers whose decimal representation never stops and never becomes periodic. Examples are

$$\begin{aligned} \sqrt{2} &= 1.4142135623730950488016887242096980785696718753769480731766797379907324784621\ldots \\ \pi &= 3.14159265358979323846264338327950288419716939937510\ldots. \end{aligned}$$

Such numbers are called *irrational*[2]. All rational numbers and irrational numbers together form the set $\mathbb{R}$ of real numbers. Every irrational number can be approximated by some rational number in the following sense: if a real number $x$ is given to you and an error bound $\varepsilon$ is given as well, then you can find a rational number $\frac{m}{n}$ such that the distance $|x - \frac{m}{n}|$ is smaller than $\varepsilon$. In fact, you can find many such fractions $\frac{m}{n}$.

**DIY:** *Find how to do this with $x = \pi$ from above and $\varepsilon = 10^{-8}$.*

### Powers

If $x$ is a real number and $n \in \mathbb{N}_+$, then $x^n = x \cdot x \cdot \ldots \cdot x$ with $n$ factors on the right-hand side. In case of $n = 0$, we make the agreement $x^0 := 1$, for every $x \in \mathbb{R}$. And if $n \in \mathbb{Z}$ is negative, then we define

$$x^n := \frac{1}{x^{-n}} \quad \text{provided that} \quad x \neq 0.$$

As an example with $n = -3$, this means $x^{-3} := \frac{1}{x^3}$ provided $x \neq 0$.

The above agreements then allow us to conclude that

$$x^{m+n} = x^m \cdot x^n, \quad \text{for all} \quad x \in \mathbb{R} \setminus \{0\} \quad \text{and all} \quad m, n \in \mathbb{Z}. \tag{2.1}$$

We also have the rule

$$(x \cdot y)^m = x^m \cdot y^m, \tag{2.2}$$

valid for all $x$, $y \in \mathbb{R}$ and all $m \in \mathbb{Z}$, subject to the restriction that if $m$ is negative, then $x = 0$ is forbidden and $y = 0$ is also forbidden because you would divide by zero.

And furthermore, we have the rule

$$\Big(x^m\Big)^n = x^{m \cdot n}, \qquad \forall x \in \mathbb{R} \setminus \{0\}, \quad \forall m, n \in \mathbb{Z}. \tag{2.3}$$

The upside-down A means *for all*. This is just an abbreviation.

The three formulas (2.1)–(2.3) are very powerful, and in what follows we wish to generalise them as much as we can. In particular, we wish to allow for more general $m$ and $n$.

### Roots. And Powers with Arbitrary Exponents

We never take square roots of negative numbers. This means that $\sqrt{-9}$ does not exist. We also do not take cube roots of negative numbers, hence also $\sqrt[3]{-8}$ is undefined. Although we could set $\sqrt[3]{-8} = -2$, below we will learn that this is a bad idea.

The general definition is: if $x \in \mathbb{R}$ is greater than or equal to zero, and if $p \in \mathbb{N}$, then $\sqrt[p]{x}$ is defined as that non-negative real number $y$ which has the property $y^p = x$. Instead of writing $y = \sqrt[p]{x}$, we may also write $y = x^{1/p}$.

Observe that $y$ is required to be *non-negative*. This means that $\sqrt{25} = +5$, and we never have $\sqrt{25} = -5$. Any statement of the form $\sqrt{25} = \pm 5$ is *wrong*.

$$\boxed{\text{Any root always has only one single value, and this value is always } \geq 0.}$$

[2] In earlier millennia, the word was *inrational*, which means *not rational*. For reasons of pronunciation, the $n$ became an $r$ afterwards.

Now comes something nice: by the very definition of the $p$th root, we have

$$\left(x^{1/p}\right)^p = y^p = x = x^1 = x^{\frac{1}{p}\cdot p},$$

which very much resembles (2.3) if you choose $m = 1/p$ and $n = p$. This is inspiring.

We come to the definition of $x^r$ if $x > 0$ is real and $r$ is rational and *positive*:

$$\text{If } x > 0 \quad \text{and} \quad r = \frac{p}{q} \quad \text{with} \quad p, q \in \mathbb{N}_+, \quad \text{then we define } x^r := \left(x^p\right)^{\frac{1}{q}} = \sqrt[q]{\left(x^p\right)}.$$

Here we need to be careful because each positive rational number $r$ has many representations as fractions, and an example is $\frac{2}{3} = \frac{4}{6} = \frac{6}{9} = \frac{8}{12} = \dots$, and therefore it becomes necessary to prove that all the four numbers

$$\left(x^2\right)^{\frac{1}{3}} = \sqrt[3]{x^2}, \qquad \left(x^4\right)^{\frac{1}{6}} = \sqrt[6]{x^4}, \qquad \left(x^6\right)^{\frac{1}{9}} = \sqrt[9]{x^6}, \qquad \left(x^8\right)^{\frac{1}{12}} = \sqrt[12]{x^8}$$

are actually equal. But they are. We skip the proof of this fact.

We come to the definition of $x^r$ if $x > 0$ is real and $r$ is rational and *negative*:

$$\text{If } x > 0 \quad \text{and} \quad r = \frac{p}{q} < 0 \quad \text{with} \quad p, q \in \mathbb{Z}, \quad \text{then we define } x^r := \frac{1}{x^{-r}}.$$

Having defined $x^r$ for all positive real $x$ and all rational $r$, we now celebrate the following formulas which generalise (2.1)–(2.3):

$$x^{r+s} = x^r \cdot x^s, \qquad \forall x > 0, \quad \forall r, s \in \mathbb{Q}, \tag{2.4}$$

$$\left(x \cdot y\right)^r = x^r \cdot y^r, \qquad \forall x > 0, \quad \forall y > 0, \quad \forall r \in \mathbb{Q}, \tag{2.5}$$

$$\left(x^r\right)^s = x^{r \cdot s}, \qquad \forall x > 0, \quad \forall r, s \in \mathbb{Q}. \tag{2.6}$$

We skip their proofs as well.

And now we can understand why defining $\sqrt[3]{-8} = -2$ would be a bad idea:

$$+2 = \sqrt[6]{64} = \left(64\right)^{\frac{1}{6}} = \left((-8)^2\right)^{\frac{1}{6}} \overset{(2.6)}{=} \left(-8\right)^{2\cdot\frac{1}{6}} = \left(-8\right)^{\frac{2}{6}} = \left(-8\right)^{\frac{1}{3}} = \sqrt[3]{-8} = -2. \qquad ☹$$

If we were to define $\sqrt[3]{-8} := -2$, then we have a price to pay, which is that the equation in (2.6) becomes wrong for $x < 0$. We don't want to pay this price.

And finally we define $x^r$ for $x > 0$ and $r$ *irrational*, such as $2.35^{\sqrt{2}}$ or $4^\pi$. The idea is to approximate the irrational exponent $r$ by a sequence of rational numbers. For instance, the number $\sqrt{2} = 1.414213562\dots$ can be approximated by the sequence

$$\frac{14}{10}, \quad \frac{141}{100}, \quad \frac{1414}{1000}, \quad \frac{14142}{10000}, \quad \frac{141421}{100000}, \dots,$$

and this sequence has limit equal to $\sqrt{2}$. You will learn in the courses *Calculus* and *Multivariable Analysis* what the precise definition of a limit is. Now choose some positive real number $x$ and keep it fixed. In order to then define $x^{\sqrt{2}}$, we consider the sequence

$$x^{\frac{14}{10}}, \quad x^{\frac{141}{100}}, \quad x^{\frac{1414}{1000}}, \quad x^{\frac{14142}{10000}}, \quad x^{\frac{141421}{100000}}, \dots,$$

and we need to prove that this sequence actually has a limit. We must omit this proof (because we have not defined anywhere what a limit is).

The positive result then is that the equations in (2.4)–(2.6) stay valid for all $r$ and $s$ from $\mathbb{R}$ (not just $\mathbb{Q}$).

### Absolute Values

You obtain the absolute[3] value $|x|$ of a real number $x$ if you throw away its sign.

Hence we have $|3.2| = 3.2$ because 3.2 can be equivalently written as $+3.2$ and you throw away the $+$. And we have $|-3.7| = 3.7$. The geometrical meaning of $|x|$ is the distance of the point $x$ from the point 0 on the number line. The geometrical meaning of $|a-b|$ is the distance of the two points $a$ and $b$ from each other on the number line (it does not matter at all whether $a$ or $b$ are positive or negative).

The mathematical definition of $|x|$ is

$$|x| := \begin{cases} x & : \text{ if } x \geq 0, \\ -x & : \text{ if } x < 0. \end{cases}$$

The absolute value function has the properties

$$\begin{aligned} |x \cdot y| &= |x| \cdot |y|, \qquad \forall x, y \in \mathbb{R}, \\ |x + y| &\leq |x| + |y|, \qquad \forall x, y \in \mathbb{R}. \end{aligned}$$

**DIY:** *Deduce from these two properties the formulas*

$$\begin{aligned} \left|\frac{x}{y}\right| &= \frac{|x|}{|y|}, \qquad \forall x \in \mathbb{R}, \quad \forall y \in \mathbb{R} \setminus \{0\}, \\ |x - y| &\leq |x| + |y|, \qquad \forall x, y \in \mathbb{R}. \end{aligned}$$

**DIY:** *Explain why the formula $\sqrt{x^2} = |x|$ is correct for all $x \in \mathbb{R}$, but the formula $\sqrt{x^2} = x$ is* ***not*** *correct for all $x \in \mathbb{R}$.*

## 2.2.2 Functions etc.

### What is a Function ?

A function $f$ connects two variables. Typically we write

$$y = f(x),$$

and the variable $x$ is called *independent* variable, because it is allowed to roam around freely. In contrast to that, the variable $y$ is called *dependent* variable, because it depends on $x$. If $x$ has been chosen somehow, then $y$ can only have one value.

A typical example is

$$y = \sin(x), \quad \text{where} \quad x \in \mathbb{R}.$$

If you choose here $x = \frac{\pi}{4} \approx 0.7853981635$, then you get $y = 0.7071067810\ldots$.

You can imagine a function as a machine where you put something in (in this case the number $0.7853981635\ldots$) and then you get something out (here it is the number $0.7071067810\ldots$).

There are various ways how to represent a function:

- as a list of values
- graphically in an $x - y$ diagram
- by means of a formula.

You need to be familiar with all three of them. One thing should be kept in mind: whenever we write "$y = f(x)$", this does not necessarily mean that we are in possession of a specific formula that connects $y$ to $x$. In mathematics, you sometimes will have to work with objects (in this case the object "$f$") which you do not know or even never will know.

[3] *absolute* comes from the Latin verb *absolvo*, *absolvere*, *absolvi*, *absolutus* which means *to release, to free.* The meaning is something like *to free the number from the tyranny of its sign.*

### Representing a Function as a List of Values

Suppose a person is trying to lose some weight. At each point of time, this person has a certain weight, which is a certain real number to which we attach the unit “kilogram”. Each morning 8am, that person puts themselves on the scales, reads off a number, and writes this number down on a sheet of paper, in a table with two columns. The first column contains the day, the second the weight.

This is an example of a representation of a function using a list of values. The independent variable $x$ is the time, and this $x$ runs smoothly through some interval, without gaps. The function “weight depending on time” has many more values which we do not know, because we only have determined the values at 8am each morning. No sane person would keep standing on the scales for 24 hours a day without interruption.

Other examples of functions represented as value lists are provided by smartphones. The smartphone which I have right now has the following sensors: 3 accelerometers (which measure how fast the velocity of the phone is changing, and you need three of them because the acceleration is a vector in $\mathbb{R}^3$ of which each coordinate has to be measured separately); 3 magnetic field sensors; a battery status sensor; a phone network signal strength sensor; another one for the WiFi signal strength; and finally a microphone signal level sensor. It seems that my phone should have more sensors, but I have no app for them. Reading a value from a sensor takes some time (and battery), and therefore each of these sensors generates one number at each time step, and each such time step has a duration of a few milliseconds (I guess).

### Representing a Function Graphically

Suppose you have a set of shelves where you store all your books. Since you own many maths books, with so many pages containing formulas created for your entertainment, there is now a considerable weight on each shelf, and consequently that shelf is now sagging. Out of curiosity, you wish to determine the *bending line* of that shelf.

Suppose the shelf length is 1 metre. So you choose as unit centimetres, let the independent $x$ variable run from 0 till 100 (because 1 metre equals 100 centimetres), and the dependent $y$ variable means how much the shelf is sagging at position $x$. Clearly $y$ depends on $x$. If you wish to determine $y$ for $x = 35$, hence evaluate $y(35)$, your only method available is to make use of a folding yardstick (or a similar instrument).

Other examples of functions represented as graphs are seismograms which are being generated from seismometres: these are mechanical devices for measuring the vibrations or other motions of the earth, using some kind of pen that draws lines on a very long moving sheet of paper.

### Representing a Function by Means of a Formula

That is what you have done in school: you had to deal with functions such as $y = \sin(x^2) + x^3$, where the variable $y$ depends in a formulaic way on the variable $x$.

### How to Convert the Three Representations Into Each Other

**Given a list of values:** take the dietary example from above. If we know the Monday morning and Tuesday morning weights of 85.0kg and 85.5kg, what can be said about the weight at Monday evening 9pm ? Well, nothing. This means that we cannot turn a value list into a graph. Of course we could visualise the list of values as dots in a diagram, and then we could connect neighbouring dots by straight lines. But then we would pretend to possess information which we in fact do not have, which would be in most situations unethical. And with a similar reason, we cannot produce a formula from a list of values. The only thing that can be done is some approximation which may or may not convince a reader.

**Given a graph:** you can easily obtain a list of values from a graph: just pick some points on the graph and measure their positions using a yardstick or a ruler.

**Given a formula:** assuming the realistic situation that the formula expresses something computable, you easily can produce a list of values. And if you have sufficiently many values (in relation to your printer resolution), then you can also generate a graph.

### Inverse Functions

Suppose two variables $x$ and $y$ are related as $y = f(x)$ (which, as explained already above, does not mean that we are in a possession of a formulaic expression of $f$ — this $f$ may very well be unknown to us). The meaning is that we put $x$ into $f$, and then this "machine" $f$ produces some $y$ in some obscure black-box manner.

The question is now: can we do this "backwards" ? In other words: suppose someone has given us a specific $y = 0.2$. Are we then able to find that $x$ which (when being fed into the function $f$) results in $y = 0.2$ ?

An example: let $f = \sin$ and $y = 0.2$. We wish to know who are those $x$ for which $\sin(x) = 0.2$ becomes true ? You perhaps know already from school that there are many such $x$. If we are allowed to pick only one of these $x$ (which is a very common requirement in mathematics), which one shall it be ?

A function which performs this "backwards calculation" is called *inverse function*. An example is: if $f(x) = x^2$ is the square function, defined for $x \geq 0$, then the inverse function of $f$ is the square root function.

**DIY:** *Why did I emphasise "defined for $x \geq 0$" in this example ?*

Now we explore how to determine an inverse function:

**If $f$ is given as a list of values:** you swap both columns and then sort the table according to ascending order of the new left column. Since this table is all information you have, there is nothing more which you could do.

**If $f$ is given as a graph:** you have that graph, drawn with a thick-nibbed pen on thin paper. The two axes have one arrow each. Now you turn over the paper, and then you rotate it in such a way that you see faintly one arrow pointing to the right, and the other arrow pointing upwards. What you can now see from the backside of the paper is the graph of the inverse function.

**If $f$ is given as a formula:** you have the equation, such as $y = f(x) = \sin(x^2) + x^3$. Now you solve it for $x$. If you are successful for that $f$, tell me how you did it.

### Linear and Affine Linear Functions. Their Gradient aka Slope

The functions $y = 2x$ and $y = -3x$ are called *linear*[4] functions. The general form is $y = a \cdot x$ where $a$ is a chosen constant.

If you shift the graph of a function $y = a \cdot x$ by a constant $b$ (upwards or downwards depending on the sign of $b$), then you obtain the graph of an *affine*[5] *linear* function $y = ax + b$.

The *slope* (also called *gradient*) of a straight line in a coordinate system is a real number that tells us by how much the dependent variable changes when the independent variable grows by one. You can figure out the slope of a straight line in a coordinate system also according to the mnemonic formula

$$\text{slope} = \frac{\text{vertical change}}{\text{horizontal change}} = \frac{\text{rise}}{\text{run}}.$$

The concept of *slope* is absolutely crucial for understanding calculus and analysis, you have to be highly familiar with it. The wikipedia entry "slope" is quite informative.

[4] from the Latin adjective *linearis* which means *belonging to lines*. That name is quite natural because the graph of a function $y = a \cdot x$ obviously **is** a straight line.

[5] which comes from the Latin *adfinis* meaning *neighbouring*, and *adfinis* itself contains the intensifying *ad-* and the word *finis*, meaning *boundary*. The graph of the function $y = ax + b$ is a neighbour of the graph of the function $y = ax$.

### 2.2.3 Exponential Functions and Logarithms

EULER's[6] number $e$ is $e = 2.718281828459\ldots$, it is irrational, and the exponential function is

$$\exp(x) := e^x, \qquad \text{where} \quad x \in \mathbb{R}.$$

A first obvious question which you might have found yourself already (every teacher appreciates curious students who raise questions) is: where does this funny number $2.718281828459\ldots$ come from ? Why didn't we select a simpler number like 2 or 10 ?

The historical reason is this one:

Suppose you have 100£ in a bank account, and the bank pays 1% interest. If you keep the money there from 1 January till 31 December untouched, you will obtain at the end of the year $100 \cdot (1 + 0.01) = 101$£ .

If you withdraw the money on 30 June, the bank will pay you half the interest, so you will get $100 \cdot (1 + \frac{0.01}{2}) = 100.50$£ , and then you directly pay back this amount into your account again. Then this amount of 100.50£ will earn an interest of 0.5% for the remaining half of the year, giving you a total amount of $100 \cdot (1 + \frac{0.01}{2})^2 = 101.0025$£ on 31 December.

If you withdraw the money of 100£ (plus the earned interests) after a third of a year, you will receive $100 \cdot (1 + \frac{0.01}{3})$£ , and then you pay back this amount into your account again, and then you withdraw everything after two thirds of the year (with the interests), and then you pay what you received back into your account. On 31 December, you will have an amount of $100 \cdot (1 + \frac{0.01}{3})^3 = 101.003337037$£ . This is making us even richer (by 0.08337037p) than the midyear split approach.

The question is: how rich can we become following this scheme if we split the year into ever smaller pieces ? In mathematical terms: what is the value of

$$100 \cdot \lim_{n \to \infty} \left(1 + \frac{0.01}{n}\right)^n \quad ?$$

You can guess an answer if you take your calculator and try $n = 100$, $n = 200$, $n = 1000$ and perhaps some more $n$. The value which we try to find amounts to $100 \cdot e^{0.01} = 101.0050167\ldots$, with $e$ being this mysterious number $e = 2.718281828459\ldots$ which may be called *Banker's Constant.* Because it actually **is** a constant: it does not depend on the currency which a country has chosen or will choose in the future, it does not depend on the interest rate, it does not depend on the length of the year. Even a bank that is registered somewhere in the Andromeda Galaxy will operate with the same constant $e = 2.718281828459\ldots$ .

It does not matter mathematically whether we write $e^x$ or $\exp(x)$. The crucial properties of the exponential function exp are

$$\begin{aligned} \exp(x + y) &= \exp(x) \cdot \exp(y), \qquad \forall x, y \in \mathbb{R}, \\ \exp'(x) &= \exp(x), \qquad \forall x \in \mathbb{R}. \end{aligned}$$

The last line means that the derivative of the exponential function is again the exponential function.

Sometimes the following functions are interesting:

$$\cosh(x) := \frac{e^x + e^{-x}}{2}, \qquad \sinh(x) := \frac{e^x - e^{-x}}{2}, \qquad x \in \mathbb{R}.$$

The pronunciations are "hyperbolic cosine" and "hyperbolic sine", respectively. There are people in this world who rhyme cosh with "bosh" and sinh with "shine" — such persons are barbarians.

The inverse function to the exponential function is called *natural logarithm*[7], written as ln.

$$\text{If} \quad x \in \mathbb{R} \quad \text{and} \quad y = \exp(x), \quad \text{then} \quad y > 0 \quad \text{and} \quad x = \ln(y).$$

The crucial rules of the natural logarithm are

$$\begin{aligned} \ln(x \cdot y) &= \ln(x) + \ln(y), \qquad \forall x > 0, \quad \forall y > 0, \\ \ln\left(\frac{x}{y}\right) &= \ln(x) - \ln(y), \qquad \forall x > 0, \quad \forall y > 0, \\ \ln(x^r) &= r \cdot \ln(x), \qquad \forall x > 0, \quad \forall r \in \mathbb{R}, \\ \ln'(x) &= \frac{1}{x}, \quad \forall x > 0. \end{aligned}$$

[6] LEONHARD EULER, 1707–1783, was one of the greatest mathematicians of all times. Please do not mispronounce neither part of his name.

[7] The word *logarithm* comes from the Greek λογαριθμός, which is an artificial word invented in 1614 by JOHN NAPIER (who lived 1550–1617 and is one of the local academic heroes of Edinburgh) from the Greek words λογός meaning *word, reason, ratio* and ἀριθμός meaning *number.*

In particular the first formula has been of highest importance at Napier's time because it allows to reduce multiplying numbers (which was hard at that time) to adding numbers (which is much easier); you only need two tables of values (one table for exp and one table for ln). The invention of logarithms allowed for a tremendous speed-up of calculations, an early example of the impact of mathematics upon society.

Certainly you know the song "Wonderful World" by Sam Cooke whose lyrics contain this verse:

> "...
> Don't know much about geography
> Don't know much trigonometry
> Don't know much about algebra
> Don't know what a slide rule is for
> ..."

Now what is a *slide rule*, and what is it good for ? Slide rules are computing devices that had been used by scientists and engineers over more than 300 years. See Figure 2.1, and also the exhibition in the William Arroll Building.

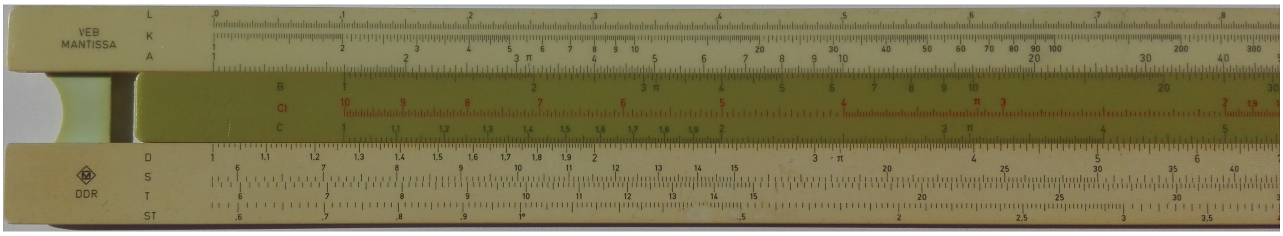


Figure 2.1: A slide rule. Look at the two scales called "A" and "B", and you can read off the identities $1.6 \cdot 2 = 3.2$, $1.6 \cdot 5 = 8$ or $1.6 \cdot 19 \approx 30.5$. Both are logarithmic scales, making it easy to perform multiplication via the identity $\ln(xy) = \ln(x) + \ln(y)$. The scales "L", "K", "D" are for taking logarithms (with base 10), cubes and square roots, the scale "CI" is for taking reciprocals of the numbers on the "D" scale, and "S", "T", "ST" are for sines, tangents, and arcs.

Finally, we mention that also tables for logarithms (with base 10) have been in use over centuries, for instance [7], which really does enable you to look up logarithms with 7 reliable digits. By the way, its compiler, James Pryde, was lecturer in mathematics at the Edinburgh School of Arts, which later became Heriot-Watt University.

### 2.2.4 Trigonometry

The *unit circle* is a circle with radius one and centre $\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$. When we say "circle", we only mean the circular line, not the area inside (should we ever speak about the inside area, we will use the words *ball* or *disk*). Angles will mostly be measured in radians, which equal the arc-lengths of angles (with vertex at the origin) in the unit circle. An angle of size $\frac{\pi}{2}$ is also called a *right angle.*

Now we explain *sines* and *cosines*. Choose an angle $\alpha$ in radians (positive or negative). Its vertex shall be the origin $\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$. On the unit circle, pick the point $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ which is on the 3pm position. From that point, walk along the unit circle. The distance you have to walk shall be $|\alpha|$, and you walk counter-clockwise if $\alpha \geq 0$, and clockwise if $\alpha < 0$. After having finished this walk, you arrive at a point on the unit circle whose coordinates are $\left[\begin{smallmatrix}\cos(\alpha)\\ \sin(\alpha)\end{smallmatrix}\right]$. These are the definitions of $\cos(\alpha)$ and $\sin(\alpha)$, where $\alpha$ can be any real number.

We quickly convince ourselves by walking that the following statements are true, for all real numbers $\alpha$:

$$\begin{aligned}
\cos(-\alpha) &= \cos(\alpha), \\
\sin(-\alpha) &= -\sin(\alpha), \\
\cos(\alpha + \pi) &= -\cos(\alpha), \\
\sin(\alpha + \pi) &= -\sin(\alpha), \\
\cos^2(\alpha) + \sin^2(\alpha) &= 1.
\end{aligned}$$

And if $0 < \alpha < \pi/2$, then you can draw $\alpha$ as an acute angle in a right-angled triangle, and you get

$$\sin(\alpha) = \frac{\text{length of opposite side}}{\text{length of hypotenuse}}, \qquad \cos(\alpha) = \frac{\text{length of adjacent side}}{\text{length of hypotenuse}}.$$

We also define $\tan(\alpha) := \frac{\sin(\alpha)}{\cos(\alpha)}$, which is valid for all $\alpha$ subject to the restriction $\cos(\alpha) \neq 0$.

We make no attempt at defining cotangent, secant and cosecant, because almost nobody except school maths teachers ever uses them[8] .

Values which you have to know are:

| $\alpha$ | $0$ | $\frac{\pi}{6} \triangleq 30°$ | $\frac{\pi}{4} \triangleq 45°$ | $\frac{\pi}{3} \triangleq 60°$ | $\frac{\pi}{2} \triangleq 90°$ |
|---|---|---|---|---|---|
| $\sin(\alpha)$ | $\frac{1}{2}\sqrt{0} = 0$ | $\frac{1}{2}\sqrt{1} = \frac{1}{2}$ | $\frac{1}{2}\sqrt{2}$ | $\frac{1}{2}\sqrt{3}$ | $\frac{1}{2}\sqrt{4} = 1$ |

You are expected to be able to figure out the corresponding values for cos via the rule $\cos(\alpha) = \sin(\pi/2-\alpha)$ and for tan.

We have the following facts:

- The area of a parallelogram with sides $a$ and $b$ and angle $\gamma$ between them equals $ab\cos(\gamma)$.
- The area of a sector of a disk with radius $r$ and angle $\alpha$ is $\frac{1}{2}r^2\alpha$.

Next we consider triangles with vertexes $A$, $B$, $C$ and sides $a$, $b$, $c$, in the sense that $A$ is opposite to $a$ etc., and we have angles $\alpha$, $\beta$, $\gamma$ at $A$, $B$, $C$. Abusing notation, we make the agreement that the letter $a$ not only denotes that side of the triangle, but also its length. Similarly for the angles. With these notations, the following holds:

- The area of this triangle equals $\frac{1}{2}ab\cos(\gamma)$.
- The law of sines:
  $$\frac{a}{\sin(\alpha)} = \frac{b}{\sin(\beta)} = \frac{c}{\sin(\gamma)},$$
  and this equals the diameter of the circumscribed circle.
- The law of cosines:
  $$c^2 = a^2 + b^2 - 2ab\cos(\gamma) \quad \text{etc.}$$

**DIY:** *Prove the laws of sines and cosines.*

Now let us tinker with angles and triangles. Take the unit circle and let its centre be christened $O$. Choose two points $B$ and $C$ on the circle, and then a triangle $\Delta OBC$ appears whose angle at $O$ shall be called $\alpha$, which is positive. Assume that $\alpha$ is small. Then the triangle $\Delta OBC$ has area $\frac{1}{2}\sin(\alpha)$, and the disk sector $\sphericalangle OBC$ has area $\frac{1}{2}\alpha$. We need one more triangle: it has a right angle at $B$ and the angle $\alpha$ at $O$. The missing vertex shall be called $D$. Then $\Delta OBD$ has area $\frac{1}{2}\tan(\alpha)$.

**DIY:** *Draw a picture.*

Because $\Delta OBC$ is contained in $\sphericalangle OBC$ which is contained in $\Delta OBD$, comparing their areas reveals us

$$\sin(\alpha) < \alpha < \tan(\alpha), \quad \text{provided} \quad 0 < \alpha < \frac{\pi}{2}.$$

If we take reciprocals everywhere, the inequalities turn around:

$$\frac{1}{\sin(\alpha)} > \frac{1}{\alpha} > \frac{1}{\tan(\alpha)}, \qquad \forall \alpha \in \left(0, \frac{\pi}{2}\right).$$

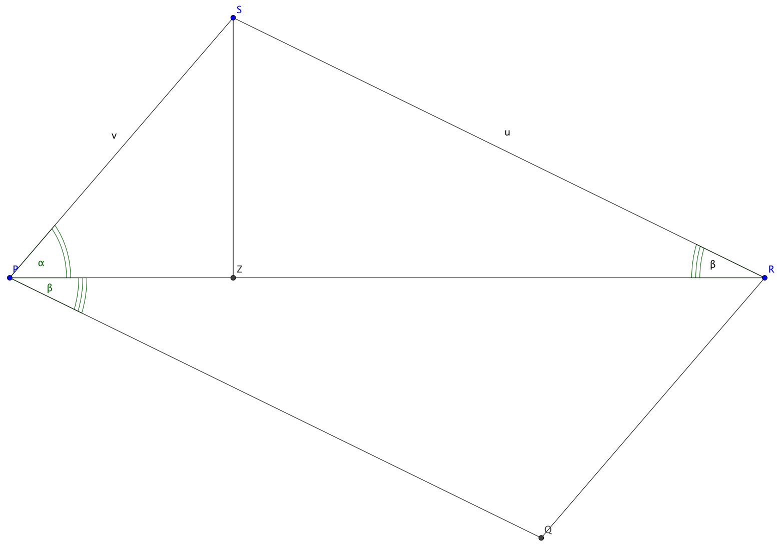
Now we wish to multiply all this by $\sin(\alpha)$, which is allowed because $\sin(\alpha)$ is positive for the mentioned $\alpha$. And we re-order the double inequality:

$$\cos(\alpha) < \frac{\sin(\alpha)}{\alpha} < 1, \qquad \forall \alpha \in \left(0, \frac{\pi}{2}\right).$$

This becomes particularly interesting if $\alpha$ is very small, something like $\alpha \approx 0.001$: then $\cos(\alpha) \approx 1$, and therefore $\frac{\sin(\alpha)}{\alpha} \approx 1$. This is definitely not obvious; because in this fraction the bottom part is basically zero, and the top part is also basically zero as well, and therefore it is initially hard to guess what the approximate value of this fraction would be. But now we know that the value of that fraction approximates 1 if $\alpha$ is close to 0. We will come back to this when we calculate the derivative of $\sin(x)$.

[8] In fact, outside the UK these functions are basically unknown. Why is the expression $\frac{1}{\sin}$ so important to deserve a name of its own ?

**DIY:** *Have a look at this parallelogram:*



*Calculate its area in two different ways (perhaps you will need several formulas for the height $ZS$). Deduce from these two formulas for the area then the formula*

$$\sin(\alpha+\beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta), \qquad \forall \alpha, \beta \in \left(0, \frac{\pi}{2}\right). \tag{2.7}$$

This formula is valid for all $\alpha$ and $\beta$ from $\mathbb{R}$, not just from the interval $(0, \frac{\pi}{2})$.

It would be advantageous to have a similar formula for $\cos(\alpha+\beta)$ as well. What can we do ? Well, we can reduce cos(stuff) to $\sin(\frac{\pi}{2} - \text{stuff})$ and then hope for the best:

$$\begin{aligned}
\cos(\alpha+\beta) &= \sin\left(\frac{\pi}{2} - (\alpha+\beta)\right) = \sin\left(\left(\frac{\pi}{2} - \alpha\right) + (-\beta)\right) && \Big| \text{ apply now the formula (2.7)} \\
&= \sin\left(\frac{\pi}{2} - \alpha\right)\cos(-\beta) + \cos\left(\frac{\pi}{2} - \alpha\right)\sin(-\beta) && \Big| \text{ apply formulas you certainly know} \\
&= \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta).
\end{aligned}$$

The inverse functions to sin, cos, tan are arcsin, arccos, arctan (pronounced *arcus sine*, *arcus cosine*, *arcus tangent*). In order to determine the value of $\arcsin(-0.3)$, we have to find all those $y \in \mathbb{R}$ for which $\sin(y) = -0.3$. Out of all those $y$, one of them is being selected, which will then be the value of $\arcsin(-0.3)$. This selection is necessary because the function arcsin cannot have two or more values at $x = -0.3$. The standard selection procedure is called *principal branch* of arcsin, arccos, arctan, compare the Figure 2.2 and 2.3:

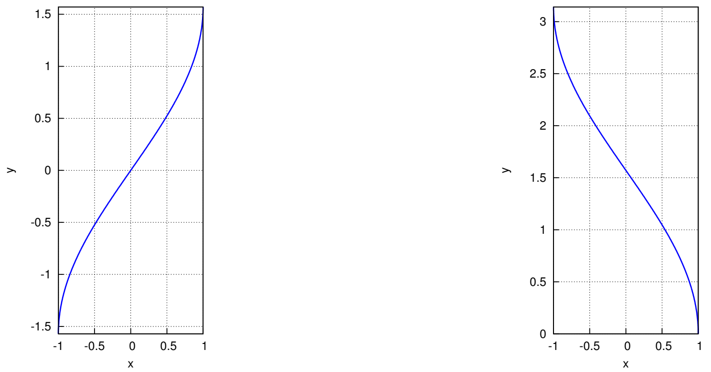


Figure 2.2: The graphs of the principal branches of the functions $y = \arcsin(x)$, $y = \arccos(x)$.

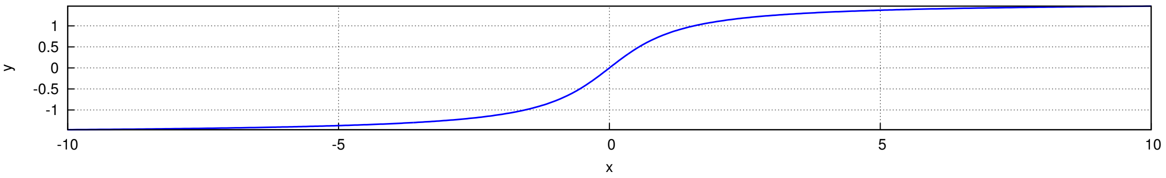


Figure 2.3: The graph of the principal branch of the function $y = \arctan(x)$.

Finally, we mention that the derivatives are

$$\frac{\mathrm{d}\sin(x)}{\mathrm{d}x} = \sin'(x) = \cos(x), \qquad \frac{\mathrm{d}\cos(x)}{\mathrm{d}x} = \cos'(x) = -\sin(x), \qquad \frac{\mathrm{d}\tan(x)}{\mathrm{d}x} = \tan'(x) = 1 + \tan^2(x).$$

### 2.2.5 Geometry

#### What are Points ? What are Vectors ?

What is $\mathbb{R}^3$ ? It is the set of all points

$$P = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

where $x_1$ and $x_2$ and $x_3$ are real numbers. These entries are called the *coordinates* of that point $P$.

Now let us consider a more general situation: the set $\mathbb{R}^n$ contains all points

$$P = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [x_1, x_2, \dots, x_n]^\top \quad \text{where the } x_j \text{ are arbitrary real numbers.}$$

The second notation as a horizontal row with a $\top$ at the right upper corner saves space and means the same thing. For $n = 1$ we obtain the number line $\mathbb{R}^1 = \mathbb{R}$, for $n = 2$ we obtain a plane, and for $n = 3$ we obtain the usual space that surrounds us.

Now let us be given two points $P$ and $Q$ from $\mathbb{R}^n$. We ask for that displacement that shifts $Q$ onto $P$. This displacement is given by a vector $\vec{x}$ whose coordinates can be found like this:

$$\vec{x} = P - Q,$$

where $P$ corresponds to the tip of the arrow, and $Q$ to the rear end of the arrow. The calculation is done as follows:

$$\text{If} \quad P = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix} \quad \text{then} \quad \vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} p_1 - q_1 \\ p_2 - q_2 \\ \vdots \\ p_n - q_n \end{pmatrix} \quad \text{shifts } Q \text{ to } P.$$

We use square brackets $\Big[\;\Big]$ for points and round brackets $\Big(\;\Big)$ for displacement.

A vector $\vec{x}$ is uniquely determined by its direction and its length. The length of a vector is also called *norm* of that vector. Different starting points can correspond to the same vector, compare Figure 2.4.

#### What Can we Do With Points ? And With Vectors ?

There is not much what you can do with **two points**: you can not add them or multiply them because it makes no sense. But on the other hand, if you have two points, then you can ask for that displacement that shifts the first point to the second point. We have introduced this operation above as $\vec{x} = P - Q$.

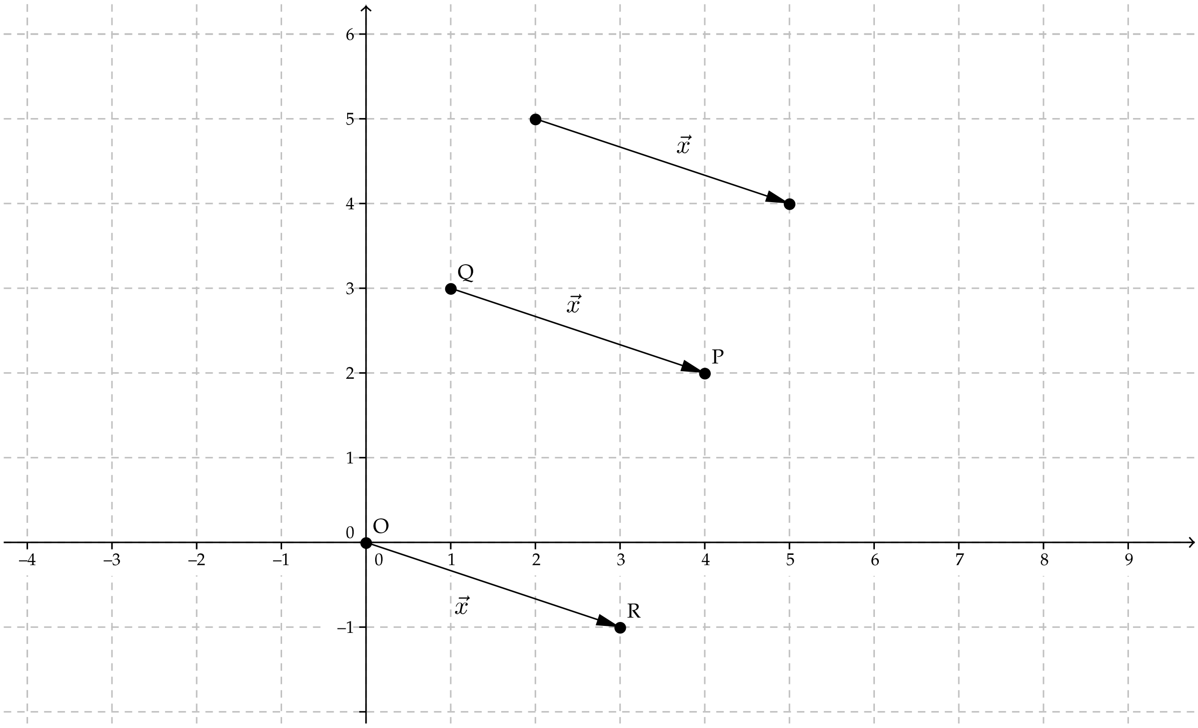


Figure 2.4: Each of the three arrows is the same vector $\vec{x} = (3, -1)^\top$.

Now let us be given **a point and a vector**: we can read the vector as a displacement which we apply to the point. For instance if the point $P = \left[\begin{smallmatrix} 2 \\ 3 \end{smallmatrix}\right]$ and the vector $\vec{x} = \left(\begin{smallmatrix} 7 \\ -1 \end{smallmatrix}\right)$ are given, then $P + \vec{x}$ means that point at which we arrive when we shift $P$ seven steps to the right and one step down:

$$\begin{bmatrix} 2 \\ 3 \end{bmatrix} + \begin{pmatrix} 7 \\ -1 \end{pmatrix} = \begin{bmatrix} 9 \\ 2 \end{bmatrix}.$$

Adding a point and a vector produces a point.

The next situation is this: we have **two vectors**, and each of them is construed as a displacement. For instance, $\vec{x} = \left(\begin{smallmatrix} 1 \\ -5 \end{smallmatrix}\right)$ shifts one grid position to the right and 5 grid positions down, and $\vec{y} = \left(\begin{smallmatrix} -3 \\ 4 \end{smallmatrix}\right)$ shifts three grid positions to the left and 4 grid positions upwards. Nobody prevents us from first performing the displacement $\vec{x}$ and afterwards performing the displacement $\vec{y}$. The result then is another displacement, which can be calculated like this:

$$\begin{pmatrix} 1 \\ -5 \end{pmatrix} + \begin{pmatrix} -3 \\ 4 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \end{pmatrix}.$$

Adding two vectors produces again a vector.

And finally, let us be given **a number and a vector**. For instance the number 7 and the vector $\vec{x} = \left(\begin{smallmatrix} 2 \\ 3 \end{smallmatrix}\right)$. The vector $\vec{x}$ stands for a displacement. Now we could — if we wish so — shift a point $P$ in direction as pointed by the vector $\vec{x}$, but over a distance seven times as long as the length of $\vec{x}$. This is the meaning of the multiplication of a vector by a number:

$$7 \cdot \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 14 \\ 21 \end{pmatrix}.$$

Multiplying a number times a vector produces a vector.

### How to Describe Straight Lines and Planes in $\mathbb{R}^2$ and $\mathbb{R}^3$

Let us change our notation: instead of $x_1$, $x_2$, $x_3$, we will now write $x$, $y$, $z$ because this is more familiar.

A **line** $\mathcal{L}$ **in** $\mathbb{R}^2$ can be handled mathematically at least in two ways. One way is by means of an equation: all the real numbers $x$ and $y$ that solve the equation $3x + 4y = 7$ are the coordinates of certain points

$P = \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$ that are located along a straight line $\mathcal{L}$ in the usual coordinate system. The vector $\left(\begin{smallmatrix} 3 \\ 4 \end{smallmatrix}\right)$ is perpendicular to that straight line $\mathcal{L}$.

**DIY:** *Check these two statements in a diagram.*

And the second way uses vectors: consider a point $P_0 = \left[\begin{smallmatrix} 2 \\ 3 \end{smallmatrix}\right]$ and a vector $\vec{v} = \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)$. If we now let a real number $t$ run through all the real numbers, from $-\infty$ to $+\infty$, then all the points

$$P = P_0 + t \cdot \vec{v} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} + t \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{bmatrix} 2+t \\ 3+t \end{bmatrix}$$

are located along a straight line $\mathcal{L}$, and the vector $\vec{v}$ is parallel to that line $\mathcal{L}$.

**DIY:** *Check these two statements in a diagram.*

A **line** $\mathcal{L}$ **in** $\mathbb{R}^3$ is being handled mathematically typically only in one way, using a point $P_0$ and a vector $\vec{v}$: consider $P_0 = [2, 1, 3]^\top$ and $\vec{v} = (1, 0, 2)^\top$. If we now let a real number $t$ run through all the real numbers, from $-\infty$ to $+\infty$, then all the points

$$P = P_0 + t \cdot \vec{v} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} + t \cdot \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} = \begin{bmatrix} 2+t \\ 1 \\ 3+2t \end{bmatrix}$$

are located along a straight line $\mathcal{L}$, and the vector $\vec{v}$ is parallel to that line $\mathcal{L}$.

Finally, we consider the case of a **plane** $\mathcal{P}$ **in** $\mathbb{R}^3$**:** there are at least two ways. The first one is by means of an equation: all the real numbers $x$, $y$, $z$ that solve the equation $3x + 4y - 5z = 17$ are the coordinates of certain points $P = [x, y, z]^\top$ that are located on a certain plane $\mathcal{P}$ in $\mathbb{R}^3$. In our situation, this plane $\mathcal{P}$ goes through the three points

$$Q = \begin{bmatrix} \frac{17}{3} \\ 0 \\ 0 \end{bmatrix}, \qquad R = \begin{bmatrix} 0 \\ \frac{17}{4} \\ 0 \end{bmatrix}, \qquad S = \begin{bmatrix} 0 \\ 0 \\ \frac{17}{-5} \end{bmatrix}.$$

You can find these three points on the coordinate axes. Additionally, the vector $(3, 4, -5)^\top$ is perpendicular to the plane $\mathcal{P}$.

The second way of describing a plane in $\mathbb{R}^3$ involves a point $P_0$ on that plane and two vectors $\vec{v}$, $\vec{w}$ parallel to that plane. To have an example, let us consider that above plane $\mathcal{P}$. We choose

$$P_0 := Q = \begin{bmatrix} \frac{17}{3} \\ 0 \\ 0 \end{bmatrix}, \quad \vec{v} = R - Q = \begin{bmatrix} 0 \\ \frac{17}{4} \\ 0 \end{bmatrix} - \begin{bmatrix} \frac{17}{3} \\ 0 \\ 0 \end{bmatrix} = \begin{pmatrix} -\frac{17}{3} \\ \frac{17}{4} \\ 0 \end{pmatrix}, \quad \vec{w} = S - Q = \begin{bmatrix} 0 \\ 0 \\ \frac{17}{-5} \end{bmatrix} - \begin{bmatrix} \frac{17}{3} \\ 0 \\ 0 \end{bmatrix} = \begin{pmatrix} -\frac{17}{3} \\ 0 \\ \frac{17}{-5} \end{pmatrix},$$

and then the plane $\mathcal{P}$ contains all points $P$ that can be written as

$$P = P_0 + t \cdot \vec{v} + s \cdot \vec{w} = \begin{bmatrix} \frac{17}{3} \\ 0 \\ 0 \end{bmatrix} + t \cdot \begin{pmatrix} -\frac{17}{3} \\ \frac{17}{4} \\ 0 \end{pmatrix} + s \cdot \begin{pmatrix} -\frac{17}{3} \\ 0 \\ \frac{17}{-5} \end{pmatrix} = \begin{pmatrix} \frac{17}{3} \cdot (1 - t - s) \\ \frac{17}{4} \cdot t \\ \frac{17}{-5} \cdot s \end{pmatrix}.$$

Here the parameters $t$ and $s$ run independently from each other through $\mathbb{R}$, hence from $-\infty$ to $+\infty$. For instance, if you choose $t = s = 0$, then you get the point $Q$. If you choose $t = 1$ and $s = 0$, then you get the point $R$. And if you choose $t = 0$ and $s = 1$, then you obtain the point $S$.

### What is the Geometrical Meaning of Solving Linear Systems

**Example:** *We wish to find the intersection points of three planes $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$ in $\mathbb{R}^3$ that are described by the following three equations:*

$$\left.\begin{array}{lrl} \mathcal{P}_1: & 2x + \phantom{2}y + \phantom{3}z = 1 & \boxed{\textit{1}} \\ \mathcal{P}_2: & 3x + \phantom{2}y + \phantom{3}z = 2 & \boxed{\textit{2}} \\ \mathcal{P}_3: & 4x + 2y + 3z = 0 & \boxed{\textit{3}} \end{array}\right\} \qquad (2.8)$$

***Wanted*** *are the common solutions of all three equations, or, in other words, all solutions of the system* (2.8) *of equations. Compare Figure 2.5.*

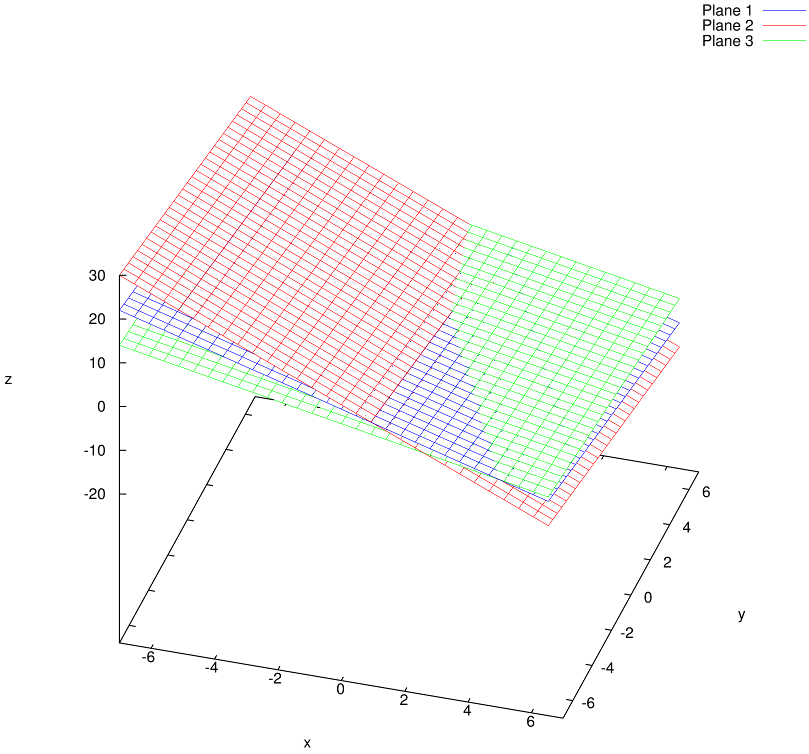


Figure 2.5: The three planes $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$ intersect in exactly one point. This point is the unique solution of the system (2.8).

*This is how to do it: Calculating* $2\cdot\boxed{\mathit{2}}-3\cdot\boxed{\mathit{1}}$ *yields:*

$$\begin{array}{rcl|l} 2\cdot\boxed{\mathit{2}} & : & 6x+2y+2z=4 & \\ 3\cdot\boxed{\mathit{1}} & : & 6x+3y+3z=3 & - \\ \hline & & -y-z=1 & \boxed{\mathit{4}} \end{array}$$

*Now we compute like this:*

$$\begin{array}{rcl|l} \boxed{\mathit{3}} & : & 4x+2y+3z=0 & \\ 2\cdot\boxed{\mathit{1}} & : & 4x+2y+2z=2 & - \\ \hline & & z=-2 & \boxed{\mathit{5}} \end{array}$$

*Substituting* $\boxed{\mathit{5}}$ *into* $\boxed{\mathit{4}}$ *reveals* $-y+2=1$, *hence* $y=1$. *Substituting this* $\boxed{\mathit{1}}$ *yields* $2x+1-2=1$, *therefore* $x=1$.

*The intersection point is* $P=[1,1,-2]^\top$, *and the system is uniquely solvable. Geometrically, this means that there is exactly one point who is simultaneously on all three planes.*

**Example:** *Now we look for all intersection points of two planes* $\mathcal{P}_1$ *and* $\mathcal{P}_2$ *in* $\mathbb{R}^3$, *compare Figure 2.6:*

$$\left.\begin{array}{lll} \mathcal{P}_1: & 2x+y+z=1 & \boxed{\mathit{1}} \\ \mathcal{P}_1: & 3x+y+z=2 & \boxed{\mathit{2}} \end{array}\right\} \tag{2.9}$$

*As above, calculating* $2\cdot\boxed{\mathit{2}}-3\cdot\boxed{\mathit{1}}$ *brings us to the equation*

$$-y-z=1 \quad \boxed{\mathit{4}},$$

*and there are no further conditions. We introduce some new variable* $t$ *that runs through* $\mathbb{R}$ *and set* $z=t$. *Then* $\boxed{\mathit{4}}$ *implies*

$$-y-t=1, \quad \textit{hence} \quad y=-1-t.$$

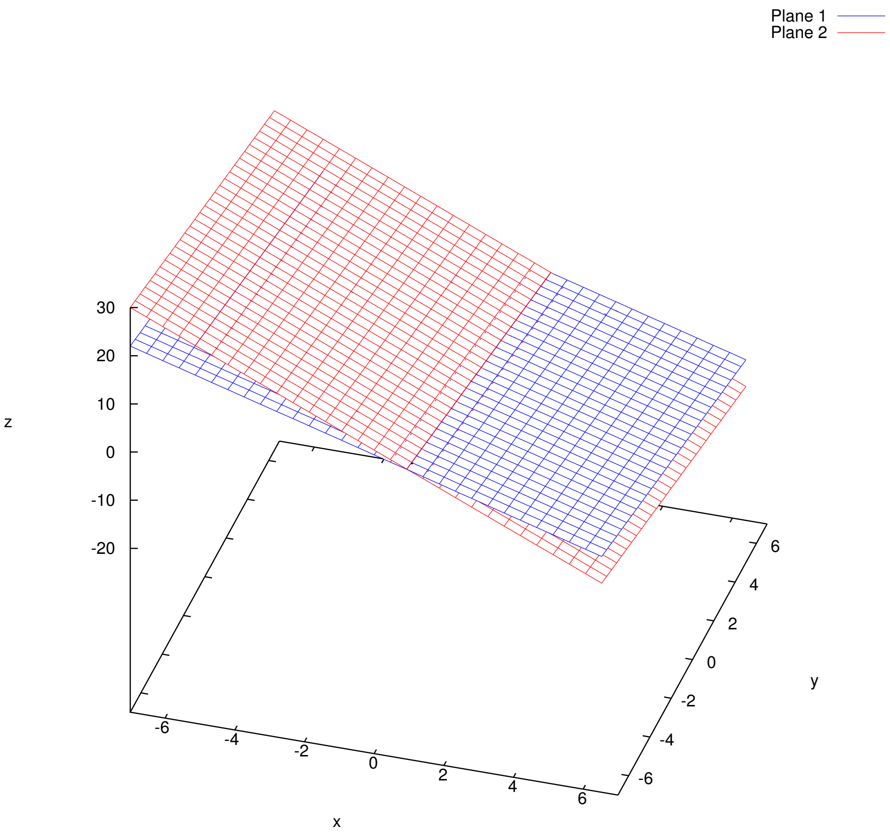


Figure 2.6: The two planes intersect along one straight line. Each point on that line is a solution to the system (2.9).

*And from* $\boxed{1}$ *we then obtain*

$$1 = 2x + (-1 - t) + t = 2x - 1, \quad \textit{hence} \quad x = 1.$$

*Consequently, all intersection points $P$ of the two planes $\mathcal{P}_1$ and $\mathcal{P}_2$ are given as follows:*

$$P = \begin{bmatrix} 1 \\ -1-t \\ t \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} + t \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}, \quad t \in \mathbb{R}.$$

*The set of solutions of the system (2.9) forms a straight line. Along this line, we always have $x = 1$, which is also (more or less) visible in Figure 2.6.*

**Example:** *Again we wish to find all the intersection points of three planes, but now we replace the plane $\mathcal{P}_3$ against plane $\mathcal{P}_4$:*

$$\left.\begin{array}{lrl} \mathcal{P}_1: & 2x + y + z = 1 & \boxed{1} \\ \mathcal{P}_2: & 3x + y + z = 2 & \boxed{2} \\ \mathcal{P}_4: & x + y + z = 3 & \boxed{3'} \end{array}\right\} \tag{2.10}$$

*As above, we calculate* $2 \cdot \boxed{2} - 3 \cdot \boxed{1}$ *and find*

$$-y - z = 1 \qquad \boxed{4}.$$

*But we can also calculate* $2 \cdot \boxed{3'} - \boxed{1}$, *which brings us to*

$$y + z = 5 \qquad \boxed{5}.$$

*Now we add* $\boxed{4}+\boxed{5}$, *and a surprise awaits us:*

$$0 = 6.$$

*This is not possible, and therefore the system (2.10) possesses no solution. See Figure 2.7.*

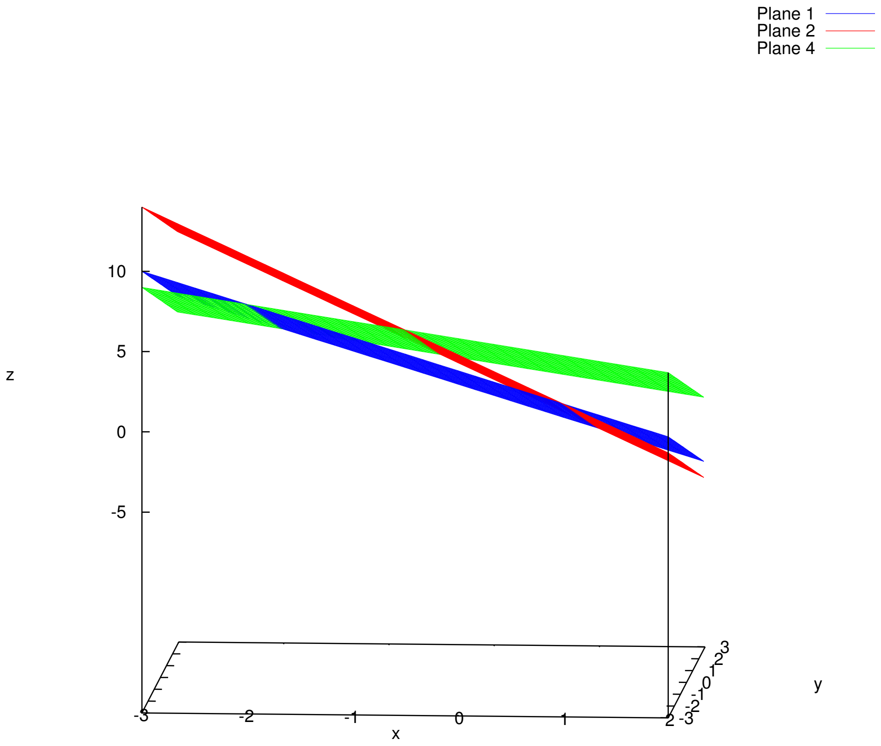


Figure 2.7: The three planes have no point in common. There is an intersection line formed by $\mathcal{P}_2$ and $\mathcal{P}_4$, but the plane $\mathcal{P}_1$ is parallel to that intersection line. Hence the solution set to the system (2.10) is the empty set.

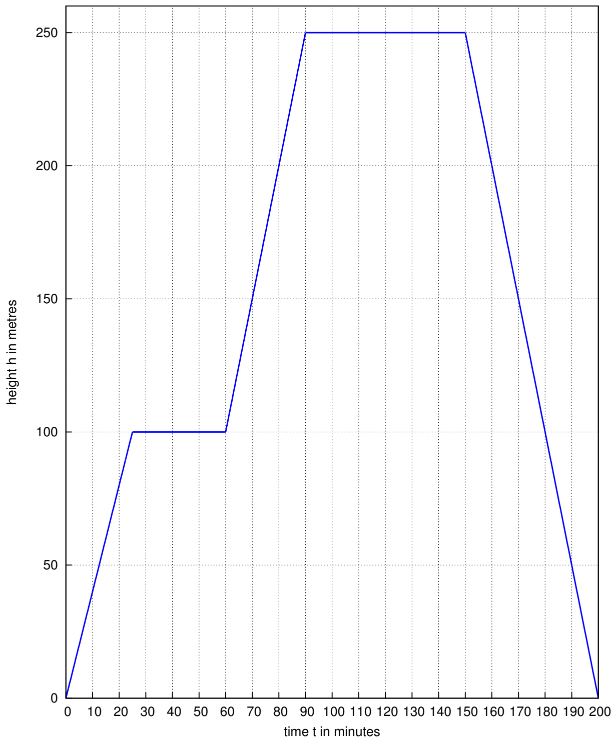
## 2.3 Some Key Ideas of Calculus and Geometry

### 2.3.1 Taking Derivatives

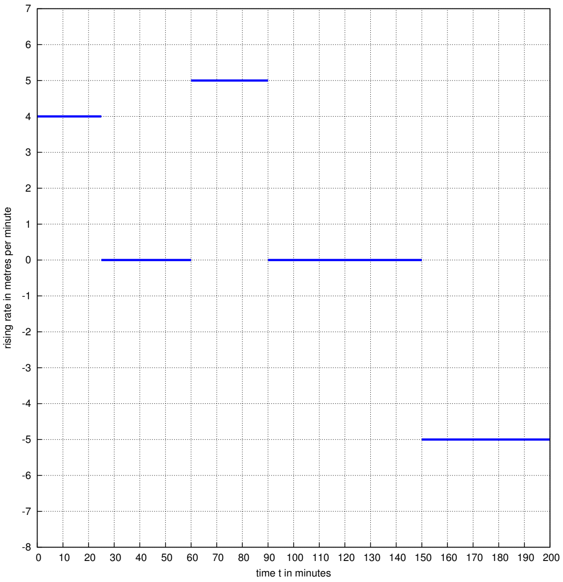
**What are Derivatives ?**

Suppose the sun has been shining for many hours, a veritable heatwave with temperatures up to 23°C has arrived at Scotland, and a person takes the health risk of staying outdoors all day, and is even so ambitious as to take a balloon ride. The diagram on the right shows the altitude of the balloon above ground as a function of time.

This is the graph of a function $h = h(t)$, and this function is piecewise affine linear.



The “rising rate” is a function depending on time that tells us how quick the height $h(t)$ changes at the point of time $t$. Recall the definition of a slope of an affine linear function. Then the rising rate has the following diagram:



This is the graph of the derivative $h'(t)$ (it should be noted that — pedantically — we should exclude those points of time where $h'(t)$ jumps up or down).

The derivative $h'(t)$ of a function $h(t)$ is a tool that tells us how quick this function $h$ changes near that chosen $t$.

### How to Determine the Derivative of a Function $f$

A function can be given in three ways:

- as a table of values
- as a graph
- by a formulaic expression

### If $f$ is given as a value table only:

Assume that there is a function $f$ of which we have only a small amount of knowledge — only a list of values such as this:

| $x$ | $f(x)$ |
|---|---|
| 0.0 | 1.57079632679 |
| 0.05 | 1.52077546999 |
| 0.1 | 1.47062890563 |
| 0.15 | 1.42022805402 |
| 0.2 | 1.369438406 |
| 0.25 | 1.31811607165 |
| 0.3 | 1.26610367278 |
| 0.35 | 1.21322522315 |
| 0.4 | 1.15927948073 |
| 0.45 | 1.10403098775 |
| 0.5 | 1.0471975512 |
| 0.55 | 0.988432088926 |
| 0.6 | 0.927295218002 |
| 0.65 | 0.86321189007 |
| 0.7 | 0.795398830184 |
| 0.75 | 0.722734247813 |

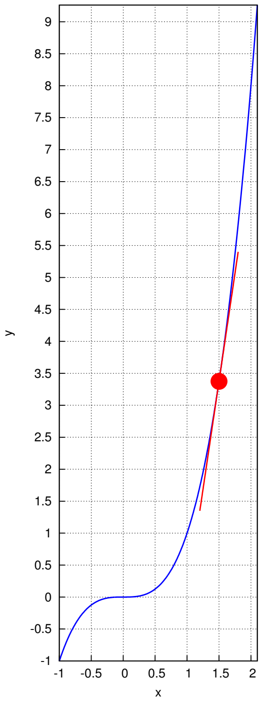
Somebody asks you for $f'(0.35)$. The only thing you can do is some approximation and hope for the best:

$$f'(0.35) \approx \frac{f(0.4) - f(0.35)}{0.4 - 0.35} = \frac{1.1592794807 - 1.21322522315}{0.05} = -1.0789\ldots.$$

and we do not know how good this approximation is.

**If $f$ is given as a graph only:**

Assume that there is a function $f$ of which we only possess a graph such as this:



and we have the task of finding $f'(1.5)$ approximately. The only thing you can do is to pick the point $\left[\begin{smallmatrix}1.5\\ f(1.5)\end{smallmatrix}\right]$ (which is red in the diagram), draw there a short tangent line (guided by visual intuition), measure there the rising angle $\alpha$ of the tangent line, and then calculate $\tan(\alpha)$. The reason for this approach is that the red tangent line is the graph of an affine linear function, and therefore its slope is defined as

$$\text{slope of red line} = \frac{\text{rise}}{\text{run}},$$

which is also the formula for $\tan(\alpha)$. In our case, $\alpha \approx 81°$, hence $f'(1.5) \approx \tan(81°) = 6.31\ldots$.
A warning seems appropriate here: if you measure angles, make sure that both axes are spaced with the same distances. Otherwise the angle will be wrong because the diagram is distorted.

**If $f$ is given as a formula:**

We know a formula for a function $f$, and we know a specific point $x$. Since this specific point is something special, we call it $x_*$ instead of just $x$ (stars are always something special).

We wish to determine $f'(x_*)$. If $f$ were given as a graph, we would draw a tangent[9] line at the graph through the point $\left[\begin{smallmatrix}x_*\\ f(x_*)\end{smallmatrix}\right]$ in such a way that "it seems to look like a tangent line". Such an approach is hand-wavy, but there is only so much you can do if all you have available is the graph.

But now we have more than a graph — we even have a formula for the function $f$. How can we determine now $f'(x_*)$, which is to be understood as the slope of the tangent line through $\left[\begin{smallmatrix}x_*\\ f(x_*)\end{smallmatrix}\right]$ ?

This is how to do it: we take another point $\left[\begin{smallmatrix}x\\ f(x)\end{smallmatrix}\right]$ which is very close to the point $\left[\begin{smallmatrix}x_*\\ f(x_*)\end{smallmatrix}\right]$ we are interested in. Two distinct points always determine a straight line, which is called *secant line*[10] (not to be confused with the useless trigonometric function $\sec(\alpha)$). The intuition is that the secant line and the tangent line should have almost the same slope if $x$ is very near to $x_*$. The error is expected to become arbitrarily small if $x$ approaches $x_*$. Now we have

$$\text{slope of secant line} = \frac{\text{rise}}{\text{run}} = \frac{f(x) - f(x_*)}{x - x_*}.$$

And therefore

$$f'(x_*) \approx \frac{f(x) - f(x_*)}{x - x_*} \qquad \text{if} \qquad x \approx x_*. \tag{2.11}$$

[9] *tangent* comes from the Latin *tangens*, *tangentis*, which is the present participle of *tango*, *tangere*, *tetigi*, *tactus* which means *to touch*.

[10] The Latin verb here is *seco*, *secare*, *sectus* which means *to cut*. The present participle *secans*, *secantis* then means *cutting*. A secant line cuts through a circle.

We remark that the scientific definition of $f'(x_*)$ is

$$f'(x_*) = \lim_{x \to x_*} \frac{f(x) - f(x_*)}{x - x_*},$$

and the courses on *Calculus* and *Analysis* will explain what this actually means. We may rewrite (2.11) as

$$f'(x_*) \approx \frac{f(x_* + k) - f(x_*)}{k} \qquad \text{if} \qquad k \approx 0. \tag{2.12}$$

**Easy Examples**

Consider the function $f(x) = \sqrt{x}$ for $x > 0$. We wish to figure out its derivative, and a deeper understanding will be obtained if we utilise all three representations of the square root function.

**As a list of values:** The square root function has the left value list,

| $x$ | $f(x) = \sqrt{x}$ |
|---|---|
| 0.0 | 0.0 |
| 0.05 | 0.22360679775 |
| 0.1 | 0.316227766017 |
| 0.15 | 0.387298334621 |
| 0.2 | 0.4472135955 |
| 0.25 | 0.5 |
| 0.3 | 0.547722557505 |
| 0.35 | 0.59160797831 |
| 0.4 | 0.632455532034 |
| 0.45 | 0.67082039325 |
| 0.5 | 0.707106781187 |
| 0.55 | 0.74161984871 |
| 0.6 | 0.774596669241 |
| 0.65 | 0.80622577483 |
| 0.7 | 0.836660026534 |
| 0.75 | 0.866025403784 |
| 0.8 | 0.894427191 |
| 0.85 | 0.921954445729 |
| 0.9 | 0.948683298051 |
| 0.95 | 0.974679434481 |
| 1.0 | 1.0 |
| 1.05 | 1.0246950766 |
| 1.1 | 1.04880884817 |
| 1.15 | 1.07238052948 |
| 1.2 | 1.09544511501 |
| 1.25 | 1.11803398875 |
| 1.3 | 1.1401754251 |
| 1.35 | 1.16189500386 |
| 1.4 | 1.18321595662 |
| 1.45 | 1.20415945788 |
| 1.5 | 1.22474487139 |
| 1.55 | 1.2449899598 |
| 1.6 | 1.26491106407 |
| 1.65 | 1.28452325787 |
| 1.7 | 1.30384048104 |
| 1.75 | 1.32287565553 |
| 1.8 | 1.3416407865 |
| 1.85 | 1.36014705087 |
| 1.9 | 1.37840487521 |
| 1.95 | 1.39642400438 |
| 2.0 | 1.41421356237 |

and according to the rule

$$f'(0.25) \approx \frac{f(0.3) - f(0.25)}{0.3 - 0.25}$$

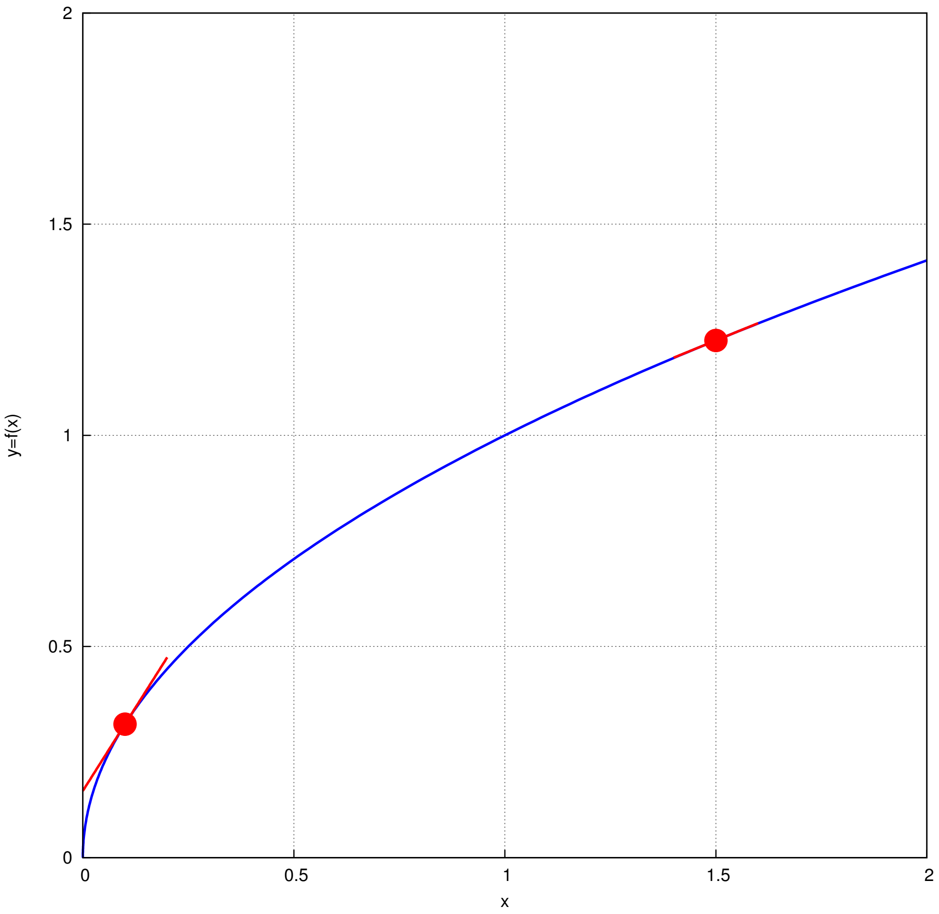
we then have a second list for the approximate values of the derivatives of the square root function:

| $x$ | $f'(x) \approx$ |
|---|---|
| 0.0 | 4.4721 |
| 0.05 | 1.8524 |
| 0.1 | 1.4214 |
| 0.15 | 1.1983 |
| 0.2 | 1.0557 |
| 0.25 | 0.9544 |
| 0.3 | 0.8777 |
| 0.35 | 0.8169 |
| 0.4 | 0.7672 |
| 0.45 | 0.7257 |
| 0.5 | 0.6902 |
| 0.55 | 0.6595 |
| 0.6 | 0.6325 |
| 0.65 | 0.6086 |
| 0.7 | 0.5873 |
| 0.75 | 0.5680 |
| 0.8 | 0.5505 |
| 0.85 | 0.5345 |
| 0.9 | 0.5199 |
| 0.95 | 0.5064 |
| 1.0 | 0.4939 |
| 1.05 | 0.4822 |
| 1.1 | 0.4714 |
| 1.15 | 0.4612 |
| 1.2 | 0.4517 |
| 1.25 | 0.4428 |
| 1.3 | 0.4343 |
| 1.35 | 0.4264 |
| 1.4 | 0.4188 |
| 1.45 | 0.4117 |
| 1.5 | 0.4049 |
| 1.55 | 0.3984 |
| 1.6 | 0.3922 |
| 1.65 | 0.3863 |
| 1.7 | 0.3807 |
| 1.75 | 0.3753 |
| 1.8 | 0.3701 |
| 1.85 | 0.3651 |
| 1.9 | 0.3603 |
| 1.95 | 0.3557 |

**DIY:** *Visualise the second table as dots in a diagram.*

**DIY:** *Why are there only 4 digits after the decimal point ?*

**As a graph:** the graph of the square root function is



and we observe that the first tangent at $x = 0.1$ is quite steep, the other tangent at $x = 1.5$ is less steep (and at $x = 1.5$, the blue graph and the red tangent are hard to discern, but that is another topic). Hence $f'(0.1)$ is positive and big, and $f'(1.5)$ is positive and not so big. On the other hand, $f'(0.001)$ appears to be brobdingnagian.

**As a formula:** We know $f(x) = \sqrt{x}$, and we have picked some $x_* > 0$. Looking at (2.12), the key question is now: what can be said about

$$\frac{f(x_* + k) - f(x_*)}{k} = \frac{\sqrt{x_* + k} - \sqrt{x_*}}{k} \qquad \text{for} \quad k \approx 0\ ?$$

We recall the formula $(a - b) \cdot (a + b) = a^2 - b^2$. Put $a = \sqrt{x_* + k}$ and $b = \sqrt{x_*}$. Substituting these values into $a - b = \frac{a^2 - b^2}{a+b}$ gives

$$\sqrt{x_* + k} - \sqrt{x_*} = \frac{\left(\sqrt{x_* + k}\right)^2 - \left(\sqrt{x_*}\right)^2}{\sqrt{x_* + k} + \sqrt{x_*}} = \frac{x_* + k - x_*}{\sqrt{x_* + k} + \sqrt{x_*}} = \frac{k}{\sqrt{x_* + k} + \sqrt{x_*}},$$

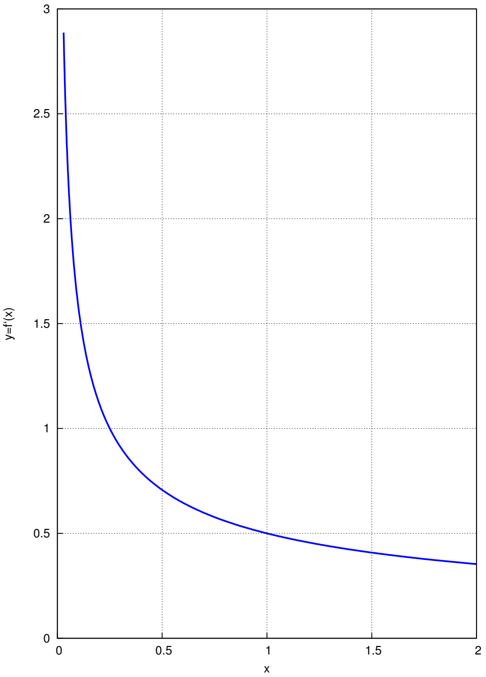
and therefore the secant line slope is

$$\frac{f(x_* + k) - f(x_*)}{k} = \frac{k}{k\left(\sqrt{x_* + k} + \sqrt{x_*}\right)} = \frac{1}{\sqrt{x_* + k} + \sqrt{x_*}}.$$

Now if $k$ goes to zero, the term $\sqrt{x_* + k}$ runs to $\sqrt{x_*}$, and hence we get

$$f'(x_*) = \lim_{k \to 0} \frac{f(x_* + k) - f(x_*)}{k} = \lim_{k \to 0} \frac{1}{\sqrt{x_* + k} + \sqrt{x_*}} = \frac{1}{2\sqrt{x_*}}.$$

The graph of the function $x \mapsto \frac{1}{2\sqrt{x}}$ is this one:



This coincides beautifully with the first two approaches.

**As an application:** how to calculate $\sqrt{150}$ mentally ?

The formula $f'(x_*) \approx \frac{f(x_*+k)-f(x_*)}{k}$ can be re-arranged to

$$f(x_* + k) \approx f(x_*) + k \cdot f'(x_*). \tag{2.13}$$

Everybody knows that $\sqrt{144} = 12$. Now choose $x_* = 144$, $k = 6$, $f(x) = \sqrt{x}$. Then

$$\sqrt{150} \approx \sqrt{144} + 6 \cdot \frac{1}{2\sqrt{144}} = 12 + 6 \cdot \frac{1}{2 \cdot 12} = 12.25.$$

The exact value is $\sqrt{150} = 12.24744\ldots$, so our error is $\approx 0.003$ in absolute terms, and in relative numbers, the error is

$$\frac{\left|12.25 - 12.24744\right|}{12.24744} \cdot 100\% = 0.021\%.$$

Not bad, considering that we have arrived at the approximate value 12.25 without a calculator.

**Another example:** what is $\sin'(x_*)$ ?

Again we need a better understanding of the fraction

$$\frac{\sin(x_* + k) - \sin(x_*)}{k} \qquad \text{for} \quad k \approx 0.$$

Recall $\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)$, hence

$$\begin{aligned}\frac{\sin(x_* + k) - \sin(x_*)}{k} &= \frac{\sin(x_*)\cos(k) + \cos(x_*)\sin(k) - \sin(x_*)}{k} \\ &= \sin(x_*) \cdot \frac{\cos(k) - 1}{k} + \cos(x_*) \cdot \frac{\sin(k)}{k}.\end{aligned}$$

We remember that the second fraction $\frac{\sin(k)}{k}$ approaches 1 if $k$ approaches 0. For the first fraction, we need some clever idea, which is again the formula $a - b = \frac{a^2-b^2}{a+b}$. Consequently,

$$\frac{\cos(k) - 1}{k} = \frac{\Big(\cos(k)\Big)^2 - 1}{k \cdot (\cos(k) + 1)} = \frac{-\sin^2(k)}{k \cdot (\cos(k) + 1)} = -\frac{\sin(k)}{k} \cdot \sin(k) \cdot \frac{1}{\cos(k) + 1},$$

and for $k \approx 0$, this is approximately equal to $-1 \cdot 0 \cdot \frac{1}{1+1} = 0$, and therefore

$$\frac{\sin(x_* + k) - \sin(x_*)}{k} \approx \sin(x_*) \cdot 0 + \cos(x_*) \cdot 1 = \cos(x_*).$$

This sketches the proof of $\sin'(x) = \cos(x)$, for all $x \in \mathbb{R}$.

Similarly, we can show $\cos'(x) = -\sin(x)$, for all $x \in \mathbb{R}$.

### 2.3.2 Working with Derivatives

**The Derivatives of Inverse Functions**

What is the derivative of $\arccos(x)$ ? Answering this becomes easier when we recall what the statement $y = \arccos(x)$ actually means: it means $x = \cos(y)$ and $0 \leq y \leq \pi$. This specification of the interval $[0, \pi]$ comes from our choice of the principal branch of the arcus cosine.

Now we are ready to start.

**Looking at the arcus cosine as a list of values:**

We have a table for $x = \cos(y)$:

| $y$ | $x = \cos(y)$ |
|---|---|
| 0.0 | 1.0 |
| 0.1 | 0.995004165278 |
| 0.2 | 0.980066577841 |
| 0.3 | 0.955336489126 |
| 0.4 | 0.921060994003 |
| 0.5 | 0.87758256189 |
| 0.6 | 0.82533561491 |
| 0.7 | 0.764842187284 |
| 0.8 | 0.696706709347 |
| 0.9 | 0.621609968271 |
| 1.0 | 0.540302305868 |
| 1.1 | 0.453596121426 |
| 1.2 | 0.362357754477 |
| 1.3 | 0.267498828625 |
| 1.4 | 0.1699671429 |
| 1.5 | 0.0707372016677 |
| 1.6 | -0.0291995223013 |
| 1.7 | -0.128844494296 |
| 1.8 | -0.227202094693 |
| 1.9 | -0.323289566864 |
| 2.0 | -0.416146836547 |
| 2.1 | -0.5048461046 |
| 2.2 | -0.588501117255 |
| 2.3 | -0.66627602128 |
| 2.4 | -0.737393715541 |
| 2.5 | -0.801143615547 |
| 2.6 | -0.856888753369 |
| 2.7 | -0.904072142017 |
| 2.8 | -0.942222340669 |
| 2.9 | -0.97095816515 |
| 3.0 | -0.9899924966 |
| 3.1 | -0.999135150273 |

Swapping the columns and re-ordering according to the new left column gives us a table for $y = \arccos(x)$:

| $x$ | $y = \arccos(x)$ |
|---|---|
| -0.999135150273 | 3.1 |
| -0.9899924966 | 3.0 |
| -0.97095816515 | 2.9 |
| -0.942222340669 | 2.8 |
| -0.904072142017 | 2.7 |
| -0.856888753369 | 2.6 |
| -0.801143615547 | 2.5 |
| -0.737393715541 | 2.4 |
| -0.66627602128 | 2.3 |
| -0.588501117255 | 2.2 |
| -0.5048461046 | 2.1 |
| -0.416146836547 | 2.0 |
| -0.323289566864 | 1.9 |
| -0.227202094693 | 1.8 |
| -0.128844494296 | 1.7 |
| -0.0291995223013 | 1.6 |
| 0.0707372016677 | 1.5 |
| 0.1699671429 | 1.4 |
| 0.267498828625 | 1.3 |
| 0.362357754477 | 1.2 |
| 0.453596121426 | 1.1 |
| 0.540302305868 | 1.0 |
| 0.621609968271 | 0.9 |
| 0.696706709347 | 0.8 |
| 0.764842187284 | 0.7 |
| 0.82533561491 | 0.6 |
| 0.87758256189 | 0.5 |
| 0.921060994003 | 0.4 |
| 0.955336489126 | 0.3 |
| 0.980066577841 | 0.2 |
| 0.995004165278 | 0.1 |
| 1.0 | 0.0 |

According to the principle

$$\begin{aligned}\arccos'(-0.904072142017) &\approx \frac{\arccos(-0.856888753369)-\arccos(-0.904072142017)}{(-0.856888753369)-(-0.904072142017)}\\ &= \frac{2.6-2.7}{(-0.856888753369)-(-0.904072142017)}\\ &= -2.1193899561988876\end{aligned}$$

we then get the following table for the approximate values of $\arccos'(x)$:

| $x$ | $\arccos'(x) \approx$ |
|---|---|
| -0.999135150273 | 10.9377 |
| -0.9899924966 | 5.2536 |
| -0.97095816515 | 3.4799 |
| -0.942222340669 | 2.6212 |
| -0.904072142017 | 2.1193 |
| -0.856888753369 | 1.7938 |
| -0.801143615547 | 1.5686 |
| -0.737393715541 | 1.4061 |
| -0.66627602128 | 1.2857 |
| -0.588501117255 | 1.1953 |
| -0.5048461046 | 1.1274 |
| -0.416146836547 | 1.0769 |
| -0.323289566864 | 1.0407 |
| -0.227202094693 | 1.0166 |
| -0.128844494296 | 1.0035 |
| -0.0291995223013 | 1.0006 |
| 0.0707372016677 | 1.0077 |
| 0.1699671429 | 1.0253 |
| 0.267498828625 | 1.0541 |
| 0.362357754477 | 1.0960 |
| 0.453596121426 | 1.1533 |
| 0.540302305868 | 1.2298 |
| 0.621609968271 | 1.3316 |
| 0.696706709347 | 1.4676 |
| 0.764842187284 | 1.6530 |
| 0.82533561491 | 1.9139 |
| 0.87758256189 | 2.2999 |
| 0.921060994003 | 2.9175 |
| 0.955336489126 | 4.0436 |
| 0.980066577841 | 6.6945 |
| 0.995004165278 | 20.01667 |

Now we should look closer at the first table of the previous page. Forget for a moment the arcus cosine, and look only at the cosine itself. We have $\cos(2.6) = -0.856888753369$ and $\cos(2.7) = -0.904072142017$. Appealing to (2.11) with $f = \cos$ we then have

$$\cos'(2.6) \approx \frac{\cos(2.7)-\cos(2.6)}{2.7-2.6} = \frac{(-0.904072142017)-(-0.856888753369)}{2.7-2.6}.$$

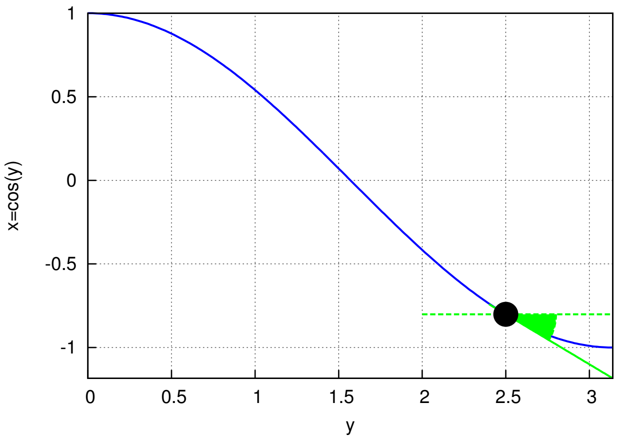
Now look at the top of the current page. You will be thoroughly amazed.

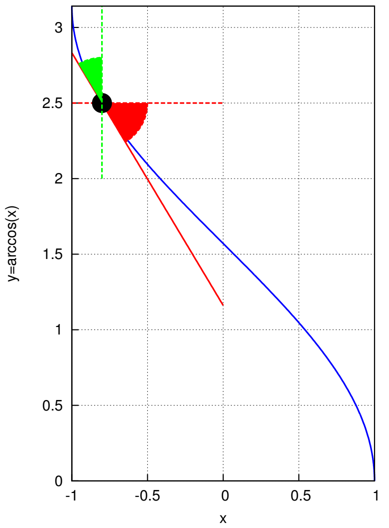
The conjecture is:

$$\text{If}\quad 0 < y_* < \pi \quad\text{and}\quad x_* = \cos(y_*) \quad\text{then}\quad \arccos'(x_*) = \frac{1}{\cos'(y_*)}. \tag{2.14}$$

**Looking at the arcus cosine as a graph:**

Here are the graphs of cos and arccos, with two tangent lines drawn at corresponding positions.





Recall that the graph of the inverse function is obtained by turning over the paper and then rotating it suitably. We see: the slope angles of the red and the green tangent lines add up to $-\pi/2$ (if both slope angles are negative, as in these two pictures), or they add up to $+\pi/2$ (if they are both positive, assuming both functions are increasing). Now

$$\text{slope of a straight line} = \tan(\text{its slope angle}),$$

and we have

$$\begin{aligned} \tan\left(\frac{\pi}{2} - \alpha\right) &= \frac{1}{\tan(\alpha)} \quad : \quad \text{if} \quad 0 < \alpha < \frac{\pi}{2}, \\ \tan\left(-\frac{\pi}{2} - \alpha\right) &= \frac{1}{\tan(\alpha)} \quad : \quad \text{if} \quad -\frac{\pi}{2} < \alpha < 0. \end{aligned}$$

This proves our conjecture (2.14).

**Looking at the arcus cosine as a formula:**

We have $0 < y_* < \pi$ and $x_* = \cos(y_*)$. We ask for another formula for $\arccos'(x_*)$, because (2.14) has the disadvantage of bringing $y_*$ into the discussion.

Let us start with

$$\arccos'(x_*) = \frac{1}{\cos'(y_*)} = \frac{1}{-\sin(y_*)}.$$

Now what do we know about this $y_*$ which we want to get rid of ? It holds $0 < y_* < \pi$, and therefore $\sin(y_*) > 0$, and this allows to conclude that

$$\sin(y_*) = \sqrt{1 - \cos^2(y_*)} = \sqrt{1 - x_*^2},$$

hence bringing us to the final result

$$\arccos'(x_*) = \frac{1}{-\sqrt{1 - x_*^2}}, \qquad \text{if} \quad -1 < x_* < 1.$$

**DIY:** *You know the derivatives of the functions*

$$x \mapsto \sin(x), \quad x \mapsto \tan(x), \quad x \mapsto \exp(x), \quad x \mapsto x^2.$$

*Figure out the derivatives of their inverse functions.*

*Here the notation $x \mapsto f(x)$ shall serve as a gentle reminder that a function $f$ is basically a machine where you put a number (called $x$) in and get another number (called $f(x)$) out.*

### The Product Rule

We know already the derivatives of the functions $x \mapsto x^2$ and $x \mapsto \sin(x)$, and these are the functions $x \mapsto 2x$ and $x \mapsto \cos(x)$.

What is then the derivative of their product $x \mapsto x^2 \sin(x)$ ?

In order to have an alternative notation, we give names:

$$u(x) := x^2, \qquad v(x) := \sin(x), \qquad f(x) := u(x)v(x) = x^2 \sin(x).$$

We choose some special argument, namely $x_* := 1.2$. Now the question becomes: what is $f'(x_*) = f'(1.2)$ ?

Appealing to (2.12), we have

$$f'(x_*) \approx \frac{f(x_* + 0.01) - f(x_*)}{0.01} = \frac{1.21^2 \sin(1.21) - 1.2^2 \sin(1.2)}{0.01},$$

and we not only wish to calculate the right-hand side (which would be easy), but we also aim for a deeper understanding.

We see two products here: $1.21^2 \sin(1.21)$ and $1.2^2 \sin(1.2)$. Both products have two factors each, and both factors "move": when we go from $x_*$ to $x_* + 0.01$, then $1.2^2$ changes into $1.21^2$, and $\sin(1.2)$ changes into $\sin(1.21)$.

It is easier to understand what is going on here if we do one change at a time, because the fraction then is less confusing. For this purpose we insert a fertile zero:

$$\begin{aligned} f'(x_*) &\approx \frac{1.21^2 \sin(1.21) - 1.2^2 \sin(1.21) + 1.2^2 \sin(1.21) - 1.2^2 \sin(1.2)}{0.01} \\ &= \underbrace{\frac{1.21^2 - 1.2^2}{0.01}}_{\text{item1}} \cdot \underbrace{\sin(1.21)}_{\text{item2}} + 1.2^2 \cdot \underbrace{\frac{\sin(1.21) - \sin(1.2)}{0.01}}_{\text{item3}} . \end{aligned}$$

Now we recall how to approximate the values of derivatives:

**Concerning *item 1*:** we have

$$\frac{1.21^2 - 1.2^2}{0.01} = \frac{u(1.21) - u(1.2)}{0.01} \approx u'(1.2) = 2 \cdot 1.2,$$

because $u(x) = x^2$ possesses the derivative $u'(x) = 2x$.

**Concerning *item 2*:** recalling (2.13), we can write

$$\begin{aligned} \sin(1.21) = v(1.21) &= v(1.2 + 0.01) \approx v(1.2) + v'(1.2) \cdot 0.01 \\ &= \sin(1.2) + \cos(1.2) \cdot 0.01. \end{aligned}$$

**Concerning *item 3*:** we recall once more (2.12) and obtain whence

$$\frac{\sin(1.21) - \sin(1.2)}{0.01} = \frac{v(1.21) - v(1.2)}{0.01} \approx v'(1.2) = \cos(1.2).$$

Having taken care of all three items, we now substitute them and find

$$\begin{aligned} f'(x_*) &\approx 2 \cdot 1.2 \cdot \Big( \sin(1.2) + \underbrace{\cos(1.2) \cdot 0.01}_{\text{to be neglected soon}} \Big) + 1.2^2 \cdot \cos(1.2) \\ &\approx 2x_* \sin(x_*) + x_*^2 \cos(x_*). \end{aligned}$$

We repeat all this, but now on a more general level. Forget about the special number 1.2 and take $x_*$

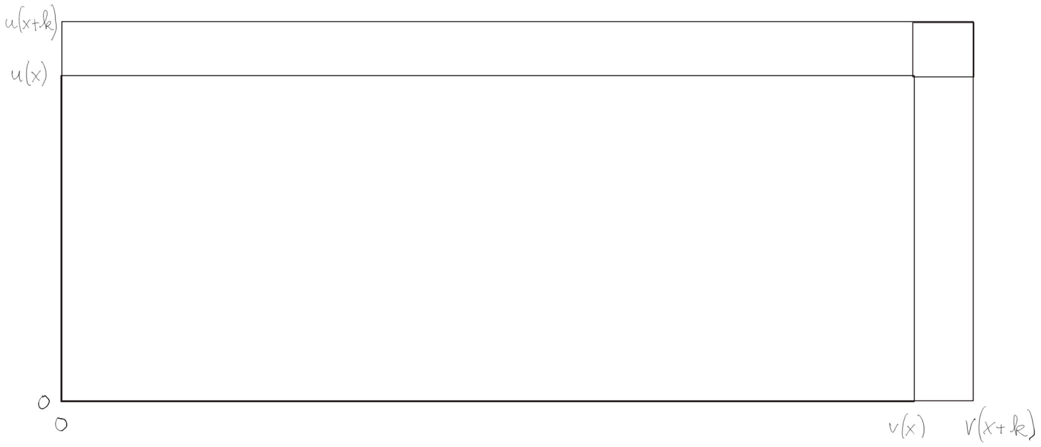
instead. Choose some small number $k$, such as $k = 0.01$. Then we have

$$\begin{aligned}
f'(x_*) &\approx \frac{f(x_*+k)-f(x_*)}{k} = \frac{u(x_*+k)v(x_*+k)-u(x_*)v(x_*)}{k} \\
&= \frac{u(x_*+k)v(x_*+k) - u(x_*)v(x_*+k) + u(x_*)v(x_*+k) - u(x_*)v(x_*)}{k} \\
&= \frac{u(x_*+k)-u(x_*)}{k}\cdot v(x_*+k) + u(x_*)\cdot\frac{v(x_*+k)-v(x_*)}{k} \\
&\approx u'(x_*)\cdot v(x_*+k) + u(x_*)\cdot v'(x_*) \\
&\approx u'(x_*)\cdot\Big(v(x_*)+v'(x_*)\cdot k\Big) + u(x_*)\cdot v'(x_*) \\
&= u'(x_*)\cdot v(x_*) + u(x_*)\cdot v'(x_*) + \underbrace{u'(x_*)\cdot v'(x_*)\cdot k}_{\text{vanishes if } k\to 0}.
\end{aligned}$$

And if we now change the notation to a generic $x$ instead of the special $x_*$:

$$\boxed{\Big(u(x)v(x)\Big)' = u'(x)\cdot v(x) + u(x)\cdot v'(x)}$$

To explain it once more, we construe products such as $u(x)\cdot v(x)$ as formula for the area of a rectangle. Have a look at the picture.



Then we get

$$\begin{aligned}
f(x+k)-f(x) &= \underbrace{u(x+k)\cdot v(x+k)}_{\text{all 4 rectangles}} - \underbrace{u(x)\cdot v(x)}_{\text{big lower left rectangle}} \\
&= \Big(\text{long narrow top left rectangle}\Big) + \Big(\text{long narrow lower right rectangle}\Big) \\
&\quad + \big(\text{tiny top right rectangle}\big) \\
&= \Big(u(x+k)-u(x)\Big)\cdot v(x) + u(x)\cdot\Big(v(x+k)-v(x)\Big) \\
&\quad + \Big(u(x+k)-u(x)\Big)\cdot\Big(v(x+k)-v(x)\Big).
\end{aligned}$$

Now we divide by $k$, and then we perform the limit $k \to 0$:

$$\underbrace{\frac{f(x+k)-f(x)}{k}}_{\text{if}\,\big\downarrow\,k\to 0} = \underbrace{\frac{u(x+k)-u(x)}{k}\cdot v(x)}_{\text{if}\,\big\downarrow\,k\to 0} + \underbrace{u(x)\cdot\frac{v(x+k)-v(x)}{k}}_{\text{if}\,\big\downarrow\,k\to 0} + \underbrace{\frac{u(x+k)-u(x)}{k}\cdot\Big(v(x+k)-v(x)\Big)}_{\text{if}\,\big\downarrow\,k\to 0}$$

$$f'(x) = \overbrace{u'(x)\cdot v(x)} + \overbrace{u(x)\cdot v'(x)} + \overbrace{u'(x)\cdot 0}$$

However, pictures can be misleading. If for instance $u(x) > 0$ and $v(x+k) < v(x)$, then the above picture is wrong, we need another picture, and now some rectangles have to be counted negative. But the calculation with the limits is correct in all combinations of the signs, and the picture is just a visualisation.

### The Chain Rule

We have the two functions $x \mapsto u(x) = x^2$ and $x \mapsto v(x) = \sin(x)$. We could add them, or multiply them, or put one function into the other, in which case we get a new function

$$w(x) := v\Big(u(x)\Big) = \sin\Big(x^2\Big).$$

We call this the *composition* of the two functions $v$ and $u$, and for obvious reasons $v$ is called *outer* function, and $u$ is called *inner* function. Instead of $v(u(x))$, we also may write $(v \circ u)(x)$, and the $\circ$ could be read as "after", because you apply the machine $v$ after the machine $u$.

The order matters: $v \circ u$ is not the same as $u \circ v$. An everyday analogy might be helpful: if $u$ is the operation "put on the socks" and $v$ is the operation "put on the shoes", then $v \circ u$ corresponds to standard behaviour of a person, but $u \circ v$ would be awkward.

Going back to the above functions $u(x) = x^2$ and $v(x) = \sin(x)$, what is now $w'(x)$ ? Pick some special value $x_*$. Then we have, for $x$ near $x_*$, the well-known approximation

$$w'(x_*) \approx \frac{w(x) - w(x_*)}{x - x_*} = \frac{v\Big(u(x)\Big) - v\Big(u(x_*)\Big)}{x - x_*}.$$

Now we need a clever idea (as it often happens in mathematics): we expand the fraction, introducing a fertile factor one:

$$w'(x_*) \approx \frac{v(u(x)) - v(u(x_*))}{u(x) - u(x_*)} \cdot \frac{u(x) - u(x_*)}{x - x_*}.$$

Perhaps the reasoning becomes clearer if we introduce two new names:

$$y := u(x), \qquad y_* := u(x_*).$$

We substitute these two names in the first fraction:

$$w'(x_*) \approx \frac{v(y) - v(y_*)}{y - y_*} \cdot \frac{u(x) - u(x_*)}{x - x_*}.$$

Our assumption is that $x$ is near to $x_*$. But then also $u(x)$ is near $u(x_*)$, because (2.13) gives us the approximation $u(x) \approx u(x_*) + u'(x_*) \cdot (x - x_*)$, and the second item in this sum is negligible when compared to the first item $u(x_*)$.

But then also $y$ is near $y_*$, because these are only new names for $u(x)$ and $u(x_*)$. Now we appeal to (2.11) twice (once for $v$, and once for $u$):

$$v'(y_*) \approx \frac{v(y) - v(y_*)}{y - y_*}, \qquad u'(x_*) \approx \frac{u(x) - u(x_*)}{x - x_*}.$$

We substitute these approximations into our formula above and obtain

$$w'(x_*) \approx v'(y_*) \cdot u'(x_*).$$

In the courses on *Calculus* and *Analysis* you will learn that this approximation is in fact an identity, which we re-write as

$$\boxed{\Big(v \circ u\Big)'(x_*) = v'\Big(u(x_*)\Big) \cdot u'(x_*)}$$

Be aware that the expression $v'(u(x))$ means "figuring out what the derivative $v'$ is, and then substituting the value $u(x_*)$ into $v'$". In our case, we have

$$v'(y_*) = \sin'(y_*) = \cos(y_*) = \cos(x_*^2), \qquad u'(x_*) = 2x_*,$$

and therefore (renaming $x_*$ to $x$)

$$\Big(\sin(x^2)\Big)' = \cos(x^2) \cdot 2x.$$

**DIY:** *Now practise all the rules for derivatives:*

- *Prove that the derivative of the function $x \mapsto x$ is the function $x \mapsto 1$. Using (2.11) should make this part trivial*[11].
- *Use the product rule and the previous result to prove rigorously that the functions $x \mapsto x^2$ and $x \mapsto x^3$ have the derivatives $x \mapsto 2x$ and $x \mapsto 3x^2$. Then give a reason why the function $x \mapsto x^n$ has the derivative $x \mapsto nx^{n-1}$ for $n \in \mathbb{N}$.*
- *Build on the previous result, and use the rule for derivatives of inverse functions to determine the derivative of the function $x \mapsto x^{1/n}$ for $n \in \mathbb{N}_+$ and $x > 0$.*
- *Building on the previous result and on the chain rule, determine (with proof) the derivative of the function $x \mapsto x^\alpha$ if $\alpha = \frac{p}{q}$ is a positive rational number and $x$ is positive.*
- *Building on the previous result and the product rule, determine (with proof) the derivative of the function $x \mapsto x^\alpha$ for a* negative *rational exponent $\alpha$, and positive $x$, using $x^\alpha \cdot x^{-\alpha} = 1$.*

### 2.3.3 Integrals

#### The Meaning of an Integral

Recall what a derivative *means:* the derivative $f'(x)$ of a function $f(x)$ is a tool that tells us how quick this function $f$ changes near that chosen point $x$. As an example: $f'(2)$ tells you how quick the function $f$ changes near the point $x = 2$. And $f'(2)$ tells you absolutely nothing how the function $f$ behaves near the point $x = 47$.

When you ask for the value of $f'(2)$, you do basically the following: you take the function $f$, you ignore everything which happens outside some small interval (such as $[1.999, 2.001]$) around the point $x = 2$, you pretend that inside this interval the graph of the function is approximately a straight line, of which you then determine the slope. By definition, $f'(2)$ is then the value of that slope. It is the *rate of change* of the function $f$, locally near that selected point.

So, in some sense, taking the derivative means to **zoom in**, because you choose to ignore all the behaviour of the function that is "far away" from that chosen point $x = 2$.

Constructing an integral means the opposite: you **zoom out**. You suppose that at every point $x$, you know the *rate of change* (which is a local information), and from that you then rebuild the *global change.*

Let us take an example. I like driving really fast cars.

| At the time | $t = 0$ | $t = 1$ | $t = 2$ | $t = 3$ | $t = 4$ | $t = 5$ | seconds, |
|---|---|---|---|---|---|---|---|
| my car has the velocity | 0 | 3 | 8 | 14 | 20 | 25 | metres per second. |

How far has the car gone during these five seconds ? In other words, we ask for the *global change* of the position of the car over the duration of five seconds. We assume that the car accelerates, which means that the velocity always goes up. In particular, over the duration of the third second, the velocity is never below $8\frac{m}{s}$, and never above $14\frac{m}{s}$. The velocity is the *rate of change*, which is a piece of local information. Now we will take these many pieces of local information and rebuild from them the global change.

We start with an estimate from below:

**During the first second:** the velocity has been at least $0\frac{m}{s}$, and therefore the position of the car has changed by at least 0 metres, over the duration of the first second.

**During the second second:** the velocity has been at least $3\frac{m}{s}$, and therefore the position of the car has changed by at least 3 metres, over the duration of the second second.

**During the third second:** the velocity has been at least $8\frac{m}{s}$, and therefore the position of the car has changed by at least 8 metres, over the duration of the third second.

[11] *trivial* comes from *tres* meaning *three* and *via* meaning *road.* Therefore *trivialis* is a Latin adjective with the meaning *that which pertains to the intersection of three roads.* In ancient times, the seven liberal arts were divided into the *trivium* grammar, logic, rhetoric (to be studied first) and the *quadrivium* music, arithmetic, geometry, astronomy (to be studied afterwards). Philosophy and theology were to be studied after that because these two disciplines are even harder.

**During the fourth second:** the velocity has been at least $14\frac{m}{s}$, and therefore the position of the car has changed by at least 14 metres, over the duration of the fourth second.

**During the fifth second:** the velocity has been at least $20\frac{m}{s}$, and therefore the position of the car has changed by at least 20 metres, over the duration of the fifth second.

In total, the car has travelled at least $0 + 3 + 8 + 14 + 20 = 45$ metres over the duration of these five seconds. Hence the *global change* of the position of the car is at least 45 metres.

And we do an estimate from above:

**During the first second:** the velocity has been at most $3\frac{m}{s}$, and therefore the position of the car has changed by at most 3 metres, over the duration of the first second.

**During the second second:** the velocity has been at most $8\frac{m}{s}$, and therefore the position of the car has changed by at most 8 metres, over the duration of the second second.

**During the third second:** the velocity has been at most $14\frac{m}{s}$, and therefore the position of the car has changed by at most 14 metres, over the duration of the third second.

**During the fourth second:** the velocity has been at most $20\frac{m}{s}$, and therefore the position of the car has changed by at most 20 metres, over the duration of the fourth second.

**During the fifth second:** the velocity has been at most $25\frac{m}{s}$, and therefore the position of the car has changed by at most 25 metres, over the duration of the fifth second.

In total, the car has travelled at most $3 + 8 + 14 + 20 + 25 = 70$ metres over the duration of these five seconds. Hence the *global change* of the position of the car is at most 70 metres.

Let us visualise the two estimates.

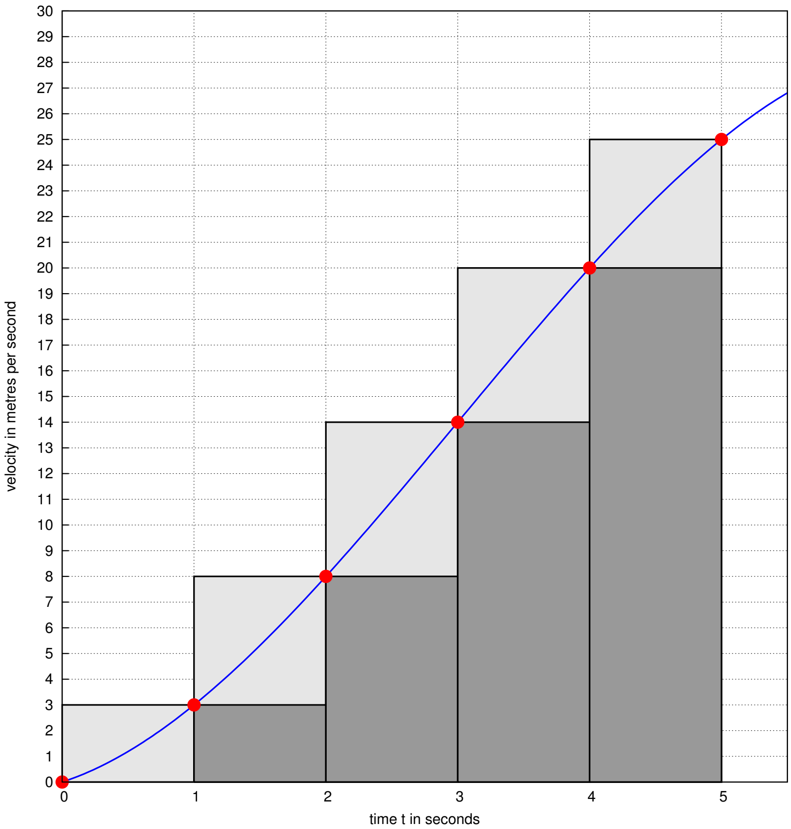
The data points from the table are the red dots in the diagram. The connecting blue line shall be the velocity of the car as a function of time $t$, and this blue line has been made up by me.

The dark grey rectangles together have the area $3+8+14+20 = 45$, and they visualise the estimate from below.

The dark grey rectangles plus the light grey visualise the estimate from above, and their joint area is $3+8+14+20+25 = 70$.

The total distance that has been travelled by the car during the five seconds would be the area below the blue curve.

Observe that the units fit nicely: seconds (width of the rectangle) multiplied by metres per second (height of rectangle) actually yield metres, as they should do.



The gap between the upper estimate 70 and the lower estimate 45 is 25, and this is the joint area of the light grey rectangles. We could shift all light grey rectangles horizontally to the left axis, and then we get a column of width 1 and height 25 that comprises all these light grey rectangles stacked on top of each other.

If we are of the opinion that this gap 25 is too big, what can we do ? We could measure the current velocity more often (which means to have more red dots in the diagram), and then repeat the previous calculation. Now the column of stacked light grey rectangles would become narrower, resulting in better estimates.

**DIY:** *Give a geometric reason why the lower estimate 45 becomes bigger if the time step size is being reduced from 1 second to* 0.5 *second. And similarly explain geometrically why the upper estimate 70 becomes lower when the time step size is being halved.*

Now let us hoist these considerations onto a more general level, which will involve formulas. The blue line shall be the graph of a velocity function $v = v(t)$, which we assume to be known. The interval $[0, 5]$ shall be replaced by an interval $[a, b]$ with $a < b$. We ask for the area of the region that is bounded from above by the blue line, bounded from below by the interval $[a, b]$, bounded from the left by the line $t = a$, and from the right by the line $t = b$. We assume that always $v(t) \geq 0$, and the function $v(t)$ is going up (which means that the car shall get faster and faster).

We choose a large natural number $N$, and we split the interval $[a, b]$ into $N$ pieces of equal length. This brings us intermediate points

$$a + \frac{b-a}{N}, \quad a + 2 \cdot \frac{b-a}{N}, \quad a + 3 \cdot \frac{b-a}{N}, \quad a + 4 \cdot \frac{b-a}{N}, \ \ldots, \quad a + (N-1) \cdot \frac{b-a}{N}.$$

The estimates from below and above involve various rectangles which have all equal width $\frac{b-a}{N}$. The area of such a rectangle is computed as height times width, and consequently the lower estimate is

$$v(a) \cdot \frac{b-a}{N} + v\left(a + \frac{b-a}{N}\right) \cdot \frac{b-a}{N} + v\left(a + 2 \cdot \frac{b-a}{N}\right) \cdot \frac{b-a}{N} + \cdots + v\left(a + (N-1) \cdot \frac{b-a}{N}\right) \cdot \frac{b-a}{N}.$$

I have trouble getting this written on one line, and exactly for such purposes the summation symbol $\Sigma$ has been invented that abbreviates this long sum as

$$\sum_{j=0}^{N-1} v\left(a + j \cdot \frac{b-a}{N}\right) \cdot \frac{b-a}{N}.$$

Let us give it a name:

$$D_{N,\text{lower}} := \sum_{j=0}^{N-1} v\left(a + j \cdot \frac{b-a}{N}\right) \cdot \frac{b-a}{N}.$$

Here $D$ stands for "distance travelled", $N$ remembers us that we have split the interval $[a, b]$ into $N$ pieces, and the subscript "lower" needs no explanation.

Similarly we have the upper estimate

$$D_{N,\text{upper}} := \sum_{j=1}^{N} v\left(a + j \cdot \frac{b-a}{N}\right) \cdot \frac{b-a}{N}.$$

The only differences are the terminal values for the running counter $j$.

Here it has been very helpful that the function $v(t)$ is assumed as increasing, because then $v$ attains its smallest value over a narrow sub-interval at its left end-point, and its biggest value at the right end-point of that sub-interval.

Because $D_{N,\text{lower}}$ and $D_{N,\text{upper}}$ are estimates of the desired area from below and from above, we have the inequalities

$$D_{N,\text{lower}} \leq (\text{desired area}) \leq D_{N,\text{upper}}.$$

The gap between both estimates is

$$D_{N,\text{upper}} - D_{N,\text{lower}} = v\left(a + N \cdot \frac{b-a}{N}\right) \cdot \frac{b-a}{N} - v\left(a + 0 \cdot \frac{b-a}{N}\right) \cdot \frac{b-a}{N} = \Big(v(b) - v(a)\Big) \cdot \frac{b-a}{N},$$

which is the light grey column once again.

We change the notation slightly: the interval $[a,b]$ is being understood as a time interval, and we have split it into pieces of duration $\frac{b-a}{N}$. Let us call this by a new name,

$$\Delta t := \frac{b-a}{N},$$

and the meaning of $\Delta t$ is "tiny change of the variable $t$". We re-write our approximations:

$$D_{N,\text{lower}} = \sum_{j=0}^{N-1} v(a+j\cdot\Delta t)\cdot\Delta t, \qquad D_{N,\text{upper}} = \sum_{j=1}^{N} v(a+j\cdot\Delta t)\cdot\Delta t.$$

Now let $N$ become bigger and bigger. The variable $t$ runs through the interval $[a,b]$ from left to right, and both estimates get ever closer. The letter $\Sigma$ is the Greek $S$, which may be written as $\int$ (which is the convention introduced by Leibniz[12]), and then we have the value $D$ of the area under the blue line:

$$D = \int_{t=a}^{b} v(t)\,\mathrm{d}t.$$

Here the Greek letter $\Delta$ (delta) has become a " d". The mathematical definition of this expression is

$$\int_{t=a}^{b} v(t)\,\mathrm{d}t := \lim_{N\to\infty}\left(\sum_{j=0}^{N-1} v(a+j\cdot\Delta t)\cdot\Delta t\right), \qquad \text{where} \quad \Delta t := \frac{b-a}{N}.$$

Today is not the right moment to explain what a limit actually *is*. And it is also not very clear whether this limit even *exists*. Just because some mathematician writes something with ink on paper does not mean that this "something" exists at all. The existence of this "something" has to be proved (however, not today). But the good news is that everything is fine if the function $v(t)$ is continuous on the interval $[a,b]$ (being continuous means roughly that the graph of $v$ is bounded from above and from below, and that the graph does not jump). The existence of the limit does not require the function $v$ to be increasing everywhere; continuity is enough.

If the function $v(t)$ is continuous on the interval $[a,b]$, then the integral $\int_{t=a}^{b} v(t)\,\mathrm{d}t$ exists, and it can be approximated *via*

$$\int_{t=a}^{b} v(t)\,\mathrm{d}t \approx \left(\sum_{j=0}^{N-1} v(a+j\cdot\Delta t)\right)\cdot\Delta t, \qquad \text{where} \quad \Delta t := \frac{b-a}{N},$$

and $N$ is a huge positive integer, and the error of this approximation can be made arbitrarily small by choosing $N$ sufficiently big.

Note that $v$ may take negative values (which means that the blue line is sometimes below the $t$-axis), and then the relevant part between the blue line and the $t$-axis is counted negative.

If we are interested in a rough upper estimate of the absolute value of the integral, here is one:

$$\left|\int_{t=a}^{b} v(t)\,\mathrm{d}t\right| \le \left(\max_{t\in[a,b]} |v(t)|\right)\cdot(b-a). \tag{2.15}$$

The geometric interpretation of the product on the RHS[13] is: we have put the area under consideration into a rectangular box of width $b-a$ and of a height which is the maximal value of $|v(t)|$ when $t$ runs through the interval $[a,b]$.

Now let us be given three numbers $a$, $b$, $c$ with $a<b<c$. What can be said about the three integrals $\int_{t=a}^{b} v(t)\,\mathrm{d}t$, $\int_{t=b}^{c} v(t)\,\mathrm{d}t$, and $\int_{t=a}^{c} v(t)\,\mathrm{d}t$ ? To answer this, it helps to get back at the physical interpretation. The variable $t$ is to be understood as running time, and $v(t)$ is the velocity of a car on a straight

[12] Gottfried Wilhelm Leibniz without t, 1646–1716

[13] right-hand side

road (no curves in the road). Then $\int_{t=a}^{b} v(t)\,\mathrm{d}t$ calculates the distance that this car has travelled during the time interval $[a, b]$. Similarly for the other two integrals, and we are quickly convinced that

$$\left(\int_{t=a}^{b} v(t)\,\mathrm{d}t\right) + \left(\int_{t=b}^{c} v(t)\,\mathrm{d}t\right) = \int_{t=a}^{c} v(t)\,\mathrm{d}t. \tag{2.16}$$

In the sequel, we will drop the big parentheses on the LHS.

**DIY:** *Give a geometrical interpretation of this formula.*

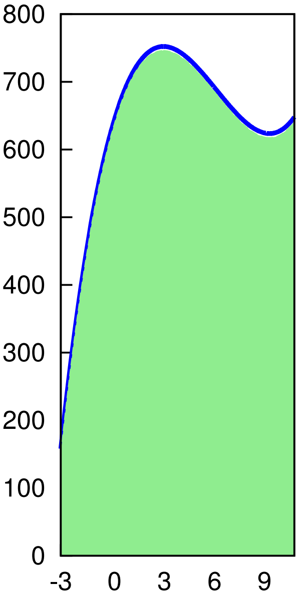
Let us make our life a bit easier — we wish to get rid of the precondition that the upper terminal at the integral symbol shall be a bigger number than the lower terminal:

**Definition 2.1.** *If $v(t)$ is a continuous function and $b < a$, then we define*

$$\int_{t=a}^{b} v(t)\,\mathrm{d}t := -\int_{t=b}^{a} v(t)\,\mathrm{d}t.$$

*We also define $\int_{t=a}^{a} v(t)\,\mathrm{d}t := 0$.*

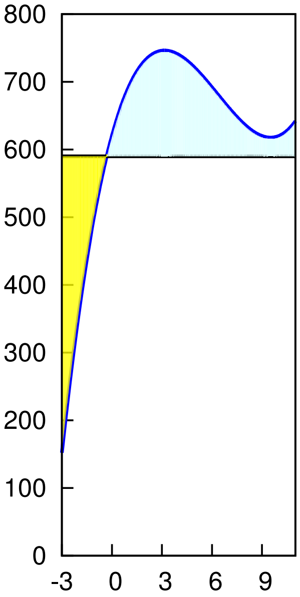
The key advantage of this definition is now that (2.16) becomes true irrespective of the order of the numbers $a$, $b$, $c$. We need no longer assume that $a < b < c$.



The blue line in the left picture is the graph of a function $v(t)$ over the interval $[a, b] = [-3, 11]$. The green area equals $\int_{t=-3}^{11} v(t)\,\mathrm{d}t$.

Now we ask whether we can find a rectangle with the same area as the green region, and with the same width $14 = 11 - (-3)$.

We can see such a rectangle in the right picture. The horizontal black line has to be chosen at such a height that the yellow region has the same area as the light blue region. Note that the black line intersects the blue line at a point $\approx \left[\begin{smallmatrix}-0.3\\ v(-0.3)\end{smallmatrix}\right]$. Call $\tau$ the $t$-value of that intersection point. Then the area of the rectangle is $v(\tau)\cdot 14$ (height times width).



Hence we have convinced ourselves of the following result:

**Proposition 2.2.** *Let $v(t)$ be a continuous function on the interval that is between the numbers $a$ and $b$. Then there is an intermediate point of time $\tau$ between $a$ and $b$ for which the following identity becomes true (irrespective of $a < b$ or $b < a$):*

$$\int_{t=a}^{b} v(t)\,\mathrm{d}t = v(\tau)\cdot(b-a).$$

### The Fundamental Theorem of Calculus

Let us be given a continuous function $f$ on the interval $[0, 10]$. If we calculate the integral $\int_{t=0}^{10} f(t)\,\mathrm{d}t$, then this equals (by its very definition) the area between the graph of $f$ and the interval $[0, 10]$. But we could integrate over a shorter interval: choose some $x$ between 0 and 10, and consider the integral $\int_{t=0}^{x} f(t)\,\mathrm{d}t$. We could keep the upper terminal $x$ flexible — it need not be fixed. Let us christen the integral:

$$F(x) := \int_{t=0}^{x} f(t)\,\mathrm{d}t, \qquad \text{where} \quad 0 < x < 10.$$

This $F$ is now a function of $x$. How does it change if $x$ changes ? The answer is in the following theorem.

**Theorem 2.3** (**Fundamental Theorem of Calculus, First Version**). *Let $f(t)$ be a given continuous function on the interval $[a,b]$, and define $F(x) := \int_{t=a}^{x} f(t)\,\mathrm{d}t$, for $a \le x \le b$.*

*Then $F'(x) = f(x)$, for every $x \in [a,b]$.*

*Proof.* Pick some $x_* \in [a,b]$ and keep it fixed. We wish to show $F'(x_*) = f(x_*)$. Appealing to (2.12), we need to investigate the fraction

$$\frac{F(x_*+k) - F(x_*)}{k}, \qquad \text{where} \quad |k| \quad \text{is tiny.}$$

From (2.16) and the very definition of $F$, we have

$$\frac{F(x_*+k) - F(x_*)}{k} = \frac{1}{k}\int_{t=x_*}^{x_*+k} f(t)\,\mathrm{d}t.$$

Now we apply Proposition 2.2 to the RHS and deduce that

$$\frac{F(x_*+k) - F(x_*)}{k} = f(\tau) \qquad \text{for some} \quad \tau \quad \text{between} \quad x_* \quad \text{and} \quad x_* + k.$$

Now we perform the limit $k \to 0$. Then $\tau$ must approach $x_*$, because $\tau$ is squeezed in between $x_*$ and $x_* + k$. This completes the pseudo-proof. □

The first version of the fundamental theorem says: if we have a function $f$ of which we **at first** calculate the integral $\int_{t=a}^{x} f(t)\,\mathrm{d}t$ with *movable* upper terminal $x$, and we then **subsequently** take the derivative, then we get the value $f(x)$ back. In that sense, the two activities “calculating an integral with movable upper terminal” and “calculation a derivative” are inverse to each other, when they are performed in the mentioned order.

The second version of the fundamental theorem takes care of the reverse order.

**Theorem 2.4** (**Fundamental Theorem of Calculus, Second Version**). *Let $g(t)$ be a continuous function on some interval $[a,b]$, and let $G(t)$ be an* arbitrary *function with the property $G'(t) = g(t)$ for every $t \in [a,b]$. Then*

$$\int_{t=a}^{b} g(t)\,\mathrm{d}t = G(b) - G(a).$$

We introduce the abbreviation $G(t)\Big|_{t=a}^{t=b} := G(b) - G(a)$.

As an example, take $[a,b] = [0,\pi]$ and $g(t) = \sin(t)$. Now we need another function $G(t)$ with the property that $G'(t) = \sin(t)$ for all $t$. One such function is $G(t) = 37 - \cos(t)$. Then we obtain

$$\int_{t=0}^{\pi} \sin(t)\,\mathrm{d}t = \Big(37 - \cos(\pi)\Big) - \Big(37 - \cos(0)\Big) = -\cos(\pi) + \cos(0) = -(-1) + 1 = 2.$$

Instead of 37, you could have taken any other fixed number. The final result $\int_{t=0}^{\pi} \sin(t)\,\mathrm{d}t = 2$ does not depend on that number.

*Pseudo–Proof of Theorem 2.4.* We have the functions $g$ and $G$, and now we define one more function $F$:

$$F(x) := \int_{t=a}^{x} g(t)\,\mathrm{d}t, \quad a \le x \le b.$$

Then we have $\int_{t=a}^{b} g(t)\,\mathrm{d}t = F(b)$, and we only need to show that $F(b) = G(b) - G(a)$.

Now the First Version tells us $F'(x) = g(x)$ for all $x \in [a,b]$, an our assumption has been that $G'(x) = g(x)$ for all $x \in [a,b]$. Let us subtract these two equations:

$$\Big(F - G\Big)'(x) = 0, \qquad \forall x \in [a,b].$$

This means that $x \mapsto (F(x) - G(x))$ is a function which has everywhere the derivative equal to zero, and therefore this function is constant, which means that the function $F(x) - G(x)$ takes everywhere the same value. In particular, the function $F - G$ has the same value at the left endpoint $a$ of the interval as at the right endpoint $b$:

$$F(a) - G(a) = F(b) - G(b).$$

We can re-arrange this into the identity

$$G(b) - G(a) = F(b) - F(a).$$

Now what is $F(a)$ ? By definition of $F$, we have $F(a) = \int_{t=a}^{a} g(t)\,\mathrm{d}t$, and therefore $F(a) = 0$. Hence we have shown $G(b) - G(a) = F(b)$, which concludes the pseudo–proof (a rigorous proof would have to explain more in detail why $F(x) - G(x)$ is constant). □

The second version can obviously be written as

$$\int_{t=a}^{b} G'(t)\,\mathrm{d}t = G(t)\Big|_{t=a}^{t=b}, \tag{2.17}$$

and another perspective will perhaps bring more clarity. We attempt to pseudo-prove (2.17) once again. We start from its LHS, we recall the definition of an integral and its approximation by a finite sum. To this end, we choose some large integer $N$, and we split the interval $[a, b]$ into $N$ pieces of equal length $\Delta t := \frac{b-a}{N}$. Then we have the approximation

$$\int_{t=a}^{b} G'(t)\,\mathrm{d}t \approx \left(\sum_{j=0}^{N-1} G'(a + j \cdot \Delta t)\right) \cdot \Delta t.$$

Now what is $G'(a+j\cdot\Delta t)$ ? It is the derivative of $G$, and this derivative is to be evaluated at $t = a+j\cdot\Delta t$. We know already how to approximate derivatives, by (2.12):

$$G'(a + j \cdot \Delta t) \approx \frac{G(a + j \cdot \Delta t + k) - G(a + j \cdot \Delta t)}{k}, \qquad \text{where} \quad k \quad \text{is tiny.}$$

We are allowed to select $k$, as long as it is reasonably small. So why not taking $k := \Delta t$ ? Let us do it:

$$G'(a + j \cdot \Delta t) \approx \frac{G(a + j \cdot \Delta t + \Delta t) - G(a + j \cdot \Delta t)}{\Delta t}.$$

We plug this into the above approximation of the integral:

$$\int_{t=a}^{b} G'(t)\,\mathrm{d}t \approx \left(\sum_{j=0}^{N-1} \frac{G(a + (j+1) \cdot \Delta t) - G(a + j \cdot \Delta t)}{\Delta t}\right) \cdot \Delta t.$$

The $\Delta t$ obviously cancel, whence:

$$\int_{t=a}^{b} G'(t)\,\mathrm{d}t \approx \sum_{j=0}^{N-1} \Big(G(a + (j+1) \cdot \Delta t) - G(a + j \cdot \Delta t)\Big).$$

We should perhaps expand the RHS:

$$\begin{aligned}
&\sum_{j=0}^{N-1} \Big(G(a + (j+1) \cdot \Delta t) - G(a + j \cdot \Delta t)\Big) \\
&= G(a + 1 \cdot \Delta t) - G(a + 0 \cdot \Delta t) && \Big| \quad \text{here} \quad j = 0, \\
&\quad + G(a + 2 \cdot \Delta t) - G(a + 1 \cdot \Delta t) && \Big| \quad \text{here} \quad j = 1, \\
&\quad + G(a + 3 \cdot \Delta t) - G(a + 2 \cdot \Delta t) && \Big| \quad \text{here} \quad j = 2, \\
&\quad \ldots
\end{aligned}$$

$$\begin{aligned}
&\quad + G(a+(N-1)\cdot\Delta t) - G(a+(N-2)\cdot\Delta t) && \Big| \text{ here } \quad j=N-2,\\
&\quad + G(a+N\cdot\Delta t) - G(a+(N-1)\cdot\Delta t) && \Big| \text{ here } \quad j=N-1,\\
&= G(a+N\cdot\Delta t) - G(a) && \Big| \text{ now apply definition of } \quad \Delta t\\
&= G(b) - G(a).
\end{aligned}$$

This makes us happy.

### The Rule of Partial Integration

The Fundamental Theorem of Calculus, applied to a function $w(x)$, says

$$w(b) - w(a) = w(x)\Big|_{x=a}^{x=b} = \int_{x=a}^{b} w'(x)\,\mathrm{d}x.$$

Now let us suppose that $w(x) = u(x)\cdot v(x)$, with two other functions $u(x)$ and $v(x)$:

$$u(x)\cdot v(x)\Big|_{x=a}^{x=b} = \int_{x=a}^{b} (uv)'(x)\,\mathrm{d}x.$$

But the product rule of differentiation says $(uv)' = u'v + uv'$, and therefore we can write

$$u(x)\cdot v(x)\Big|_{x=a}^{x=b} = \int_{x=a}^{b} u'(x)\cdot v(x)\,\mathrm{d}x + \int_{x=a}^{b} u(x)\cdot v'(x)\,\mathrm{d}x.$$

Re-arranging then gives the *rule of partial integration:*

$$\boxed{\int_{x=a}^{b} u'(x)\cdot v(x)\,\mathrm{d}x = u(x)\cdot v(x)\Big|_{x=a}^{x=b} - \int_{x=a}^{b} u(x)\cdot v'(x)\,\mathrm{d}x}$$

The idea behind the name *partial integration* is something like: in the integrand, we have a product $u'v$ to be integrated. First we integrate only the left factor (and keep the right factor), which results in the item $u(x)v(x)|_{x=a}^{x=b}$. But this is not the correct final result because we did not treat the right factor correctly, and we have to compensate for this mistake with another integral, which then involves the integrand $uv'$. When you compare both integrals, you will observe that the derivative has wandered from one factor to the other factor (and you have acquired also a minus sign). Hopefully the second integral is easier to evaluate than the first integral.

After having done the following exercise, you will understand why partial integration is useful:

**DIY:** *Calculate the following integrals using partial integration:*

$$\begin{aligned}
&\int_{x=2}^{4} x\exp(x)\,\mathrm{d}x \quad \textit{with} \quad u'(x)=\exp(x), \quad v(x)=x,\\
&\int_{x=1}^{3} x\sin(x)\,\mathrm{d}x \quad \textit{with} \quad u'(x)=\sin(x), \quad v(x)=x,\\
&\int_{x=17}^{18} \ln(x)\,\mathrm{d}x \quad \textit{with} \quad u'(x)=1, \quad v(x)=\ln(x).
\end{aligned}$$

### Colin Maclaurin and his Series

Colin Maclaurin[14] was a Scottish mathematician, and the Maths building of Heriot-Watt is named after him. You can find his grave at Greyfriars Kirk.

Maclaurin has become famous for his *Maclaurin Series* which we attempt to present now. Assume that some arithmetic operations are considered *easy*, and some others are considered *difficult.* The easy ones

[14] 1698–1746

are adding, subtracting, multiplying; however everything else (for example taking sine and cosine, taking logarithms, even dividing by an expression that contains $x$) is considered difficult.

Now the question is: can we approximate a difficult arithmetic operation by many easy ones ? Or in other words, how can we calculate $\ln(x)$ for $x = 1.2$ with only the following operations: adding, subtracting, multiplying terms with $x$, dividing by numbers, but no other operations ? Questions of this type are relevant for the real world because the designers of microprocessors in computers have exactly this problem: the electronic devices in the processors are typically transistors, and these transistors can only do two things: switch the electric current on, and switch the electric current off. Now how does your calculator do the logarithm ? Switching things on and off is a primitive way of counting something. The processor designers have to arrange many transistors in such a way that this whole circuit can do additions, out of counting operations. And then transistors have to be arranged in another way, in order to build multiplications out of additions. And from that, then all the other functions such as logarithm and sine have to be built, somehow. That is one reason why modern processors in your laptop have between one billion and five billion transistors.

The solution of our mathematical problem comes from the derivative of the logarithm: $\ln'(1+x) = \frac{1}{1+x}$. If we integrate this equation from $a = 0$ to $b = 0.2$, we get

$$\ln(1.2) - \ln(1) = \int_{x=0}^{x=0.2} \frac{1}{1+x}\,\mathrm{d}x.$$

But $\ln(1) = 0$, and let us write $b$ for 0.2:

$$\ln(1+b) = \int_{x=0}^{b} \frac{1}{1+x}\,\mathrm{d}x.$$

We wish to apply here the rule of partial integration, for which we need two factors in the integrand:

$$\frac{1}{1+x} = \underbrace{1}_{u'(x)} \cdot \underbrace{\frac{1}{1+x}}_{v(x)}.$$

Now $u'(x) = 1$, so what is $u(x)$ ? Having clever ideas is often helpful, and so we choose $u(x) = x - b$. We check that this function $u(x)$ indeed has the derivative $u'(x) = 1$. Now we perform partial integration, and we introduce the habit of writing all *difficult* operations in red colour, and all *easy* operations in black:

$$\begin{aligned}\ln(1+b) &= \int_{x=0}^{b} u'(x)\cdot v(x)\,\mathrm{d}x = u(x)\cdot v(x)\Big|_{x=0}^{x=b} - \int_{x=0}^{b} u(x)\cdot v'(x)\,\mathrm{d}x \\ &= (x-b)\cdot\frac{1}{1+x}\Big|_{x=0}^{x=b} + \int_{x=0}^{b}(x-b)\cdot\frac{1}{(1+x)^2}\,\mathrm{d}x \\ &= b + \int_{x=0}^{b}(x-b)\cdot\frac{1}{(1+x)^2}\,\mathrm{d}x.\end{aligned}$$

At first glance, this does not look very good. But the first item on the RHS is $b$ (which we can calculate trivially because $b$ is given anyway, namely $b = 0.2$). And the integral turns out to be quite small: it holds

$$\left|(x-b)\cdot\frac{1}{(1+x)^2}\right| \le b \qquad \text{if } 0 \le x \le b,$$

and therefore — owing to (2.15) — the value of the integral is at most $b \cdot b$, which is in our case 0.04. Therefore, we have determined that $\ln(1.2)$ is somewhere between 0.16 and 0.24.

The trick is now to do partial integration again, with new functions $u'$ and $v$ (not the same as above):

$$\underbrace{(x-b)}_{u'(x)}\cdot\underbrace{\frac{1}{(1+x)^2}}_{v(x)}, \qquad \text{hence} \quad u(x) = \frac{1}{2}(x-b)^2, \qquad v'(x) = \frac{-2}{(1+x)^3}.$$

We then end up with

$$\begin{aligned}\ln(1+b) &= b + u(x)\cdot v(x)\Big|_{x=0}^{x=b} - \int_{x=0}^{b} u(x)\cdot v'(x)\,\mathrm{d}x \\ &= b + \frac{1}{2}(x-b)^2\cdot\frac{1}{(1+x)^2}\Big|_{x=0}^{x=b} + \int_{x=0}^{b}\frac{1}{2}(x-b)^2\cdot\frac{2}{(1+x)^3}\,\mathrm{d}x \\ &= b - \frac{b^2}{2} + \int_{x=0}^{b}(x-b)^2\cdot\frac{1}{(1+x)^3}\,\mathrm{d}x.\end{aligned}$$

And now this integral has an unknown value (and is therefore written in red), but it is even smaller than the previous unknown integral, because of

$$\left|(x-b)^2\cdot\frac{1}{(1+x)^3}\right| \le b^2 \qquad \text{if } 0\le x\le b,$$

and consequently the value of the integral is at most $b^2\cdot b$, which is in our case 0.008. Hence we have determined that $\ln(1.2)$ is somewhere between $0.2-0.02-0.008=0.172$ and $0.2-0.02+0.008=0.188$. It seems that we are making progress.

And now we perform partial integration a third time, but with $u'(x)=(x-b)^2$ and $v(x)=\frac{1}{(1+x)^3}$, hence $u(x)=\frac{1}{3}(x-b)^3$ and $v'(x)=\frac{-3}{(1+x)^4}$. If you do the maths, you get

$$\ln(1+b) = b-\frac{b^2}{2}+\frac{b^3}{3}+\int_{x=0}^{b}(x-b)^3\cdot\frac{1}{(1+x)^4}\,\mathrm{d}x.$$

We quickly figure out that now this red integral is bounded by $b^3\cdot b=b^4=0.0016$, and therefore

$$\ln(1+b)\approx b-\frac{b^2}{2}+\frac{b^3}{3} \qquad \text{with error at most} \quad b^4,$$

which shall be precise enough for our purposes.

Now what is the famous MACLAURIN series ? If $f=f(x)$ is a function for which all the derivatives appearing in the next formula exist everywhere, then

$$f(x)\approx f(0)+f'(0)\cdot x+\frac{1}{1\cdot 2}f''(0)\cdot x^2+\frac{1}{1\cdot 2\cdot 3}f'''(0)\cdot x^3+\frac{1}{1\cdot 2\cdot 3\cdot 4}f''''(0)\cdot x^4+\dots,$$

and in the courses on *Calculus* and *Analysis* you will learn what the $\approx$ and the dots at the end actually mean. A *series* is, roughly spoken, a sum with an infinite number of items to be added (the dots at the end hint at the many missing items). The hope is that most of these infinitely many items in the summation can be neglected (if they cannot because they are big, you are in trouble). In our case, we have $f(x)=\ln(1+x)$. The proof of the Maclaurin formula goes as above.

**DIY:** *Explain why the first function $u(x)$ had been chosen as $u(x)=x-b$, instead of (the at first glance perhaps more attractive choice) $u(x)=x$ ?*

### 2.3.4 More Geometry

Suppose that there are two curves in the plane, and these two curves intersect somehow:

How can we determine the angle under which they intersect ? We could use a protractor but this requires that these two curves actually exist in the physical world; and that it is safe enough to go to the intersection point and hold a protractor there.

How about the situation that the two curves are given by means of formulas, but they do not exist in physical reality, and we have to figure out their intersection angle by way of mathematical reasoning ?

First of all a clarification is needed: when mathematicians speak about curves, they mean lines that may be heavily bent or slightly bent or even perfectly straight lines. Allowing a curve to be non-bent and still calling it “curve” is mathematically much easier.

Second: describing curves needs knowledge from the courses on *Calculus* and *Analysis* (because you need to handle the curvature somehow), but figuring out angles needs knowledge from courses on *Geometry.*

So we have to combine knowledge from different disciplines of mathematics. Bringing knowledge together from various parts of mathematics, and from other scientific fields, will be unavoidable in your future, but for today let us cheat a bit and pretend that the two curves are actually straight lines:

Then we can find the answer inside geometry alone, and calculus is not needed. Let the intersection point be $O$, and let $A$ be a point on one straight line, and $B$ a point on the other straight line. In the triangle $\triangle OAB$, we wish to figure out the angle at $O$, without using a protractor.

Let us devise a co-ordinate system, which has its centre (its origin) at the point $O$, and suppose that we know the coordinates of all three points:

$$O = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad A = \begin{bmatrix} x_a \\ y_a \end{bmatrix}, \qquad B = \begin{bmatrix} x_b \\ y_b \end{bmatrix}.$$

We also introduce two vectors: $\overrightarrow{a} = \overrightarrow{OA}$ points from $O$ to $A$, and $\overrightarrow{b} = \overrightarrow{OB}$ points from $O$ to $B$. Their coordinates are

$$\overrightarrow{a} = \begin{bmatrix} x_a \\ y_a \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{pmatrix} x_a \\ y_a \end{pmatrix}, \qquad \overrightarrow{b} = \begin{bmatrix} x_b \\ y_b \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{pmatrix} x_b \\ y_b \end{pmatrix}.$$

The lengths of the vectors $\overrightarrow{a}$ and $\overrightarrow{b}$ can be determined citing Pythagoras:

$$\|\overrightarrow{a}\| = \sqrt{x_a^2 + y_a^2}, \qquad \left\|\overrightarrow{b}\right\| = \sqrt{x_b^2 + y_b^2},$$

and the expression $\|\overrightarrow{a}\|$ is called the *norm of the vector* $\overrightarrow{a}$. It is exactly equal to the length of the line segment $\overline{OA}$.

We still need to figure out the angle at the point $O$. Call it $\gamma$. An ingenious idea will help us: we calculate the length $\left\|\overrightarrow{AB}\right\|$ of the vector $\overrightarrow{AB}$ **twice**, and then we compare the results.

**First calculation:** the vector $\overrightarrow{AB}$ has coordinates

$$\overrightarrow{AB} = B - A = \begin{bmatrix} x_b \\ y_b \end{bmatrix} - \begin{bmatrix} x_a \\ y_a \end{bmatrix} = \begin{pmatrix} x_b - x_a \\ y_b - y_a \end{pmatrix},$$

and therefore its length can be determined by appealing to Pythagoras:

$$\left\|\overrightarrow{AB}\right\| = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2}.$$

**Second calculation:** the cosine theorem in the triangle $\triangle OAB$ tells us

$$\left\|\overrightarrow{AB}\right\|^2 = \left\|\overrightarrow{OA}\right\|^2 + \left\|\overrightarrow{OB}\right\|^2 - 2\left\|\overrightarrow{OA}\right\| \cdot \left\|\overrightarrow{OB}\right\| \cdot \cos(\gamma).$$

**Comparison of both results:** Squaring the first result and setting it equal to the second result:

$$(x_b - x_a)^2 + (y_b - y_a)^2 = \left\|\overrightarrow{OA}\right\|^2 + \left\|\overrightarrow{OB}\right\|^2 - 2\left\|\overrightarrow{OA}\right\| \cdot \left\|\overrightarrow{OB}\right\| \cdot \cos(\gamma).$$

Let us substitute the shorter names:

$$(x_b - x_a)^2 + (y_b - y_a)^2 = \|\overrightarrow{a}\|^2 + \left\|\overrightarrow{b}\right\|^2 - 2\|\overrightarrow{a}\| \cdot \left\|\overrightarrow{b}\right\| \cdot \cos(\gamma).$$

We expand the LHS, and the first two items on the RHS:

$$\left(x_b^2 - 2x_b \cdot x_a + x_a^2\right) + \left(y_b^2 - 2y_b \cdot y_a + y_a^2\right) = \left(x_a^2 + y_a^2\right) + \left(x_b^2 + y_b^2\right) - 2\|\overrightarrow{a}\| \cdot \left\|\overrightarrow{b}\right\| \cdot \cos(\gamma).$$

We eliminate equal terms on both sides:

$$-2x_b \cdot x_a - 2y_b \cdot y_a = -2\|\overrightarrow{a}\| \cdot \left\|\overrightarrow{b}\right\| \cdot \cos(\gamma).$$

We solve for $\cos(\gamma)$:

$$\cos(\gamma) = \frac{x_a \cdot x_b + y_a \cdot y_b}{\|\overrightarrow{a}\| \cdot \left\|\overrightarrow{b}\right\|} = \frac{x_a \cdot x_b + y_a \cdot y_b}{\left(\sqrt{x_a^2 + y_a^2}\right) \cdot \left(\sqrt{x_b^2 + y_b^2}\right)}.$$

**We celebrate:** every item on the RHS is known, and now we only have to do the arcus cosine in order to obtain $\gamma$. This angle is from the interval $[0, \pi]$.

We introduce a notation: the expression

$$\overrightarrow{a} \cdot \overrightarrow{b} := x_a \cdot x_b + y_a \cdot y_b$$

is called *scalar product of the vectors $\overrightarrow{a}$ and $\overrightarrow{b}$*, and we also have the second formula

$$\overrightarrow{a} \cdot \overrightarrow{b} = \|\overrightarrow{a}\| \cdot \left\|\overrightarrow{b}\right\| \cdot \cos\left(\angle(\overrightarrow{a}, \overrightarrow{b})\right).$$

In our above case, the vectors $\overrightarrow{a}$ and $\overrightarrow{b}$ live in $\mathbb{R}^2$, but this is not a restriction — they can be members of some $\mathbb{R}^n$ with a large $n$. Let us fix the notation for that case:

$$\overrightarrow{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}, \qquad \overrightarrow{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix},$$

and then the scalar product becomes $\overrightarrow{a} \cdot \overrightarrow{b} := a_1b_1 + a_2b_2 + \ldots + a_nb_n$, and the angle $\angle(\overrightarrow{a}, \overrightarrow{b})$ between $\overrightarrow{a}$ and $\overrightarrow{b}$ is determined by means of

$$\cos\left(\angle(\overrightarrow{a}, \overrightarrow{b})\right) = \frac{\overrightarrow{a} \cdot \overrightarrow{b}}{\|\overrightarrow{a}\| \cdot \left\|\overrightarrow{b}\right\|}.$$

An application of scalar products will come up in Chapter 7, where we will discuss how a mobile phone connects to the base station, and how this mobile phone is able to reliably reconstruct the electronic signal that has been sent from the antenna of the base station to the antenna of the mobile phone. Reconstructing this signal (and removing electronic noise) is necessary because otherwise you cannot hear what the person at the other end of the conversation is talking. Then the vectors $\overrightarrow{a}$ and $\overrightarrow{b}$ will be members of $\mathbb{R}^{16}$, and in that sixteen dimensional space we will then need to use a scalar product.

Maths applications appear in the most unexpected places.

**DIY:** *Let us be given two planes $\mathcal{P}_1$ and $\mathcal{P}_2$ in $\mathbb{R}^3$. Those two planes are being described by the equations $3x - 2y + z = 1$ (for $\mathcal{P}_1$) and $2x + 4y + 3z = 17$ (for $\mathcal{P}_2$). Figure out their angle of intersection.*

## 2.4 Some Easy Aspects of Physics

### 2.4.1 Motions in $\mathbb{R}^1$

We consider $\mathbb{R}^1$ first because it is easier. Assume that there is a street without curves, and we consider the motion of a vehicle along that road. In order to do mathematics, we need numbers, and these numbers typically refer to a coordinate system.

Now what is a coordinate system ? You have a coordinate system on that street after you have specified

**the origin:** just select a certain location on that street and say out loud “Here is zero”

**the unit of length:** most people will choose “metres”

**a direction:** this means that you have to clarify how to distinguish between “going forward” and “going backward”.

Physics people call it “frame of reference”, but we shall not adopt that terminology.

Now the purpose of the vehicle is to travel as time goes on, and hence you need to choose a unit of time, which is seconds for us.

The position of the vehicle at time $t$ shall be $p(t)$, which is a real number. We consider the vehicle as a point (otherwise we would have to specify whether $p(t)$ refers to the front of the car, or the back, or whatever). Because $p(t)$ refers to a location along the road, its unit will be *automatically* metres.

The velocity of the vehicle at time $t$ then is $v(t)$, and by definition this is

$$v(t) := p'(t).$$

What is the unit of $v(t)$ ? According to (2.12), we have

$$v(t) \approx \frac{p(t+k)-p(t)}{k}, \qquad \text{where} \quad k \quad \text{is tiny,}$$

and the top of the fraction has unit "metres", the bottom of the fraction has unit[15] "seconds", and then the velocity $v(t)$ has *automatically* the unit "metres divided by seconds".

We take an example: at time $t = 0$, the vehicle is at position 3 (metres), which we write as $p(0) = 3$. Assuming we know $v(t)$, where is the vehicle at time $t = 10$ (hence ten seconds later) ?

To find this out, we integrate the equation $v(t) = p'(t)$ over the time interval $[0, 10]$:

$$\int_{t=0}^{10} v(t)\,\mathrm{d}t = \int_{t=0}^{10} p'(t)\,\mathrm{d}t,$$

and now we apply the Fundamental Theorem of Calculus on the RHS:

$$\int_{t=0}^{10} v(t)\,\mathrm{d}t = p(10) - p(0) = p(10) - 3.$$

Let us re-arrange this:

$$\underbrace{p(10)}_{\text{final position}} \quad = \quad \underbrace{3}_{\text{initial position}} \quad + \quad \underbrace{\int_{t=0}^{10} v(t)\,\mathrm{d}t}_{\text{distance travelled}} \quad .$$

$$\boxed{\text{The velocity is the rate of change of the position.}}$$

Whenever we speak about velocity, we always refer to the *instantaneous velocity*, and never to the *average velocity*, which is defined as

$$\text{average velocity} = \frac{\text{distance travelled during long time interval}}{\text{duration of that long time interval}}.$$

In case that the velocity $v(t)$ is constant, both have the same value. But in many cases, the velocity $v(t)$ changes over time, and then the instantaneous velocity is not the same as the average velocity.

**DIY:** *One of these two velocities is the slope of the secant line in the graph of $p(t)$, the other is the slope of the tangent line. Who is who ?*

How can the velocity change ? The vehicle could get faster (called acceleration), or it could get slower (called deceleration). To have a formula, we say that the acceleration of the vehicle at time $t$ is called $a(t)$, and by definition this is

$$a(t) := v'(t).$$

What is the unit of $a(t)$ ? According to (2.12), we have

$$a(t) \approx \frac{v(t+k)-v(t)}{k}, \qquad \text{where} \quad k \quad \text{is tiny,}$$

and the top of the fraction has unit "metres divided by seconds", the bottom of the fraction has unit "seconds", and then the acceleration $a(t)$ has *automatically* the unit "metres per (seconds squared)".

We take an example: at time $t = 0$, the vehicle has velocity $23\frac{\text{metres}}{\text{seconds}}$, which we write as $v(0) = 23$. Assuming we know $a(t)$, what is the velocity of the vehicle at time $t = 17$ (hence 17 seconds later) ?

[15] The reason why the bottom of the fraction has unit "seconds" is also automatic: the variable $t$ has seconds as unit, and therefore the variable $k$ must have the same unit as $t$, because otherwise you cannot add $t + k$. Think about it: you cannot add 7 seconds to 230 Volts, because it makes no sense.

To find this out, we integrate the equation $a(t) = v'(t)$ over the time interval $[0, 17]$:

$$\int_{t=0}^{17} a(t)\,\mathrm{d}t = \int_{t=0}^{17} v'(t)\,\mathrm{d}t,$$

and now we apply the Fundamental Theorem of Calculus on the RHS:

$$\int_{t=0}^{17} a(t)\,\mathrm{d}t = v(17) - v(0) = v(17) - 23.$$

Let us re-arrange this:

$$\underbrace{v(17)}_{\text{final velocity}} = \underbrace{23}_{\text{initial velocity}} + \underbrace{\int_{t=0}^{17} a(t)\,\mathrm{d}t}_{\text{change of velocity}} .$$

$$\boxed{\text{The acceleration is the rate of change of the velocity.}}$$

And also now, we always refer to the *instantaneous acceleration*, never to some *average acceleration*.

If $a(t) > 0$, then the velocity of the car (with respect to the chosen direction of the road) is getting higher at the time $t$. And if $a(t) < 0$, then the car decelerates.

Now $a(t)$ is the derivative of $v(t)$, which is the derivative of $p(t)$, and therefore $a(t)$ can be understood as *second derivative* of the position $p(t)$:

$$a(t) = p''(t).$$

Physicists prefer to express time derivatives using dots over the function name, and their notation then looks as follows:

$$v(t) = \dot{p}(t), \qquad a(t) = \dot{v}(t), \qquad a(t) = \ddot{p}(t),$$

and this means exactly the same as $v(t) = p'(t)$, $a(t) = v'(t)$, $a(t) = p''(t)$. There are also the notations

$$v(t) = \frac{\mathrm{d}p(t)}{\mathrm{d}t} = \frac{\mathrm{d}p}{\mathrm{d}t}(t) = \frac{\mathrm{d}}{\mathrm{d}t}p(t), \qquad a(t) = \frac{\mathrm{d}v(t)}{\mathrm{d}t}, \qquad a(t) = \frac{\mathrm{d}^2 p(t)}{\mathrm{d}t^2}.$$

Pay attention to where the superscripts “2” are written.

Now consider an example: At time $t = 0$, a car is at the initial position $p(0) = 5$, and it has the initial velocity $v(0) = 25$. There is an obstacle at position $c = 40$. The driver applies the brake, and therefore the acceleration is $a(t) = -6$ for all $t$.

Will the car hit the obstacle ? When ? Interpret your answers and discuss their validity. The letter $c$ stands for “cockroach”, why ?

### 2.4.2 Motions in $\mathbb{R}^3$

Having understood motions along a line, we now discuss motions in three dimensional space as it surrounds us. Imagine a satellite travelling around the earth, or a shotput at a sports event. Again we need a coordinate system, for otherwise we can’t do mathematics.

We will have a coordinate system in three dimensional space after we have specified

**the origin:** just select a certain location and say out loud “Here is $[0, 0, 0]^\top$”

**the unit of lengths:** most people will choose “metres”

**three basis directions:** take your right hand, and move your fingers in such a way that you have three right angles between your thumb, your index finger, your middle finger. Hold this hand at the origin, and then thumb, index finger, middle finger will point along three axes. On these three axes, you then draw the base vectors of length one.

This is a “frame of reference”.

Now consider a motion, for instance a satellite travelling around the earth in a circular manner, or some other Unspecified Flying Object. How to choose the coordinate system ?

The origin should be chosen as the centre, because this is the most natural way. And for the choice of the three axes, we note that the motion actually happens inside a plane, so we hold our hand in such a way that thumb and index finger are inside that plane, and the middle finger perpendicular to that plane. Almost every other choice would make our life much harder. If now the object flies along a circle with radius $R = 17$, and if it needs $2\pi \approx 6.28$ seconds for one revolution, then the position at time $t$ is

$$P(t) = \begin{bmatrix} 17\cos(t) \\ 17\sin(t) \\ 0 \end{bmatrix}.$$

This formula is natural because the functions $\cos(t)$ and $\sin(t)$ have been introduced in Section 2.2.4 exactly in this fashion. The third component is zero for all times $t$ because the motion always stays inside that plane that is being spanned by our first two fingers. And we write square brackets $[\dots]$ instead of round brackets $(\dots)$ because $P$ refers to a point in space, and you cannot add two points.

And if the object needs $T$ seconds for one revolution, then the formula for its position is

$$P(t) = \begin{bmatrix} R\cos(2\pi t/T) \\ R\sin(2\pi t/T) \\ 0 \end{bmatrix}.$$

This expression $\frac{2\pi}{T}$ will appear now in many places, and therefore we introduce an abbreviation:

$$\omega = \frac{2\pi}{T},$$

with $\omega$ pronounced *omega*. If $f$ is the number of revolutions during one second, then we also have the formula $\omega = 2\pi f$. This $\omega$ is called *angular frequency*, and its unit is $\frac{1}{\text{seconds}}$ because $T$ has unit “seconds”.

Next we discuss velocities. By definition, *the velocity is the rate of change of position*, hence the velocity of that flying object is

$$\overrightarrow{v}(t) = P'(t) = \frac{\mathrm{d}}{\mathrm{d}t}P(t) = \frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} R\cos(\omega t) \\ R\sin(\omega t) \\ 0 \end{bmatrix} = \begin{pmatrix} -R\cdot\sin(\omega t)\cdot\omega \\ R\cdot\cos(\omega t)\cdot\omega \\ 0 \end{pmatrix} = \begin{pmatrix} -R\omega\sin(\omega t) \\ R\omega\cos(\omega t) \\ 0 \end{pmatrix},$$

because the derivative of the function $t \mapsto R\cos(\omega t)$ is exactly $-R\cdot\sin(\omega t)\cdot\omega$, by the chain rule, which explains the top component of the vector $\overrightarrow{v}$. We write now round brackets $(\dots)$ because two velocities can be added meaningfully.

**DIY:** *Select $\omega = 1$ for simplicity and figure out (for various choices of time $t$), in which direction the vector $\overrightarrow{v}(t)$ points.*

*Determine the length $\|\overrightarrow{v}(t)\|$ of the vector $\overrightarrow{v}(t)$. Does this length depend on the value of the time $t$ ?*

*Determine the unit of the length $\|\overrightarrow{v}(t)\|$.*

Ultimately, we come to accelerations. By definition, *the acceleration is the rate of change of velocity*, hence the acceleration of that flying object is

$$\overrightarrow{a}(t) = \overrightarrow{v}'(t) = \frac{\mathrm{d}}{\mathrm{d}t}\overrightarrow{v}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} -R\omega\sin(\omega t) \\ R\omega\cos(\omega t) \\ 0 \end{pmatrix} = \begin{pmatrix} -R\omega\cdot\cos(\omega t)\cdot\omega \\ -R\omega\cdot\sin(\omega t)\cdot\omega \\ 0 \end{pmatrix} = \begin{pmatrix} -R\omega^2\cos(\omega t) \\ -R\omega^2\sin(\omega t) \\ 0 \end{pmatrix},$$

where we have differentiated each of the three components of $\overrightarrow{v}(t)$ separately, and applied the chain rule.

**DIY:** *Select $\omega = 1$ for simplicity and figure out (for various choices of time $t$), in which direction the vector $\overrightarrow{a}(t)$ points.*

*Determine the length $\|\overrightarrow{a}(t)\|$ of the vector $\overrightarrow{a}(t)$. Does it change with $t$ ?*

*Determine the unit of the length $\|\overrightarrow{a}(t)\|$.*

To summarise some main ideas:

- we introduce a coordinate system,
- the velocity is the rate of change of the position,
- the acceleration is the rate of change of the velocity,
- when we take time derivatives, we do it for each component of the point/vector separately.

### 2.4.3 Forces

What are forces, and how do they affect motions ? An example of a force is gravity, which pulls us towards the centre of the earth.

A general principle is: if no total force acts upon a body, then this body keeps moving undisturbed along a straight line with constant velocity — provided that the coordinate system is an *inertial frame of reference.*

Roughly spoken, an inertial frame of reference is such a coordinate system in which the equations of motion become the easiest. Examples of such inertial frames are:

**a standing person on the pavement:** the gravity pulls that person down, but the tarmac pushes the person up, both forces cancel, hence the total force equals zero. The person does not move at all. The coordinate system is chosen in such a way that the origin $[0,0,0]^\top$ is at that person.

**a stone falling in water with constant speed (observer system):** a stone is falling in a lake. The gravity pulls the stone down, the friction force between stone and water pushes upwards, both forces cancel, hence the total force is zero. The velocity does not change with $t$. The coordinate system is chosen in such a way that the origin $[0,0,0]^\top$ is with an observer standing at the shore. If the speed is two metres per second downwards, then the velocity is $(0,0,-2)^\top$, which does not depend on $t$.

**a stone falling in water with constant speed (stone system):** the same situation, but now the stone drags its own coordinate system along. Which means that the origin $[0,0,0]^\top$ is always with the stone, and the velocity is then $(0,0,0)^\top$. The velocity *always* refers to the coordinate system.

**a person travelling in a car with constant speed:** a person sits as a passenger in a car that travels along a straight road with constant speed. The gravity pulls the person down, the seat base resists the gravity and pushes the person up, both forces cancel. The coordinate system with an outside observe is an inertial frame, the passenger system (where the passenger drags the origin with themselves) is an inertial frame as well.

Non-inertial frames have the awkward property that there are mysterious forces that you cannot explain, and examples are the following two:

**accelerating car (passenger system):** now the car accelerates forward, and the passenger drags its own coordinate system along. Because the passenger is at any time at the position $[0,0,0]^\top$, their velocity is always the vector $(0,0,0)^\top$. *But* now the passenger feels a mysterious force emanating from the backrest that pushes the passenger forward, and this force cannot be explained from the passenger system alone.

**rotating bucket of water (rotating system):** consider a half-full bucket of water that rotates about its axis, and the coordinate system rotates as well. The gravity pulls the water molecules down, but the lower water molecules support the upper water molecules, hence the explainable forces cancel, and the velocity of the water molecules is $(0,0,0)^\top$ because the coordinate frame rotates with same angular velocity as the bucket. *But* now there is a mysterious force that makes the water surface curved, and the water goes up along the bucket walls, which cannot be explained from the rotating system alone.

That has been *Newton's First Law.* We also need the second one: having chosen an inertial frame of reference, the total force acting upon a body is proportional to the acceleration of that body, and the factor of proportionality equals the mass of that body:

$$\overrightarrow{F}_{\text{total}} = m \cdot \overrightarrow{a},$$

where $\overrightarrow{F}_{\text{total}}$ is the total force acting upon the body, $m$ is the mass of the body, and $\overrightarrow{a}$ is the acceleration of the body.

An example: at time $t=0$, a body with mass $m=5$ kilograms is at position $[0,0,0]^\top$, at rest. A spring pushes the body to the right, with force of 27 Newtons, where Newton $= \frac{\text{kilogramm}\cdot\text{metre}}{\text{second}^2}$. How does the body move ?

Let $P(t)$ be the position of the body at time $t$. Then

$$P(0) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad P'(0) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

because for $t=0$, the body is at the origin, and it is at rest. The force is

$$\overrightarrow{F}_{\text{total}} = \begin{pmatrix} 27 \\ 0 \\ 0 \end{pmatrix},$$

because our right-hand thumb points from left to right. Now we have

$$P''(t) = \overrightarrow{a}(t) = \frac{1}{m}\overrightarrow{F}_{\text{total}} = \frac{1}{5}\begin{pmatrix} 27 \\ 0 \\ 0 \end{pmatrix},$$

and therefore

$$P(t) = \begin{bmatrix} \frac{27}{10}t^2 \\ 0 \\ 0 \end{bmatrix},$$

and this is valid for all times $t$ for which there is a contact between the spring and the body.

**DIY:** *Verify that the above presented $P(t)$* actually is *a solution.*

We will not discuss Newton's Third Law because we do not need it.

We conclude with one more example. TV-satellites that transmit your TV programmes have to stay over the UK, always at the same position in the sky, for obvious reasons. What is their distance to the earth ?

To answer this, we need a formula for the gravitational force. If there are two bodies of masses $m_1$ and $m_2$, and distance $D$ to each other, then the gravitational force between them is

$$\gamma \cdot \frac{m_1 \cdot m_2}{D^2},$$

with $\gamma = 6.674 \cdot 10^{-11} \frac{\text{Newton}\cdot\text{metre}^2}{\text{kilogram}^2}$ being the gravitational constant.

If $P(t)$ denotes the position of the satellite, then (in a cleverly chosen coordinate system)

$$P(t) = \begin{bmatrix} D\cos(\omega t) \\ D\sin(\omega t) \\ 0 \end{bmatrix},$$

with $D$ being the distance to the earth centre, and $\omega = \frac{2\pi}{T}$, and $T$ is the time needed for one revolution, which is 23 hours, 56 minutes, 4 seconds.

**DIY:** *Explain these funny numbers.*

Therefore $T = 23 \cdot 3600 + 56 \cdot 60 + 4 = 86164$ seconds. The acceleration has been already calculated as

$$\overrightarrow{a}(t) = \begin{pmatrix} -D\omega^2\cos(\omega t) \\ -D\omega^2\sin(\omega t) \\ 0 \end{pmatrix}.$$

The gravitational force, which pulls the satellite towards the earth centre, is

$$\overrightarrow{F}_{\text{gravity}}(t) = \gamma \cdot \frac{m_{\text{earth}} \cdot m_{\text{satellite}}}{D^2}\begin{pmatrix} -\cos(\omega t) \\ -\sin(\omega t) \\ 0 \end{pmatrix}.$$

We should perhaps clarify what we know, and what we do not know:

**known:** $\gamma$, $T$, $\omega$, $m_{\text{earth}} = 5.972 \cdot 10^{24}$ kilograms,

**unknown:** $m_{\text{satellite}}$, $D$.

Now Newton's second law says $\overrightarrow{F}_{\text{gravity}} = m_{\text{satellite}} \cdot \overrightarrow{a}(t)$, hence

$$\gamma \cdot \frac{m_{\text{earth}} \cdot m_{\text{satellite}}}{D^2}\begin{pmatrix} -\cos(\omega t) \\ -\sin(\omega t) \\ 0 \end{pmatrix} = m_{\text{satellite}} \cdot \begin{pmatrix} -D\omega^2\cos(\omega t) \\ -D\omega^2\sin(\omega t) \\ 0 \end{pmatrix},$$

and therefore $\gamma \cdot m_{\text{earth}} = D^3\omega^2$, resulting in $D = 42163112$ metres, which is about 35792 kilometres above the surface of the earth (which has a radius of 6371 kilometres).

## 2.5 How to do Mathematics

### 2.5.1 The Purpose of Homework at University

- You cannot learn how to swim by reading a book about swimming.

  You cannot learn how to swim by attending a lecture about swimming in a lecture theatre.

  You cannot learn how to swim by repeatedly asking the lecturer (or the personal tutor) to explain it again and again and once again.

  If you wish to learn how to swim, you cannot avoid going into the water, which implies getting wet.

  The same applies to mathematics, and for this reason, this set of lecture notes contains many do-it-yourself (**DIY**) exercises, and it is crucial for your success that you actually do them.

  At all Scottish universities, students have a workload of 600 hours per semester, which means 150 hours per course. Assuming a term duration of 10 weeks, you are expected to spend 15 hours each week on the *Maths in Context* course. These 15 hours comprise four contact hours in the form of lectures and exercises, and consequently 11 hours of self study at home. A reasonable splitting could be to spend eight hours on the homework sheet, and three hours reading this set of lecture notes. The questions on the homework sheets have been chosen accordingly.

- Universities are places of *Higher Education*, not places of *Vocational Training.* How do they differ ?

  > “To be prepared against surprise is to be *trained.* To be prepared for surprise is to be *educated.* . . . Training repeats a completed past in the future. Education continues an unfinished past into the future.”[16]

  An obvious and unavoidable surprise for you is that nobody can know where you will be 8 years from today — in which sector of the economy, in which profession, at which company. But we do know that whatever amount of factual knowledge you acquire at university, this factual knowledge will always differ from what will be needed in your future job(s).

- At school, the teachers explained things to you, then gave you questions about these things, they told you the answers, then you learned these answers, and then you had the assessment. Assessments at school are exceptionally bizarre rituals: teachers ask you questions, but they already know the answer. So why are they asking ?

  In a company, the situation is different: your colleague/your boss/your customer asks you something, and they do not know the answer (otherwise they would not ask). Either you will know the answer, or you won’t, in which case you *will have* to find it somehow, and nobody will help you, because nobody can.

  One of the purposes of university is to prepare your mindset for situations of that nature.

- Therefore, you will have homeworks with questions that you have never seen before, and perhaps these questions deal with topics that have not yet been covered in class. That is by design.

- In the course *Maths in Context*, we don’t have homeworks of the type “Plug and Chug” because computers have been invented for a reason.

- If you see some homework question for the first time, and don’t really know where to start — this feeling is normal; most of the other students feel the same. A homework sheet is typically solvable in eight hours. Look up all the scientific words (in the lecture notes, in Wikipedia, in books). Read the lecture notes. Ask a fellow student. Ask another fellow student. Repeat this procedure the next day. After some time, you *will* understand it. And you *will* remember it, because you worked for it.

  But **never** ask the personal tutor, or the tutorial helpers of other courses.

- Some students prefer to learn from past exam papers. I have never understood the purpose of that approach. If you learn that way, you will be able to answer specific questions, provided their wording does not change. But the challenges that await you in the future will be different anyway, hence wouldn’t it be smarter to cherish those challenges, instead of sticking to past exam paper questions that will be water under the bridge next year anyway ?

[16] James Carse, *Finite and Infinite Games*, Simon and Schuster, New York 2011

### 2.5.2 Survival Tips

**Practice as much as possible**

Because everything gets easier if you practice it often enough.

**Distinguish the knowns from the unknowns**

When doing financial accounting, we must distinguish the monetary numbers according to *credit* and *debit*.

When doing a calculation, we must distinguish the mathematical objects according to *given* objects and *wanted* objects.

When doing a proof, we must distinguish the statements according to *assumption/presupposition* and *claim/assertion*.

Whatever you do — when you mix these things up, your piece of work *will* go wrong.

**Understand the problem before solving it**[17]

Look up all the scientific words. Have you seen a similar problem before ? What are the properties of the appearing mathematical objects ? Can you find statements in the lecture notes about these properties ?

If a mathematical variable appears, can you tell its unit ? If not, then the understanding is not deep enough.

**Each mathematical object has a type** The type determines what you can do with it.

A vector is not a number. You can divide by a number, but you cannot divide by a vector.

The length of a vector is not the same as this vector. A vector is not a point.

A function is not the same as a number (because a function is a machine with input and output).

**Distinguish a function and its value**

A function $f$ is a machine that takes something in (often called $t$) and produces something (which is the value). In contrast to that, the expression $f(t)$ is the *value* of the function $f$.

**Test formulas with numbers**

If you have a formula and don't know whether it is right or wrong, then test it with numbers. This will not constitute a proof, but it can give you some hints.

**Do quick estimates, or back-of-the-envelope calculations**

If you have a mathematical expression with various components, you should be able to quickly figure out which of these components is typically big, and which one is normal sized, an which is small. As an analogy: if a group of people has a boss, it is advantageous to know who that is.

**Switch points of view**

For instance, any deeper understanding of functions requires all three representations of them.

**Don't copy mindlessly**

Cooperation with fellow students can be very helpful, but you should not copy from them without thinking, because everybody gets things wrong from time to time. You will burn your name if you and seven other students submit homework claiming that $\frac{1}{4} + \frac{1}{4} = \frac{1}{8}$.

### 2.5.3 How to Write Mathematics

**Suppose**

the author of a text with mathematical contents has written for formula

$$x = e^y,$$

without leaving any other comment. What could the author have meant ?

[17] Somebody should tell this Whitehall

**Perhaps they have intended to say:**

- By assumption we have $x = e^y$.
- From the presupposition we deduce that $x = e^y$, after a short calculation.
- Owing to the Blafubni Theorem, we have $x = e^y$.
- We define $x \in \mathbb{R}$ via $x := e^y$.
- We define $y \in \mathbb{R}$ via $x =: e^y$. This is doable because the Theorem of Ernesto Binomi guarantees that $x > 0$.
- Suppose we had[18] $x = e^y$. But then blafasel would follow, which then would yield the blubb formula, in contradiction to formula (5). Therefore $x = e^y$ must be wrong.
- The equation (7) is claimed to hold for all $x$. Let us take (7) at their word, substitute $e^y$ for the variable $x$, and look where does this lead us to: ...
- The line $x = e^y$ is the first case of a distinction by cases, and half a page down, another formula $x \neq e^y$ drops down from the sky, as unmotivated as the first line.

**We have no chance of knowing !**

**Conclusion**

We read some text in order to be afterwards more wiser or more knowledgeable than before (because otherwise there is no point in reading it). Therefore, we as readers want to comprehend the line of thought of the author.

A formula without surrounding text is ambiguous. This formula could be

- a part of the assumption,
- a part of the claim,
- a statement that has been proved earlier,
- a statement that will be proved soon,
- a statement that will be refuted soon,
- a part of a distinction by cases,
- a definition (and then it can still be ambiguous which of the various objects is being defined here),
- a guidance to substitute some variable someplace else,
- something else I cannot imagine right now.

The line of thought crucially depends on which of these many bullet points applies.

It is not very polite to leave the reader guessing. Maybe the reader is able to guess what has been meant, but you should not rely on that. The reader is giving you their time, so don't let it go to waste.

Furthermore, it may well be that author and reader are the same person, with a delay of several months. It is smarter to write in such a way that the text is quick to understand.

[18] this subjunctive mood indicates the beginning of an indirect proof by contradiction

### Recommended expressions

Nobody expects you to write the King James Bible when doing homework. The good news is that you can achieve a lot already with three words:

- Put $\boxed{\cdots}$ := $\boxed{\cdots}$.
- Assume . . . .
- Consider . . . .
- $\boxed{\cdots}$ , hence $\boxed{\cdots}$ . (The first box justifies the second box.)
- $\boxed{\cdots}$ , because $\boxed{\cdots}$ . (The second box justifies the first box.)
- Proof by contradiction. Suppose . . . . (OK, occasionally four words.)
- By assumption: . . . .
- Assumption implies . . . . (This is not the same as the previous version which refers to more obvious cases.)
- From $\boxed{\cdots}$ we get $\boxed{\cdots}$ .
- We know . . . .
- Obviously . . . . (Only to be used if it is truely obvious, otherwise not polite.)
- Now we prove that . . . .
- First . . . .
- Second . . . .
- Finally . . . .
- Substitute $x = 0$ in (7): . . . .

### Further Recommendations for Writing

#### Write legibly

Once I had to mark the exam script of a student who wrote $\frac{3}{z}$ somewhere in his calculations (which was correct), and one line below they then wrote 1.5 (wrong). This student had read the letter $z$ for a number 2, and subsequently lost marks. And then I had another student who wrote 2, $x$, $\pi$, $\lambda$, and they looked all the same.

Keep in mind that (for instance) $g$ and $q$ should be distinguishable, and also 5 and $s$. A professional appearance of the script will rub off on you.

#### Give names to objects

Otherwise you will have to resort to cumbersome expressions such as “this function” or “that variable”.

#### Always write full sentences

This helps to keep your own thinking straight.

#### Fraction bars should be at the height of surrounding operations etc

As an example:

$$\frac{16}{\frac{4}{2}} = \frac{16}{4/2} = \frac{16}{2} = 8, \qquad \text{but} \quad \frac{\frac{16}{4}}{2} = \frac{16/4}{2} = \frac{4}{2} = 2.$$

#### Don’t start a sentence with a variable

It is ugly. Instead of writing “$f$ has the derivative. . . ” you could write “The function $f$ has the derivative. . . ”, i.e. just prepend the type of that object.

### 2.5.4 Recommendations for the Exercise Session

Instead of a tutorial part, the course *Mathematics in Context* has something else: an exercise session where students present their solved homework in front of all the other students. Think of it as a practicing to give presentations (but without powerpoint), in a friendly environment.

The most important when giving a presentation is the audience. Perhaps some fellow students could not answer that homework question completely, and your blackboard presentation then is the only information channel to explain them how to solve it (printed example solutions will not be provided).

Some things to keep in mind:

- Your line of thought should become clearly visible.
- Explain in spoken words what you are doing; don't just write in silence.
- Talk to your fellow students, not to the blackboard.
- Think about the font size and about the students sitting in the last row.

## 2.6 The Greek Alphabet

Please get familiar with the Greek alphabet[19]:

| letter | name | remarks |
|---|---|---|
| $\mathrm{A}$, $\alpha$ | alpha | |
| $\mathrm{B}$, $\beta$ | beta | |
| $\Gamma$, $\gamma$ | gamma | |
| $\Delta$, $\delta$ | delta | |
| $\mathrm{E}$, $\epsilon$, $\varepsilon$ | epsilon | |
| $\mathrm{Z}$, $\zeta$ | zeta | |
| $\mathrm{H}$, $\eta$ | eta | |
| $\Theta$, $\theta$, $\vartheta$ | theta | |
| $\mathrm{I}$, $\iota$ | iota | |
| $\mathrm{K}$, $\kappa$ | kappa | |
| $\Lambda$, $\lambda$ | lambda | |
| $\mathrm{M}$, $\mu$ | mu | |
| $\mathrm{N}$, $\nu$ | nu | |
| $\Xi$, $\xi$ | xi | |
| $\mathrm{O}$, $o$ | omikron | |
| $\Pi$, $\pi$ | pi | |
| $\mathrm{P}$, $\rho$, $\varrho$ | rho | |
| $\Sigma$, $\sigma$, $\varsigma$ | sigma | Write $\Sigma$ with 2 strokes because its upper line is horizontal. |
| $\mathrm{T}$, $\tau$ | tau | |
| $\Upsilon$, $\upsilon$ | upsilon | |
| $\Phi$, $\phi$, $\varphi$ | phi | |
| $\mathrm{X}$, $\chi$ | chi | pronounced as "ch" in Scottish "Loch" |
| $\Psi$, $\psi$ | psi | |
| $\Omega$, $\omega$ | omega | |

[19] I can give you the guarantee that you will not need to learn Sütterlin which was my fate when I studied in the Nineties, although I acknowledge that Sütterlin can become useful when writing vectors on the blackboard. In that script, the standard sentence "The quick brown fox jumps over the lazy dog" turns into The quick brown fox jumps over the lazy dog.

# Chapter 3

# Mathematics in Nature — Rainbows

## 3.1 General Remarks

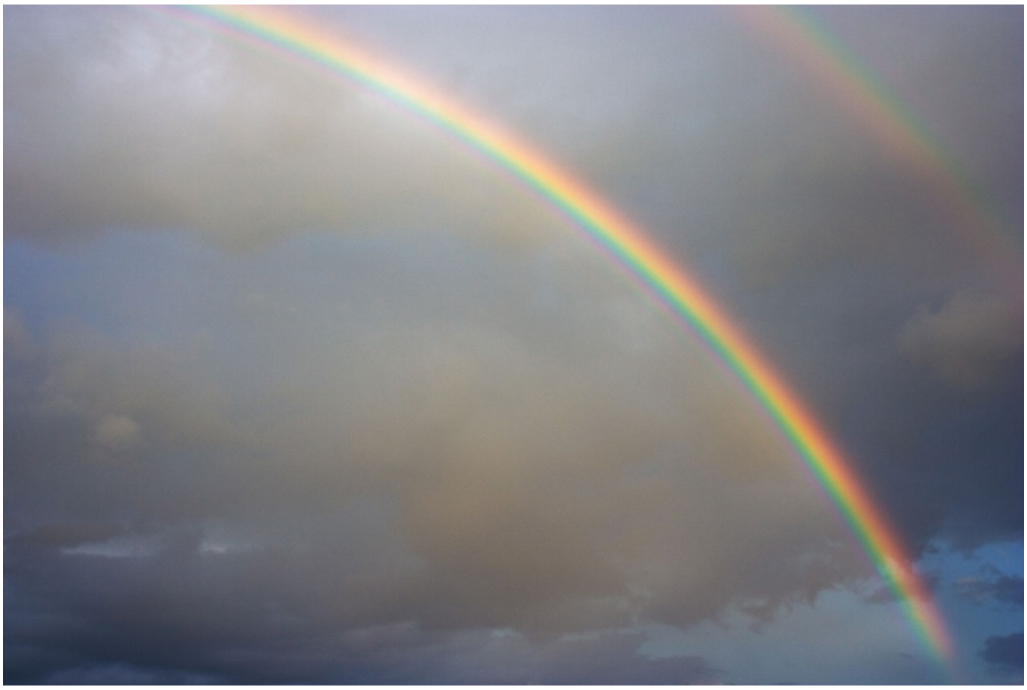


Figure 3.1: Primary (bottom) and secondary rainbow, with Alexander's band inbetween. Note that both rainbows have their colours in different order. Photo (public domain) by Petr Kratochvil.

Rainbows occur in a certain way when it rains and you have sufficiently strong sunshine at the same time. You may also utilise a garden hose.

We intend to answer several questions:

- Rainbows appear to be circular arcs. Are they really circular arcs ?
- Where is the rainbow, relative to our position and the position of the sun ?
- What makes the colours ?
- What makes the secondary rainbow which may become visible in the case of exceptionally strong sunshine ?
- Why is the region between the primary rainbow and the secondary rainbow so dark (called Alexander's band, see Figure 3.1, named after Alexander of Aphrodisias ($\approx$ 200AD)) ?

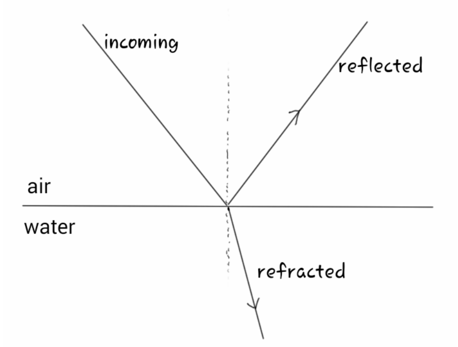
What we will learn:

- a bit about physics
- we will see a bit of how to do modelling
- elementary geometry
- elementary calculus

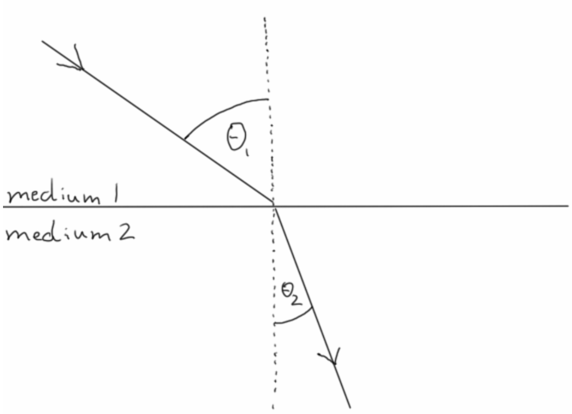
## 3.2 Physical Remarks

There is no colour "white". The white light which we typically see consists of various colours together: red, orange, yellow, green, blue, violet. In vacuum, all these colours propagate with light speed ($\approx 3 \cdot 10^8 m/s$), and in air, the speed is a bit slower. In water, on the other hand, the velocity of light is considerably slower[1] than $3 \cdot 10^8 m/s$.

What happens now if a light-ray (a beam coming from a torch, say) hits a flat water surface ? A part of the light gets reflected and stays in the air, and a part of the light enters the water and gets refracted, changing its direction.



We need to understand the refraction more in detail. By how much changes the direction of the light ? We consider the normal line[2] to the surface (dashed line) and the angles $\theta_1$ and $\theta_2$ of the light beam to the normal lines as in the figure:



Then Snell's Law holds:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2,$$

named after Willebrord Snel van Royen[3], with $n_1$, $n_2$ (which are positive real numbers, without unit) called the *refractive indices* of medium 1 and medium 2, respectively, defined as

$$n_j = \frac{\text{light speed in vacuum}}{\text{(apparent) light speed in medium } j}, \qquad j = 1, 2.$$

The numbers $n_j$ have no unit because the unit "meters per second" in the top of the fraction cancels with the unit "meters per second" in the bottom of the fraction.

However, it is worth noticing that the law of refraction of light was first discovered by Ibn Sahl in Bagdad, who lived $\approx 940 - \approx 1000$.

[1] here we are simplifying heavily, because in electrodynamics and quantum electrodynamics, the velocity of light is the same in any medium, and we only perceive the velocity as smaller because of a certain interaction of the moving particles of light (called photons) and the electrons in the medium through which the light propagates; but we will not make an attempt to dig deeper in this topic.

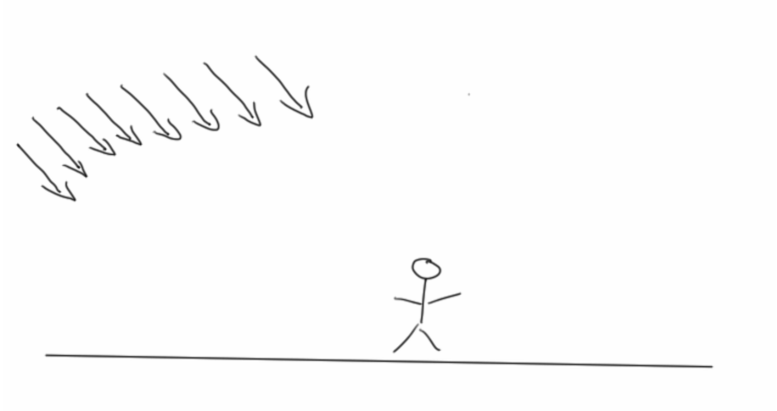
[2] "normal line to a surface" is scientific speak for a line being perpendicular to that surface

[3] 1580 – 1626

The refractive index of air is very close to one (1.0003), and the refractive index of water is approximately 1.333. Relevant for the understanding of a rainbow is now that the refractive index of light in water depends on the colour of the light [8]:

| colour | refractive index |
|---|---|
| red | 1.330 |
| orange | 1.334 |
| yellow | 1.335 |
| green | 1.338 |
| blue | 1.341 |
| violet | 1.349 |

That is almost all the knowledge from physics which we need. One more fact might be relevant: the sun has a distance from the earth of 150 million kilometres (approximately), hence we may assume that the sunlight arrives at the earth in a parallel manner, like a package of not-yet-boiled spaghetti:



## 3.3 First Considerations

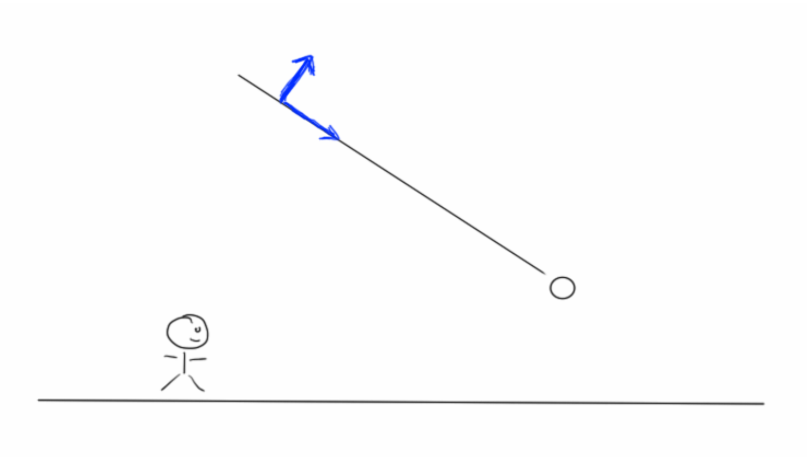
We have parallel rays of light, and we have water droplets in the air. The sunlight hits a droplet, gets reflected and refracted there (perhaps several times), eventually then leaves the droplet and then travels away from the droplet — and if then the eye of a human observer is in the path of that light ray, then this light-ray could contribute to the phenomenon of a rainbow *for that particular observer*. Therefore, we are allowed to ignore all those light-rays that will not arrive at the observer's eye eventually.

Now: there is the sun, the observer, the droplet; and these are objects in $\mathbb{R}^3$, and these three objects define a two-dimensional plane approximately (because three points always define a plane (or a line in exceptional cases)). As soon as reflected or refracted light-rays abandon that plane, they will never get back to the eye of the observer, and such light-rays cannot be relevant.

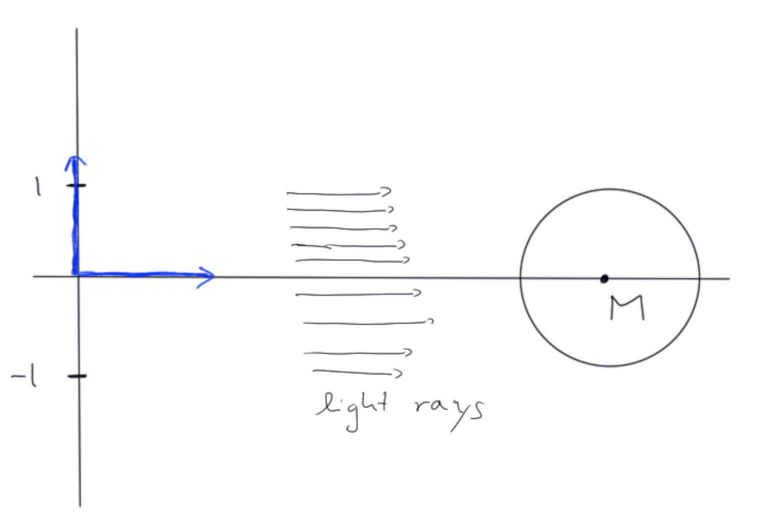
Therefore all our considerations will be done in two dimensions, and the plane that is spanned by the straight line observer–droplet centre and by the straight line observer–sun intersects the droplet in a circle, and only that circle is the relevant part of the droplet.

Now we need a coordinate system in order to describe the path of the light through the air and inside the droplet. Let us pause for a moment and take the time for a slightly philosophical consideration. There is the natural real world which contains the sun, the light, the air, the droplet; and if you believe, you will perhaps agree that it was created by some god; and if you don't believe, you will perhaps agree that this real world does exist even when we humans do not observe it. The real world has *not* been created by humans. On the other hand, coordinate systems are an entirely human convention; they were first introduced by René Descartes around 1643. The purpose of coordinate systems is to help us in describing the real world around us. Therefore, we should arrange the coordinate systems in such a way that our calculations are as easy as possible. The school approach was perhaps different — choose the coordinate systems in a crazy manner in order to make the calculations sufficiently complicated. Everything is fine in School as long as the pupils are kept busy and stay away from thinking.

Keeping this philosophical view in the back of our head, we pick one droplet and arrange the coordinate system as painted blue in the picture to the right.
One axis of the blue coordinate system sits on that sunlight ray that goes directly through the droplet centre point. The unit length is also available for choice, and we define it as the droplet radius.



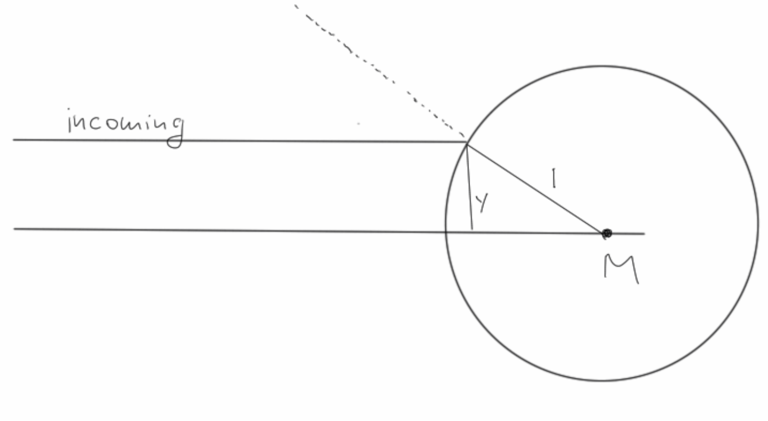
We neglect the human observer for a moment and make a new picture:
Those light-rays that are able to interfere with that chosen droplet can be characterised by a parameter $y \in (-1,1)$, with $y = 0$ meaning that the light-ray aims for the centre of the droplet, $y = 1$ means "a grazing contact at the top end" and $y = -1$ meaning "a grazing contact at the bottom end".
We don't care about the distance of the droplet to the origin of the blue coordinate system, but we let the droplet have its location in the right half-plane, the coordinates of its middle point being $(M, 0)$ with $M > 1$.



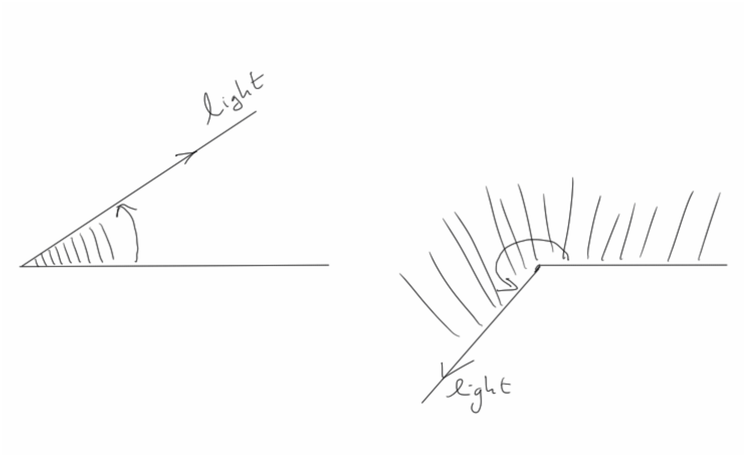
Now we come to reflection and refraction. Choose some $y \in (-1,1)$. The associated light-ray hits the droplet at a contact point $C_1$ with coordinates

$$\begin{bmatrix} M - \sqrt{1-y^2} \\ y \end{bmatrix},$$

using Pythagoras.

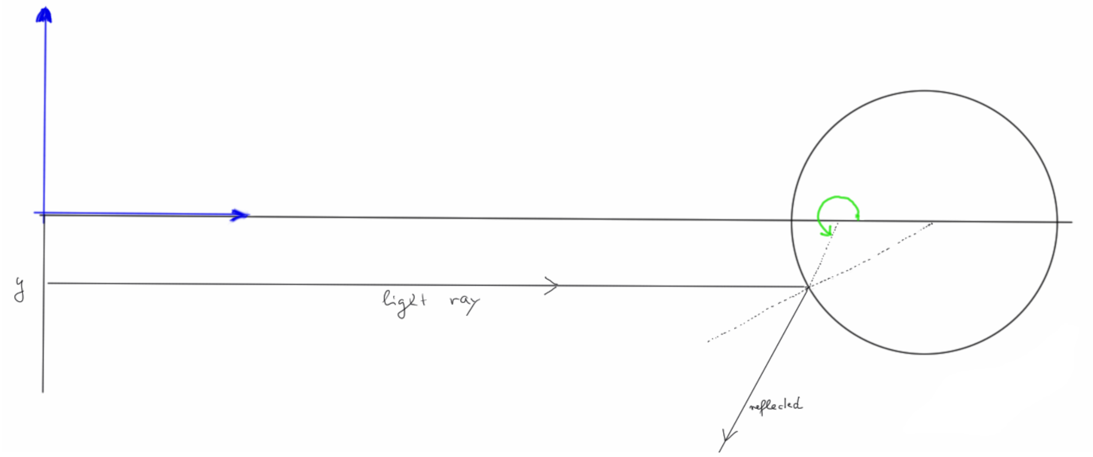


For each relevant ray, we wish to consider its angle to the direction “to the right”. The incoming light-ray has associated directional angle zero, and we count the angles counter-clockwise from 0 to $2\pi$:



Usually, the sun is in your back when you watch a rainbow, hence you expect interesting angles to be approximately in the interval $(135°, 270°)$, otherwise these rays cannot end up in your eyes.

Now we look at the first contact point $C_1$ of the light-ray with the droplet boundary, and the reflected ray there has associated angle (green)

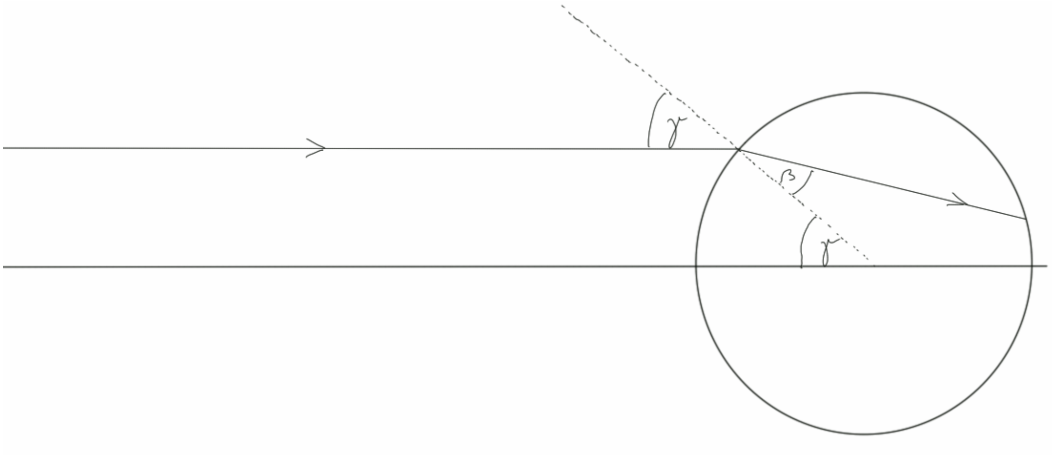


equal to $\alpha_1 = \pi - 2\arcsin y$, where we have chosen the principal branch of the arcsin function that produces values between $-\pi/2$ and $\pi/2$.

**DIY:** *Prove this formula for $\alpha_1$.*

In the following, directional angles for rays in air are called $\alpha_?$, and directional angles for rays in the liquid are called $\lambda_?$.

Consider the first ray that enters the droplet.



SNELL's Law says $n_{air} \cdot \sin\gamma = n_{water} \cdot \sin\beta$, with $\beta$ and $\gamma$ as in the picture. We abbreviate $n_w := n_{water}$ and $n_{air} = 1$, hence

$$\beta = \arcsin\left(\frac{1}{n_w}\sin\gamma\right).$$

Recall $\sin\gamma = y$, which is true also for $-1 < y \le 0$. The directional angle of the first ray in the liquid droplet then is

$$\lambda_1 = \begin{cases} 2\pi - \arcsin y + \arcsin\left(\frac{1}{n_w}y\right) & : 0 \le y < 1, \\ -\arcsin y + \arcsin\left(\frac{1}{n_w}y\right) & : -1 < y < 0. \end{cases} \tag{3.1}$$

**DIY:** *Check this formula for positive $y$ and for negative $y$.*

The right-hand side of (3.1) is to be read as distinguishing two cases: if $0 \le y < 1$, then the upper formula is to be chosen, otherwise the lower formula. Later we will have to handle a second reflection and a third one, which would perhaps then lead us to four cases and eight cases, respectively. On the other hand, the both formulas in (3.1) merely differ by $2\pi$, which corresponds to a full rotation. Distinguishing cases when it is avoidable is not something we are looking forward to with undiluted pleasure, and therefore we make an agreement.

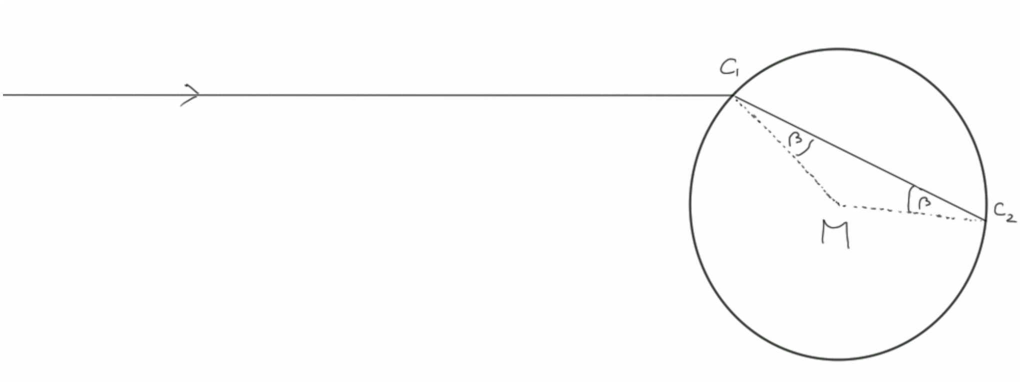
**Convention:** two angles are considered equal if they differ by integer multiples of $2\pi$.

Then we can write

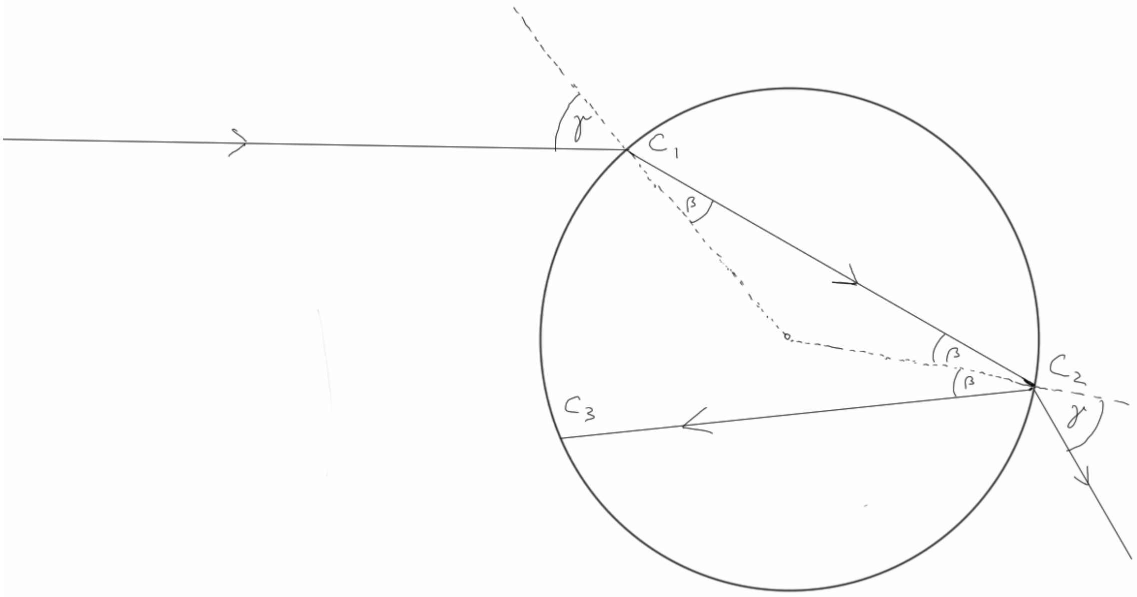
$$\lambda_1 = -\arcsin y + \arcsin\left(\frac{1}{n_w}y\right), \qquad -1 < y < 1$$

as formula for the directional angle of the first ray in the liquid.

We come to the second contact point $C_2$.



We discover an isosceles triangle $\triangle MC_1C_2$ and get the angle $\beta$ (which we know already) at the vertex $C_2$. We extend the above picture:



We apply Snell's Law at $C_2$ and get the angle $\gamma$ again. At $C_2$, a refracted ray exits from the droplet and goes into the air with a directional angle

$$\alpha_2 = \lambda_1 + \beta - \gamma.$$

With $\lambda_1$ from above and $\gamma = \arcsin y$, $\beta = \arcsin(y/n_w)$ we then get

$$\alpha_2 = 2\lambda_1 = 2\arcsin\left(\frac{1}{n_w}y\right) - 2\arcsin y.$$

**DIY:** *Check that this is also the correct formula when $-1 < y < 0$ (our picture only considers the case $y > 0$, and $\beta$ becomes negative for $y < 0$).*

The second ray in the liquid runs from $C_2$ to $C_3$ and has directional angle

$$\begin{aligned}\lambda_2 &= \lambda_1 + \pi + 2\beta\\ &= \left(\arcsin\left(\frac{1}{n_w}y\right) - \arcsin y\right) + \pi + 2\arcsin\left(\frac{1}{n_w}y\right)\\ &= 3\arcsin\left(\frac{1}{n_w}y\right) - \arcsin y + \pi\end{aligned}$$

which is again valid for positive $y$ as well as negative $y$.

And at $C_3$, a part of the light leaves the droplet by means of refraction, and that light-ray travelling through air has directional angle

$$\begin{aligned}\alpha_3 &= \lambda_2 + \beta - \gamma\\ &= 4\arcsin\left(\frac{1}{n_w}y\right) - 2\arcsin y + \pi, \qquad -1 < y < 1.\end{aligned}$$

**DIY:** *Repeat these considerations and prove that*

$$\alpha_4 = 6\arcsin\left(\frac{1}{n_w}y\right) - 2\arcsin y, \qquad -1 < y < 1.$$

You will come back to $\alpha_4$ when you discuss secondary rainbows and Alexander's band yourselves.

We should now plot those functions that give us the directional angles as functions of $y$, where we only consider the rays that go through air, and we choose $n_w = 1.333$, see Figure 3.2. For this I have chosen the programming language `python` because it is very powerful, it can do a lot of scientific mathematics, it is easy to learn, it costs nothing.... There are many tutorials on the internet for self-learning `python`.

The `python` code of the program called `rainbow.py` is here:

```
import numpy as NP                               # numpy contains many mathematical methods
import matplotlib.pyplot as PLT                  # this is some software package for mathematical plots

nw = 1.333
xvals = NP.arange(-1.0, 1.0, 0.005)              # this produces a long list of x values
yvals1 = 180.0 * (NP.pi - 2 * NP.arcsin(xvals)) / NP.pi
yvals2 = 180.0 * 2 * (NP.arcsin(xvals / nw) - NP.arcsin(xvals)) / NP.pi
yvals3 = 180.0 * (4 * NP.arcsin(xvals / nw) - 2 * NP.arcsin(xvals) + NP.pi) / NP.pi
yvals4 = 180.0 * (6 * NP.arcsin(xvals / nw) - 2 * NP.arcsin(xvals)) / NP.pi

PLT.figure(1)
PLT.subplot(221)                                 # in a 2x2 table, build the first plot
PLT.plot(xvals, yvals1)
PLT.xlabel(’y’)
PLT.ylabel(’alpha_1’)

PLT.subplot(222)
PLT.plot(xvals, yvals2)
PLT.xlabel(’y’)
PLT.ylabel(’alpha_2’)

PLT.subplot(223)
PLT.plot(xvals, yvals3)
PLT.xlabel(’y’)
PLT.ylabel(’alpha_3’)

PLT.subplot(224)
PLT.plot(xvals, yvals4)
PLT.xlabel(’y’)
PLT.ylabel(’alpha_4’)


PLT.show()                                       # now show us everything
```

and when we run it on a Linux command-line as “python rainbow.py”, then the output is this one:

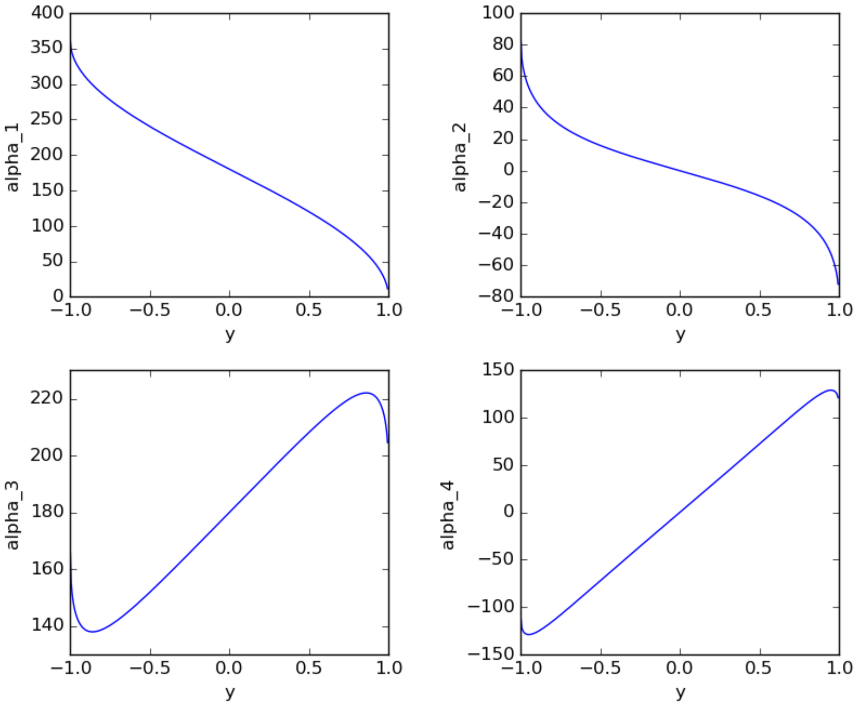


Figure 3.2: The angles $\alpha_1$, $\alpha_2$, $\alpha_3$, $\alpha_4$, converted into degrees.

Now we do a computer simulation: we split the interval $(-1, 1)$ for the variable $y$ into 20000 parts of equal length, each having a length $10^{-4}$. This gives 20000 values of $y$, all uniformly distributed over the interval. For each such value of $y$, we send one ray of light to the droplet. This gives us 20000 light-rays, and we record under which directional angle they exit the droplet at $C_1$, at $C_2$, at $C_3$, at $C_4$. We convert the angles into degrees (which is more familiar to us), and and we split the range of angles into 120 bins having equal widths of at most 3°. For instance in the case of $\alpha_2$, the relevant interval $[-80°, 80°]$ is being split into 120 equal parts. Each of the 20000 rays hits one bin, and we record how often each bin is being hit. Now look at Figure 3.3, whose vertical axes tell us how often each bin is being hit.

We would expect that each bin is being hit by approximately $\frac{20000}{120} = 167$ light rays, but this happens only for $\alpha_1$, and definitely not for $\alpha_3$. We observe for $\alpha_3$ (but not for $\alpha_1$ and $\alpha_2$) that some bins are being hit much more often than most other bins. They correspond to angles $\approx \pm 222°$, and a lot of light arrives at our eyes under these directional angles. See Figure 3.3, and the `python` code is on the next page.

These angles are the values of $\alpha_3$ for those $y$ where an extremal value is being attained. However, the extremality is not necessary here. All those values of $y$ are interesting where the derivatives $\alpha_3'(y)$ become zero. These values of $y$ are $y \approx \pm 0.87$ (see Figure 3.2). It is just a nice coincidence that these values are locations of maxima/minima, but this is not of bigger importance. An inflection point with horizontal tangent (such as $x = 0$ for the function $f(x) = x^3$) would do the same.

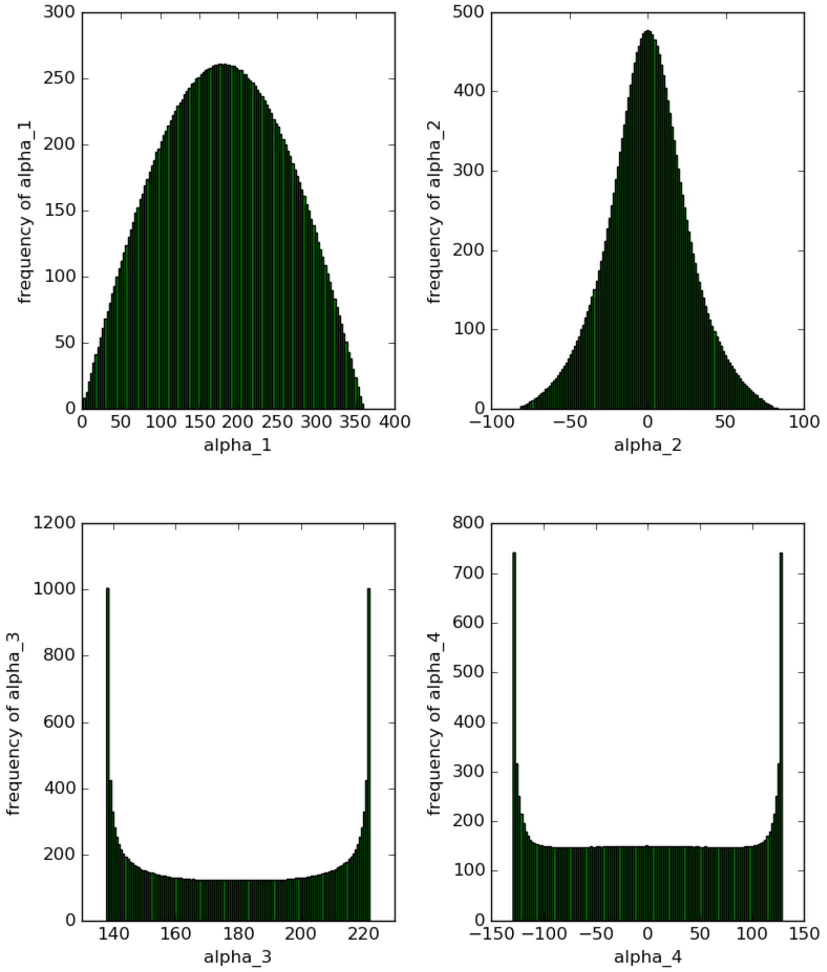
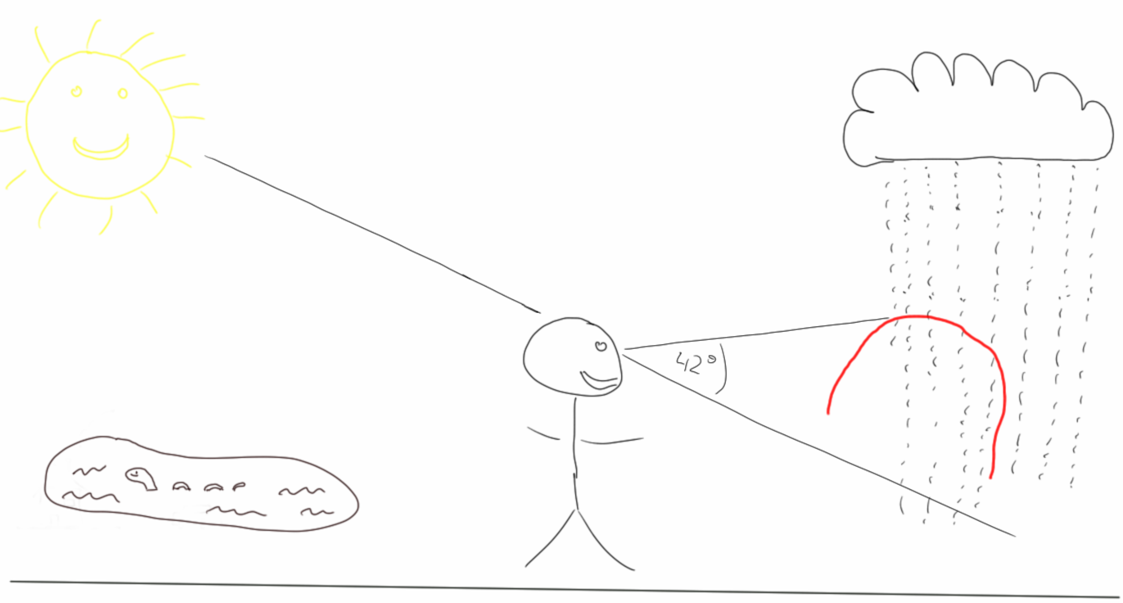


Figure 3.3: The histograms for the angles $\alpha_1$, $\alpha_2$, $\alpha_3$, $\alpha_4$, converted into degrees.

We have a first result: if the sun is in our back and we look into the rain, then we see the rainbow under an angle of 42° (since 222° − 180° = 42°).



The two sides of this 42° angle are

- the extension of the line from the sun towards us
- the line from us towards the rainbow.

Here comes the promised `python` code that generated the histograms in Figure 3.3.

```
import numpy as NP
import matplotlib.pyplot as PLT

nw = 1.333
xvals = NP.arange(-1.0, 1.0, 0.0001)
yvals1 = 180.0 * (NP.pi - 2 * NP.arcsin(xvals)) / NP.pi
yvals2 = 180.0 * 2 * (NP.arcsin(xvals / nw) - NP.arcsin(xvals)) / NP.pi
yvals3 = 180.0 * (4 * NP.arcsin(xvals / nw) - 2 * NP.arcsin(xvals) + NP.pi) / NP.pi
yvals4 = 180.0 * (6 * NP.arcsin(xvals / nw) - 2 * NP.arcsin(xvals)) / NP.pi

PLT.figure(1)
PLT.subplot(221)
PLT.hist(yvals1,bins = 120, color='green')
PLT.xlabel('alpha_1')
PLT.ylabel('frequency of alpha_1')

PLT.subplot(222)
PLT.hist(yvals2,bins = 120, color='green')
PLT.xlabel('alpha_2')
PLT.ylabel('frequency of alpha_2')

PLT.subplot(223)
PLT.hist(yvals3, bins = 120, color='green')
PLT.xlabel('alpha_3')
PLT.ylabel('frequency of alpha_3')

PLT.subplot(224)
PLT.hist(yvals4,bins = 120, color='green')
PLT.xlabel('alpha_4')
PLT.ylabel('frequency of alpha_4')

PLT.show()
```

## 3.4 Where do the Colours Come From ?

Short answer: the refractive index of water depends on the colour. The angle, under which we see that colour equals a critical value of $\alpha_3(y)$, minus $\pi$. Now what does this mean ?

**Definition 3.1.** *Let $f = f(y)$ be a differentiable function in $\mathbb{R}$. We say that $F$ is a* critical value of $f$ *if some real number $y_*$ exists with*

$$F = f(y_*), \quad \textit{and} \quad f'(y_*) = 0.$$

**Example 3.2.** *The function $f(y) = \sin(y)$ has critical values $F_1 = +1$ and $F_2 = -1$ because $+1 = \sin(\pi/2)$ and $\sin'(\pi/2) = \cos(\pi/2) = 0$. Similarly for $-1$.*

*The functions $\ln(y)$, $\exp(y)$ and $\tan(y)$ don't have any critical values.*

Recall that

$$\alpha_3 = 4 \arcsin\left(\frac{1}{n_w} y\right) - 2 \arcsin y + \pi,$$

hence we search for critical values of

$$f_3(y) = 4 \arcsin\left(\frac{1}{n_w} y\right) - 2 \arcsin y, \qquad -1 < y < 1,$$

and the refractive index $n_w$ depends on the colour of the light. This means we try to solve[4] $f_3'(y) \overset{!}{=} 0$ for the variable $y$. We recall that $\arcsin'(s) = \frac{1}{\sqrt{1-s^2}}$ for $-1 < s < 1$, and consequently the chain rule then gives us

$$f_3'(y) = 4 \cdot \frac{1}{\sqrt{1 - y^2/n_w^2}} \cdot \frac{1}{n_w} - \frac{2}{\sqrt{1-y^2}}.$$

We set this equal to zero, solve it for $y$, which has a solution $y_*$, and then we have to evaluate $f_3(y_*)$.

This is certainly doable, and you are heartily invited to pursue this plan.

Another approach is to calculate numerically. We write a little `python` program (which you can find in Appendix A.1) and the output of this program tells us the critical values of $f_3$:

```
The angle of the red part is  42.5163817211
The angle of the orange part is  41.932910149
The angle of the yellow part is  41.7881567797
The angle of the green part is  41.3565406454
The angle of the blue part is  40.9288507308
The angle of the violet part is  39.8071607468
```

We compare with the photo in Figure 3.4 and observe there that indeed red is at the outer boundary of the rainbow, and blue is at the inner boundary of the rainbow.

## 3.5 Now it is Your Turn

Explain the secondary rainbow (visible in the photo) using the function $\alpha_4$.

Explain why its colours are ordered reversely.

Explain why the region between the primary and the secondary rainbows is so dark (Alexander's band).

[4] The exclamation mark over the equality sign shall mean "I want this left-hand side to be zero".

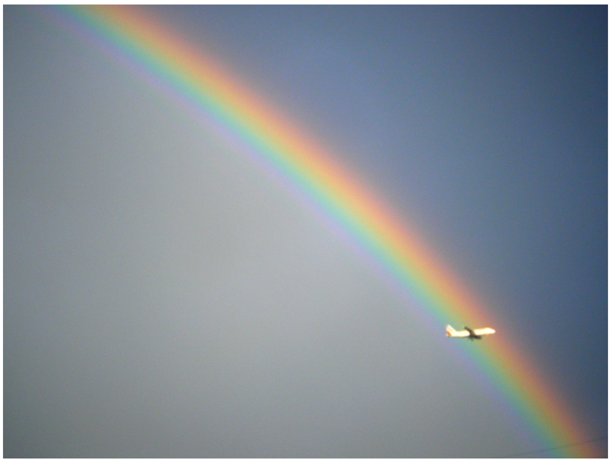


Figure 3.4: A primary rainbow, with Alexander's band in the top right corner. Violet is at the inner boundary, then comes blue, green, yellow, orange. And red is at the outer boundary. Photo (public domain) by Elodie Marnot.

# Chapter 4

# Mathematics in Sports

The material in this chapter is taken from [5] and [3].

## 4.1 How to Win a Shot Put

It is clear that a bigger start velocity (measured in metres per second) means a longer distance. The dependence on the angle is more unclear.

**Refined question:** given a starting velocity, how to choose the angle such that the distance is maximal ?

We have the modelling assumptions

- we ignore the air resistance
- hence we may regards the ball as a point
- the ball is flying in 2D, not 3D
- the starting point has height zero, and the ground is even.

We list the **known data**, and we also record their *unit* and their *type* (every mathematical object has a *type* that restricts which mathematical operations you can perform upon that object):

- the mass $m$ of the ball, with unit kilogram, and the type is a positive real number.
- the gravitational acceleration $g$ of the earth, with value (and unit) $g = 9.81\frac{m}{s^2}$, the type is a positive real number,
- the scalar value $v_{\text{init}}$ of the initial velocity, with unit metres per second, and the type being a positive real number.

We list the **unknown data**:

- the angle $\alpha$ between the initial velocity $\vec{v}_0$ and the ground, such that the ball flies as far as possible. This angle $\alpha$ has no unit and is a real number between 0 and $\pi/2$.
- the initial velocity $\vec{v}_0$ has the type of a vector in 2D, whose length is equal to $v_{\text{init}}$, and the unit is metres per second.

Further **unknown data** which seem to be relevant are

- the duration $T$ of the flight, with unit seconds, and type being a positive real number
- the maximal height $H$ of the flight, with unit metres and type a positive real number
- the length $L$ of the shot in metres, a positive real number.

**survival hint 1:** always choose the same unit for lengths (for instance metres, but do *not* mix metres with inches and centimetres). Choose always the same unit for times (seconds). Then the other units for velocity and acceleration follow automatically as $\frac{m}{s}$ and $\frac{m}{s^2}$.

**survival hint 2:** always distinguish between vectors and numbers; these are different types. The admissible operations that you can perform upon them are different. For instance, you can often divide something by a number, but you cannot never divide something by a vector.

Now we need some **input from physics**. The ball is making two motions simultaneously: one motion is "rising up, then falling down again" in the vertical direction. The other motion is "from left to right" horizontally. Both motions do not interact with each other. The only external force acting upon the ball is the gravity, which pulls downward and has magnitude $m \cdot g$, with $g$ being our gravity constant. Recall that we ignore air resistance (which seems reasonable for the shot put; badminton would be much harder).

You will learn more details of this elementary physics in the course *Applied Mathematics A* in the third semester. More advanced physics will then be taught in the course *Applied Mathematics B.*

We need more identifiers:

- the time variable is called $t$, with units seconds, and its type is a real valued variable. The expression "time $t$" means a point on the time axis, it *does not mean* a time duration of length $t$.
- the horizontal coordinate of the ball position at time $t$ is $x(t)$, with unit metres. The type of this $x(t)$ is a *function depending on* $t$ *and having values in* $\mathbb{R}$. Moreover, this function $x(t)$ is unknown.

**survival hint 3:** always distinguish between numbers and functions; these are different types. The admissible operations that you can perform upon them are different. For instance, you can often take the derivative of a function, but taking a derivative of a number is silly.

What do we know about the function $x = x(t)$ ? At time zero, the ball is in the start position, hence $x(0) = 0$. At time $T$, the ball has just landed, so $x(T) = L$. There is no acceleration in horizontal direction, which means $x''(t) = 0$ for all times $t$. Which means that the first derivative $x'(t)$ is a constant function, and its meaning is the horizontal velocity. We obtain (at the time $t = 0$) the horizontal velocity by splitting the vectorial velocity $\vec{v}_0$ into its horizontal component and its vertical component (you are encouraged to make a picture yourselves) and the result then is $x'(0) = v_{\text{init}} \cdot \cos(\alpha)$. Let us call this $v_{0x}$, hence $v_{0x} := v_{\text{init}} \cdot \cos(\alpha)$ which is unknown because $\alpha$ is unknown. Since $x'(t)$ is constant, we have $x'(t) = v_{0x}$ for all times $t$, and therefore $x(t) = v_{0x} \cdot t$ for all $t$.

Next we consider the vertical component of the motion of the ball. We need one more identifier:

- the vertical coordinate of the ball position at time $t$ is $y(t)$, with unit metres. It means the height above ground. The type of this $y(t)$ is a *function depending on* $t$ *and having values in* $\mathbb{R}$. Moreover, this function $y(t)$ is unknown.

What do we know about the function $y = y(t)$ ? At time zero, the ball is in the start position (height zero), hence $y(0) = 0$. At time $T$, the ball has just landed, so $y(T) = 0$. The gravitational acceleration is pulling downwards, which means $y''(t) = -g$, and the minus sign is explained because the function $y$ measures the height upwards, but the gravity is pulling downwards. And in the moment when we throw the ball (which means $t = 0$), its vertical velocity equals the vertical component of the initial vectorial velocity $\vec{v}_0$. Our DIY picture tells us that this vertical component of the initial vectorial velocity is $v_{0y} = v_{\text{init}} \cdot \sin(\alpha)$. We summarize: the unknown function $y = y(t)$ has the following properties

$$y(0) = 0, \qquad y'(0) = v_{0y} = v_{\text{init}} \cdot \sin(\alpha), \qquad y''(t) = -g, \qquad y(T) = 0.$$

The equation $y''(t) = -g$ is called a *differential equation*, and in the following years you will learn a lot about such differential equations. It is a differential equation of second order, because two derivatives of the unknown function $y$ are present. And we have two initial conditions for this differential equation, namely $y(0) = 0$ and $y'(0) = v_{\text{init}} \cdot \sin(\alpha)$. We wish to solve this differential equation and its two initial conditions, and the solution will be the function $y = y(t)$. At school you also solved a lot of equations, and the solutions were numbers in most cases. Now at university, the solution to an equation may be a function (not a number).

You will learn in the following years that a second order differential equation with two initial conditions (almost) always possesses exactly one solution $y = y(t)$, which is this function:

$$y(t) = -\frac{g}{2} \cdot t^2 + v_{0y} \cdot t$$

**DIY:** *check that this function indeed solves the differential equation and the two initial conditions.*

What have we achieved so far ? We have determined functions $x = x(t)$ and $y = y(t)$, but not completely, because they contain unknown parameters, namely $v_{0x}$ and $v_{0y}$. The numbers $\alpha$, $L$, $T$, are still not known.

We have not yet used that $y(T) = 0$, so we substitute $T$ for $t$ into the equation for $y(t)$ and get

$$0 = y(T) = -\frac{g}{2} \cdot T^2 + v_{0y} \cdot T = T \cdot \left(-\frac{g}{2} \cdot T + v_{0y}\right).$$

We solve this for $T$ and get two solutions: $T = 0$ (which is not relevant) and

$$T = \frac{2v_{0y}}{g}.$$

What is our goal, actually ? We wish to find an $\alpha$ for which $L$ becomes maximal. So, what do we know about $L$, by the way ? We have $L = x(T) = v_{0x} \cdot T$, which brings us to

$$L = v_{0x} \cdot \frac{2v_{0y}}{g} = \frac{2}{g} \cdot v_{0x} \cdot v_{0y} = \frac{2}{g} \cdot v_{\text{init}} \cos(\alpha) \cdot v_{\text{init}} \sin(\alpha) = \frac{v_{\text{init}}^2}{g} \cdot \sin(2\alpha),$$

where we have used $2\sin(\alpha)\cos(\alpha) = \sin(2\alpha)$. Now we are done: on the right-hand side, everything is known except $\alpha$, and we recall that the sine function gets maximal at $\pi/2$, hence we get

$$L_{\text{max}} = \frac{v_{\text{init}}^2}{g},$$

and this maximal value is attained at an angle $\alpha = \frac{\pi}{4}$ which corresponds to 45°.

We come to **Step 4** of the modelling cycle (we walked the previous three steps without announcing them explicitly) and interpret the obtained result:

- the maximal height $H$ was never needed, this should be dropped in a final write-up
- the optimal shot length $L$ is $L = \frac{v_{\text{init}}^2}{g}$ provided the starting height is zero and the initial angle is $\frac{\pi}{4}$
- for later use (in the part about how to add performances in the decathlon) we note that $L$ is proportional to the square of the initial velocity $v_{\text{init}}$.
- how realistic is the assumption of initial height zero ? From real life we guess that $y(0) \approx 1.7$ metres and $H \approx 3$ metres (or even less), and therefore the assumption $y(0) = 0$ seems a bit unrealistic.

Now we decide to be unhappy with this modelling assumption $y(0) = 0$, and instead we throw from an initial height $h_0$ which is known, has unit metres, and is a positive real number.

Now we go through the whole modelling cycle again, but we can re-use a lot of the knowledge from the previous attempt. The results are:

$$\begin{aligned}
&x(0) = 0, \qquad x(T) = L, \qquad x'(0) = v_{0x} = v_{\text{init}} \cdot \cos(\alpha),\\
&x''(t) = 0 \quad \text{for all times } t,\\
&y(0) = h_0, \qquad y(T) = 0, \qquad y'(0) = v_{0y} = v_{\text{init}} \cdot \sin(\alpha),\\
&y''(t) = -g \quad \text{for all times } t.
\end{aligned}$$

Solving the second-order differential equation $x''(t) = 0$ together with its two initial conditions $x(0) = 0$ and $x'(0) = v_{0x}$ then gives

$$x(t) = v_{0x} \cdot t.$$

**DIY:** *check that this function is indeed a solution to the differential equation $x''(t) = 0$ and its two initial conditions.*

We substitute $T$ for $t$ and get $L = v_{0x} \cdot T$.

Solving[1] the second-order differential equation $y''(t) = -g$ together with its two initial conditions $y(0) = h_0$ and $y'(0) = v_{0y}$ then gives

$$y(t) = -\frac{g}{2} \cdot t^2 + v_{0y} \cdot t + h_0.$$

**DIY:** *check that this function is indeed a solution to the differential equation $y''(t) = -g$ and its two initial conditions.*

We substitute $T$ for $t$ and get the equation

$$0 = y(T) = -\frac{g}{2} \cdot T^2 + v_{0y} \cdot T + h_0,$$

and solve this for $T$, which gives two solutions

$$T_\pm := \frac{v_{0y}}{g} \pm \sqrt{\frac{v_{0y}^2}{g^2} + \frac{2h_0}{g}},$$

where only the plus sign is relevant.

**DIY:** *what is the physical meaning of the solution $T_-$ which we ignore in what follows ?*

Now we have found $T = T_+$, and the shot length $L$ then is

$$L = v_{0x} \cdot T = v_{0x} \cdot \left( \frac{v_{0y}}{g} + \sqrt{\frac{v_{0y}^2}{g^2} + \frac{2h_0}{g}} \right).$$

Now we bring $\alpha$ into play via $v_{0x} = v_{\text{init}} \cdot \cos(\alpha)$ and $v_{0y} = v_{\text{init}} \cdot \sin(\alpha)$, and dragging some constant factors out of the root then gives

$$L = \frac{v_{\text{init}}^2}{g} \cdot \left( \cos(\alpha) \cdot \sin(\alpha) + \cos(\alpha) \cdot \sqrt{\sin^2(\alpha) + \frac{2h_0 g}{v_{\text{init}}^2}} \right).$$

The only unknown variable on the right-hand side (RHS) is $\alpha$. Everything else is known. We want to find all $\alpha$ for which $L$ becomes maximal. Our expectation from real life is that there is only one optimal $\alpha$.

Now we have a problem: to maximise the function $L = L(\alpha)$, we have learned in school to take the derivative $L'(\alpha) = \frac{\mathrm{d}L}{\mathrm{d}\alpha}(\alpha)$ and then set this derivative equal to zero, and then to solve the equation $L'(\alpha) \overset{!}{=} 0$ for the variable $\alpha$. But the function $L(\alpha)$ is already quite complicated, and then its derivative $L'(\alpha)$ will be an even bigger mess. Solving the equation $L'(\alpha) = 0$ for $\alpha$ is then quite a challenge. The approach which you learned at school turns out to be mostly useless.

Now let's be realistic. Do we need the derivative $L'(\alpha)$ at all, and is it really necessary to determine the optimal $\alpha$ up to 8 decimal digits ? The athlete has certainly other things to worry about than to reproduce an angle up to such a ridiculous accuracy. Therefore we do some quick numerical calculation using `python`. The program is below (for $v_{\text{init}} = 16.0$ and $h_0 = 1.8$), and I called it `shotput.py`.

```
import scipy as SP             # scipy contains many mathematical methods

gravity = 9.81

vinit = 16.0                   # initial velocity
hzero = 1.8                    # initial height

prefactor = vinit**2 / gravity

heightterm = 2 * hzero * gravity / (vinit**2)

alphalist = SP.arange(35.0, 55.0, 0.5)

for alpha in alphalist:
   a = alpha * 2 * SP.pi / 360
   shotlength = prefactor * SP.cos(a) * (SP.sin(a) + SP.sqrt((SP.sin(a))**2 + heightterm))
   print "alpha = ", alpha, " and shotlength = ", shotlength
```

[1] I am just telling you that this function $y = y(t)$ *is* the solution. Late courses will tell you how to find it.

Then I run it on a Linux command-line as “python shotput.py”, and the output is this table:

```
alpha =  35.0  and shotlength =  26.8682405323
alpha =  35.5  and shotlength =  26.981761224
alpha =  36.0  and shotlength =  27.0884860548
alpha =  36.5  and shotlength =  27.1883483685
alpha =  37.0  and shotlength =  27.2812846387
alpha =  37.5  and shotlength =  27.3672344475
alpha =  38.0  and shotlength =  27.4461404651
alpha =  38.5  and shotlength =  27.5179484299
alpha =  39.0  and shotlength =  27.5826071292
alpha =  39.5  and shotlength =  27.6400683805
alpha =  40.0  and shotlength =  27.6902870132
alpha =  40.5  and shotlength =  27.7332208505
alpha =  41.0  and shotlength =  27.7688306924
alpha =  41.5  and shotlength =  27.7970802981
alpha =  42.0  and shotlength =  27.8179363692
alpha =  42.5  and shotlength =  27.8313685335
alpha =  43.0  and shotlength =  27.8373493282
alpha =  43.5  and shotlength =  27.8358541838
alpha =  44.0  and shotlength =  27.8268614083
alpha =  44.5  and shotlength =  27.8103521705
alpha =  45.0  and shotlength =  27.7863104851
alpha =  45.5  and shotlength =  27.754723196
alpha =  46.0  and shotlength =  27.7155799605
alpha =  46.5  and shotlength =  27.6688732339
alpha =  47.0  and shotlength =  27.6145982526
alpha =  47.5  and shotlength =  27.5527530189
alpha =  48.0  and shotlength =  27.483338284
alpha =  48.5  and shotlength =  27.4063575322
alpha =  49.0  and shotlength =  27.3218169638
alpha =  49.5  and shotlength =  27.2297254785
alpha =  50.0  and shotlength =  27.1300946588
alpha =  50.5  and shotlength =  27.022938752
alpha =  51.0  and shotlength =  26.9082746532
alpha =  51.5  and shotlength =  26.7861218874
alpha =  52.0  and shotlength =  26.6565025912
alpha =  52.5  and shotlength =  26.5194414948
alpha =  53.0  and shotlength =  26.3749659026
alpha =  53.5  and shotlength =  26.2231056746
alpha =  54.0  and shotlength =  26.0638932068
alpha =  54.5  and shotlength =  25.8973634112
```

and we observe that the optimal angle is 43.0°, with a shot put length of 27.84 metres. If you deviate from the optimal angle by 0.5°, the thrown shot put length changes by less than 1 centimetre, which we consider negligible.

This was an example in “poor man’s numerics”, and you will learn much more sophisticated numerical methods in the various courses called *Numerical Analysis*.

## 4.2 How to Add up Points in Decathlon and Heptathlon

Let us recall some history: one of the earliest combined event competitions were the 1851 Much Wenlock Olympics in England, which was a men’s pentathlon with the events high jump, long jump, putting the 36lbs shot, running 880 yards, and climbing the 55ft rope. Then various other combined event competitions were performed over the years.

The modern versions are

**decathlon men:** 100 m, long jump, shot put, high jump, 400 m, 110 m hurdles, discuss throw, pole vault, javelin, 1500. This discipline is olympic since 1912.

**heptathlon women:** 100 m hurdles, high jump, shot put, 200 m, long jump, javelin, 800 m. This discipline became olympic very late — 1984. Before there were only olympic pentathlons whose disciplines changed heavily over the years.

**Questions:**

- How to determine a winner in such a competition ?
- How can we define “world record” ?
- What is fairness ?

Let us work scientifically and get our terminology clear and the thinking straight:

- We call a competition *poly-athlon*, generalizing pentathlon, heptathlon, decathlon.
- Each poly-athlon contains $E$ events, and $E$ is a given natural number, with $E = 5$ for the pentathlon, $E = 7$ for the heptathlon, $E = 10$ for the decathlon.

- In each competition (for instance the 2012 Olympics) we have $N$ participants, and each of them is doing all $E$ events.
- For each participant and each event, a performance is determined like this:

  **track events:** run time in seconds (smaller is better)

  **jump events:** length in metres (bigger is better)

  **throw events:** length in metres (bigger is better)
- This gives $E \cdot N$ non-negative real numbers, and we need a way of figuring out who of the $N$ participants has rank one, who has rank two, $\ldots$, who has rank $N$. This decision has to be based on these $EN$ numbers only.
- We want "fairness" (this has to be made more precise in mathematical terms), on various levels:

  **fairness inside each event:** if Alan and Bob participated in the same competition, and Alan is better than Bob in exactly one event, and in all the other events they are exactly equal, then Alan *must not* be ranked worse than Bob.

  **fairness between different events of the same competition:** the specialists in the jumping events should not be advantaged against the specialists in the running events when we do the ranking at the level of one specific poly-athlon.

  **fairness between competitions:** otherwise there is no way of determining the world record holder.

  Other aspects of "fairness" will come later.

The political wish of the officials is that only *well-rounded* athletes should win (meaning quite strong in all events). A person who is excellent in one event, but less than mediocre the others should not win.

We now present various methods that have been and are actually used, discuss their merits, and we learn a bit of elementary calculus along the way.

**Method 1**

A first approach was this one: each participant gets in a certain event as many points as the rank of that participant in that specific event. After all the events have happened, we form the sum for each participant over his points. The winner is then that person with the lowest point sum (it works similar to Formula 1, but with high and low numbers swapped). This calculation was actually used in various poly-athlons until about 1880, and in the olympic pentathlon of the men till 1924.

Discussion: we have obvious fairness in each event, but not between competitions. The winner of the Edinburgh City Council School Children Decathlon would have a better rank than the Bronze winner in the London 2012 Olympics.

This first approach is also a bit silly because it is advantageous to win by a small margin in comparison with a winning by a large margin, in order to save energy for the following events.

**Method 2**

In 1884 the following calculation was developed in the US and then employed in the olympic decathlon throughout: If a participant repeats the world record of the previous year in a specific event, then that participant gets 1000 points for that event. If a participant shows the performance of a schoolboy, then he gets one point (women were not admitted in 1884). For performances in-between, appropriate points are awarded by linear interpolation. For performances worse than a schoolboy, zero points are given. And for performances better than the world record, appropriately more points than 1000 will be awarded.

**Remark 4.1.** *Here we see an important mathematical concept: normalization. The calculation with percentages is another example of normalization. Compare also Remark 6.8.*

All the following methods of adding points in poly-athlon competitions are variations of this method, so we express it in mathematical terms:

We have $N$ participants with names $1, 2, \ldots, N$.

We have $E$ events with names $1, 2, \ldots, E$.

We want to determine functions $f_1$, $f_2$, …, $f_E$ that translate from a performance value $p$ (measured in metres for the jump and throw events, and measured in seconds for the track events) into a number of points.

In mathematics it is very helpful to give names for each relevant quantity: the total number of points for the participants 1, 2, …, $N$ shall be called $T_1$, $T_2$, …, $T_N$.

The performance of participant $i$ at event $k$ shall be $p_{ik}$, which is a real number with a unit (metres or seconds). Here $1 \le i \le N$ and $1 \le k \le E$. Then the total number of points is

$$\forall i \in \{1, 2, \dots, N\}\colon \qquad T_i := \sum_{k=1}^{E} f_k(p_{ik}).$$

We write := because each $T_i$ is being *defined* by the RHS. We need to find functions $f_k$ where $k \in \{1, 2, \dots, E\}$.

Method 2 brings the following restrictions:

$$f_k(p) = \begin{cases} 0 & : p \text{ is worse than an schoolboy performance at event } k \\ 1 & : p \text{ is just a schoolboy performance at event } k \\ 1000 & : p \text{ repeats the world record of the previous year at event } k \end{cases}$$

And in case of Method 2, this function $f_k$ is piecewise linear with exactly one "knee".

Now let us ignore for a moment this Method 2 and try to see the bigger picture, and ask for relevant properties that our functions $f_k$ shall possess in order to get *fairness.*

**if event $k$ is a jump event or a throw event,** then the function $f_k$ shall be monotonously increasing because bigger performance numbers should give more points.

**if event $k$ is a run event,** then the function $f_k$ shall be monotonously decreasing because lower performance numbers should give more points.

Assuming that all functions $f_k$ possess a derivative, we can express this requirement of monotonicity as

$$\begin{aligned} &\forall x \in \mathbb{R}_{>0}\colon \quad f_k'(p) > 0 \quad \text{if event } k \text{ is a jump/throw event} \\ &\forall x \in \mathbb{R}_{>0}\colon \quad f_k'(p) < 0 \quad \text{if event } k \text{ is a run event} \end{aligned}$$

These monotonicity requirements are necessary conditions for *fairness inside each event.*

We wish a bit more fairness however. Certainly there is some limit to what a human body can do. We don't expect ever a human to jump 20 metres in the long jump, for instance. Suppose that a long jumper is quite near to that boundary of what a human body can do, and he improves his performance by 5 centimetres. Then his points in the long jump event (which is the second event in the official decathlon order) will improve.

And there is another long jumper who jumps considerably worse than the first jumper. But also the second long jumper improves his performance by 5 centimetres. His points will improve, *but they should improve not stronger than the points of the first jumper*, because otherwise we would give a bonus to weaker athletes which looks silly.

Let's do this with numbers: in the 1912 olympics, the decathlon winner jumped 6.87 metres, and the last participant jumped 5.43 metres (just for a comparison). Hence we have the condition

$$f_2(6.65) - f_2(6.60) \overset{!}{\ge} f_2(5.95) - f_2(5.90).$$

On the LHS we have the point increment of a strong jumper when he jumps 5 centimetres farther, and on the RHS we have the point increment of a weaker jumper when he jumps 5 centimetres farther. The exclamation mark over the relation symbol shall mean "we want the LHS to be not smaller than the RHS". Let us divide by 0.05 which is positive:

$$\frac{f_2(6.65) - f_2(6.60)}{0.05} \overset{!}{\ge} \frac{f_2(5.95) - f_2(5.90)}{0.05}.$$

Recall that if $f$ is a differentiable function and $\varepsilon$ is a very small positive number, then $f'(x) \approx \frac{f(x+\varepsilon)-f(x)}{\varepsilon}$. Hence the following requirement seems reasonable

$$f_2'(6.60) \overset{!}{\geq} f_2'(5.90).$$

We could choose other numbers than 6.60 and 5.90, and do a similar reasoning. The result then will be that we wish the derivative $f_2'$ to not be falling, which is expressed as

$$\forall x \in \mathbb{R}_{>0}\colon \quad \left(f_2'\right)'(x) \overset{!}{\geq} 0, \qquad \text{or} \quad f_2''(x) \geq 0.$$

**DIY:** *Show that the same condition holds also for the run events.*

*The trouble here is that $f_k$ is falling in the run events (in contrast to $f_2$ for the long jump, which is rising).*

We summarize: the functions $f_k$ which we have to find shall satisfy the inequalities

$$\begin{aligned}
&\forall x \in \mathbb{R}_{>0}\colon f_k'(x) > 0 \quad \text{if } k \text{ corresponds to a jump/throw event,}\\
&\forall x \in \mathbb{R}_{>0}\colon f_k'(x) < 0 \quad \text{if } k \text{ corresponds to a run event,}\\
&\forall x \in \mathbb{R}_{>0}\colon f_k''(x) \geq 0.
\end{aligned}$$

Functions $f$ with $f'' \geq 0$ everywhere are called *convex functions*.

Now let us discuss the functions $f_k$ as they are specified in Method 2. Ignoring the “knee” in their graphs at the schoolboy performance, where the $f_k$ don't have a first order derivative (let alone a second order derivative), these inequalities are indeed satisfied, which is good. And also the rule “world record performance equals 1000 points” is an objective criterion and contributes to fairness between different events. On the other hand, the rule “schoolboy performance equals 1 point” is quite subjective (and would lead to heated discussions in committee meetings).

**DIY:** *take the numbers from the 1912 olympics:*

**100 m:** *11.0 seconds give 952.4 points, and 13.3 seconds give 405.0 points*

**javelin:** *50.40 metres give 878.175 points, and 32.32 metres give 380.975 points.*

*The functions $f_k$ were chosen as piecewise affine linear, as explained above.*

- *determine the formulas for these two functions $f_k$*
- *explain why specialists in throwing or jumping have an unfair advantage over the specialists in running.*

### Method 3

The IAAF[2] used a third kind of scoring tables from 1934 till approximately 1952: For each event, let experts decide by “experienced judgement” what is a performance level A (best), performance level B, …, performance level G. Then split all the levels into twenty sub-intervals. In each sub-interval, twiddle with parameters $a$ and $b$ of an appropriately chosen exponential function of the type $y(x) = \exp(ax + b)$ where $x$ is the distance (in field events) or the velocity (in track events).

### Method 4

Over the years, the performances of the athletes got better, training was being improved, the jumping techniques changed, so new scoring tables were approved by the IAAF in 1952. In the subsequent years, athletes became severely unhappy with these new scoring tables because the functions $f_k$ have become “much too convex” (or, in the words of the IAAF: “the scale has become way too progressive”), favouring specialised athletes over well-rounded athletes.

### Method 5

The IAAF approved new tables in 1962 based on the following principles:

[2] International Association of Athletics Federations

- determine world leading performance in each event (1000 points),
- determine a zero point performance,
- for the run events: calculate the average velocity as $\frac{\text{distance}}{\text{runtime}}$, and then perform a linear interpolation in a diagram whose horizontal axis is the *velocity* and whose vertical axis tells the awarded points,
- for the throw/jump events: deduce the velocity from the distance by taking the square root of the distance and then multiplying by a conversion factor. Have a look at page 73 for an explanation where the square root comes from. Then perform a linear interpolation in a diagram whose horizontal axis is the velocity and whose vertical axis tells the awarded points.

Let us discuss this method: it looks fair, also between the different events in a competition, because (in distinction to Method 2), now all diagrams are always operating with the same units: points are now always calculated from the velocity (before, the points were calculated partly from the distance, partly from the runtime).

**DIY:** [3] *explain why the first-order derivatives* $f_k'$ *have the desired sign, but some second-order derivatives* $f_k''$ have not*. This means that some functions* $f_k$ *which should have been convex are in fact concave.*

Running experts liked it, but all the other athletes became severely unhappy with this method “because the scales are not at all progressive”.

**Method 6**

From 1985 onwards, the following formulas are valid for the men’s decathlon:

| | | |
|---|---|---|
| 100 m: | $f_1(x) = 25.437 \cdot (18.0 - x)^{1.81}$, | $x$ in seconds |
| long jump: | $f_2(x) = 0.14354 \cdot (x - 220)^{1.40}$, | $x$ in centimetres |
| shot put: | $f_3(x) = 51.39 \cdot (x - 1.5)^{1.05}$, | $x$ in metres |
| high jump: | $f_4(x) = 0.8465 \cdot (x - 75)^{1.42}$, | $x$ in centimetres |
| 400 m: | $f_5(x) = 1.53775 \cdot (82 - x)^{1.81}$, | $x$ in seconds |
| 110 m hurdles: | $f_6(x) = 5.74352 \cdot (28.5 - x)^{1.92}$, | $x$ in seconds |
| discus: | $f_7(x) = 12.91 \cdot (x - 4.0)^{1.10}$, | $x$ in metres |
| pole vault: | $f_8(x) = 0.2797 \cdot (x - 100)^{1.35}$, | $x$ in centimetres |
| javelin: | $f_9(x) = 10.14 \cdot (x - 7.0)^{1.08}$, | $x$ in metres |
| 1500 m: | $f_{10}(x) = 0.03768 \cdot (480 - x)^{1.85}$, | $x$ in seconds. |

And for the women’s heptathlon, the formulas are

| | | |
|---|---|---|
| 100 m hurdles: | $f_1(x) = 9.23076 \cdot (26.7 - x)^{1.835}$, | $x$ in seconds |
| high jump: | $f_2(x) = 1.84523 \cdot (x - 75)^{1.348}$, | $x$ in centimetres |
| shot put: | $f_3(x) = 56.0211 \cdot (x - 1.5)^{1.05}$, | $x$ in metres |
| 200 m: | $f_4(x) = 4.99087 \cdot (42.5 - x)^{1.81}$, | $x$ in seconds |
| long jump: | $f_5(x) = 0.188807 \cdot (x - 210)^{1.41}$, | $x$ in centimetres |
| javelin: | $f_6(x) = 15.9803 \cdot (x - 3.8)^{1.04}$, | $x$ in metres |
| 800 m: | $f_7(x) = 0.11193 \cdot (254 - x)^{1.88}$, | $x$ in seconds. |

The general structure is $f_k(x) = A \cdot (B - x)^C$ or $f_k(x) = A \cdot (x - B)^C$, depending on whether $f_k$ is desired to be falling or rising. Where do these funny numbers come from ?

The following has been done:

- collect a lot of statistical data of poly-athlon performances, over many years, over all levels of competitions,

[3] For didactical reasons, the author (purposefully) does not explain the unfairness of Method 5 in all details. Please figure it out on your own.

- twiddle with the parameters $A$, $B$, $C$ in each $f_k$ such that "well-rounded athletes" get approximately the same amount of points in each of the $E$ events. (Mathematical tools for this work will be developed in the *Statistics* courses; today we can't even clearly phrase the question).
- all functions $f_k$ should be of the correct monotonicity, and be convex (but not "too much convex" because well-rounded athletes should be favoured over specialized athletes)
- twiddle with the parameters in such a way that world-class athletes get approximately the same sum under the old system and under the new system.

## 4.3 How do World Records Evolve Over Time

Let us have a look at the official world records in the 100 m runs:

| 1912 | 1921 | 1930 | 1936 | 1956 | 1960 | 1968 | 1991 | 1994 | 1996 | 1999 | 2005 | 2008 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10.6 | 10.4 | 10.3 | 10.2 | 10.1 | 10.0 | 9.9 | 9.86 | 9.85 | 9.84 | 9.79 | 9.77 | 9.69 |

**Question:** what can we expect for 2050 ? For 2100 ? Can we find a function that approximates as good as possible the overall trend ?

In modelling, it is a reasonable approach to always start with an easy description, hence we attempt to describe the general trend as an affine linear function

$$y = f(x) = ax + b$$

with $a$ and $b$ to be determined.

More precisely: we have given data $x_1 = 1912$, $x_2 = 1921$, $\dots$, $x_{13} = 2008$ as well as $y_1 = 10.6$, $y_2 = 10.4$, $\dots$, $y_{13} = 9.69$. In a coordinate system, this gives us a collection of $n$ points with coordinates $(x_i, y_i)$, where $i \in \{1, 2 \dots, n\}$. Next we wish to find the "best possible" approximation of this point collection by a graph of an affine linear function $y = f(x) = ax + b$. Our tuning parameters are $a$ and $b$.

What is meant by "best possible" ? This is a matter of taste, in some sense. Suppose some parameters $a$ and $b$ have been found somehow. How much does the graph of the function $y = ax + b$ deviate from the point $(x_7, y_7)$ ? We could drop the perpendicular from $(x_7, y_7)$, calculate the length of that perpendicular; then do the same for all the other $n - 1$ points, and then attempt to make all these distances as small as we can. The calculations will then be quite involved, so it has become custom to simplify it: instead of the perpendicular distance of the point $(x_i, y_i)$, we take the vertical distance, which is $|ax_i + b - y_i|$. We end up with $n$ such vertical distances, and we wish that they are "all small", whatever that means. Now we invent some meanings:

**Task 1:** given $n$ real numbers $x_1 < x_2 < x_3 < \dots < x_n$ and $n$ real numbers $y_1, \dots, y_n$, determine two numbers $a$ and $b$ in such a way that the biggest of the $n$ numbers $|ax_1 + b - y_1|$, $|ax_2 + b - y_2|$, $\dots$, $|ax_n + b - y_n|$ becomes as small as possible.

**Task 2:** given $n$ real numbers $x_1 < x_2 < x_3 < \dots < x_n$ and $n$ real numbers $y_1, \dots, y_n$, determine two numbers $a$ and $b$ in such a way that the sum of $n$ numbers $|ax_1 + b - y_1| + \dots + |ax_n + b - y_n|$ becomes as small as possible.

**Task 3:** given $n$ real numbers $x_1 < x_2 < x_3 < \dots < x_n$ and $n$ real numbers $y_1, \dots, y_n$, determine two numbers $a$ and $b$ in such a way that the sum of squares of $n$ numbers $|ax_1+b-y_1|^2+\dots+|ax_n+b-y_n|^2$ becomes as small as possible.

Which of the three tasks is most meaningful ? This is a matter of taste and depends on the problem that we intend to solve. The mathematically hardest are Task 1 and Task 2, mainly for the reason that the absolute-value-function cannot be differentiated at its minimum, and that function which picks the largest number out of $n$ numbers also cannot be differentiated.

On the other hand, Task 3 has the advantage that big deviations $|ax_i + b - y_i|^2$ are being "punished" because the square function makes numbers bigger if they are bigger than one, and makes them smaller if they are smaller than one.

The method is hence called "method of least squares" and you will learn it in the lectures on *Statistics*.

**Mathematical method (least squares method)**

**Given:** $n$ real numbers $x_1 < \cdots < x_n$ and $n$ real numbers $y_1, \ldots, y_n$

**Wanted:** real numbers $a$ and $b$ such that $\sum_{i=1}^n (ax_i + b - y_i)^2$ becomes minimal

**Solution:** find those $a$ and $b$ that solve the following system of two equations for the two unknowns $a$ and $b$:

$$\begin{cases} a\sum_{i=1}^n x_i^2 + b\sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i \\ a\sum_{i=1}^n x_i + bn = \sum_{i=1}^n y_i. \end{cases}$$

**DIY:**

- *apply this method for the 100m world records for men*
- *make a prediction for 2025, 2050, 2010*
- *discuss the quality/reliability of your prediction*

This mathematical method raises several questions:

- What do these two equations actually mean ?
- Is that system of two equations for the two unknowns $a$ and $b$ actually solvable ?
- If it is solvable, how to solve it ?

Concerning the meaning of the system: the numbers $x_i$ and $y_i$ are fixed, and only the numbers $a$ and $b$ are variable. Hence we define a function $f$ with two variables $a$ and $b$ like this:

$$f(a,b) = \sum_{i=1}^n (ax_i + b - y_i)^2.$$

We wish to minimise this function. Imagine (for a moment) also $b$ as fixed, and only $a$ is variable. Then we may form the derivative of $f$ with respect to its only remaining variable $a$, and put this derivative equal to zero. The mathematical notation is $\frac{\partial f}{\partial a}(a,b) \overset{!}{=} 0$. It turns out that this is the first equation of the system. Next we set $b$ free again, and now $a$ is temporarily frozen. Then we compute the derivative of $f$ with respect to its only remaining variable $b$, and we put this derivative equal to zero, written as $\frac{\partial f}{\partial b}(a,b) \overset{!}{=} 0$, which is the second equation of the system. Our function $f$ has two variables ($a$ and $b$), therefore it is called a *multi-variable function.* How to handle such functions (also on a more scientific level) will be the contents of the course on *multi-variable calculus* in the second year.

Concerning the second question: keep in mind that the system

$$\begin{cases} a \cdot 45 + b \cdot 15 = 15, \\ a \cdot 15 + b \cdot 5 = 10 \end{cases}$$

possesses no solution $(a,b)$. How can we be sure that we never run into a situation where $n = 5$, $\sum_{i=1}^5 x_i^2 = 45$, $\sum_{i=1}^5 x_i = 15$, $\sum_{i=1}^5 x_i y_i = 15$ and $\sum_{i=1}^5 y_i = 10$ ?

**DIY:** *Try to find distinct real numbers $x_1$, $x_2$, $x_3$, $x_4$, $x_5$ and $y_1$, $y_2$, $y_3$, $y_4$, $y_5$ with the property that*

$$\begin{aligned} x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 &= 45, \\ x_1 + x_2 + x_3 + x_4 + x_5 &= 15, \\ x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + x_5y_5 &= 15, \\ y_1 + y_2 + y_3 + y_4 + y_5 &= 10. \end{aligned}$$

*You have ten variables at your free disposal, and only four equations to satisfy, so certainly it should be possible to find these ten numbers ?*

You will learn some techniques about how to handle such questions in the *Problem Solving* course of the second semester. And the theoretical background about when linear systems are actually solvable will be covered in the *Algebra* courses.

Concerning the third question: solving a linear system with two equations for two unknowns can certainly be done by hand. For bigger systems you need numerical methods and a computer for which you have to write a computer program; details will follow in the courses on *Numerical Analysis*.

Now let us look back at the world record example once again.

- The world record data are not all equal, because the older ones are hand-clocked with an accuracy of $0.1s$, but the newer ones are automatically clocked with an accuracy of $0.01s$. Therefore, maybe it is scientifically inappropriate to treat all the points $(x_1, y_1)$, ..., $(x_n, y_n)$ on an equal footing in our calculations. More on that in the courses on *Statistics*.
- In our attempt at approximating the data $(x_i, y_i)$ for the world records, we had chosen the ansatz $y = ax + b$. Perhaps this is too simple, and another ansatz might be

  $$y = c + e^{ax+b},$$

  with parameters $a$, $b$, $c$ to play with. Then the task becomes: for given $x_1 < x_2 < \cdots < x_n$ and given $y_1$, $y_2$, ..., $y_n$, how to find real numbers $a$, $b$, $c$ such that the sum

  $$f(a, b, c) = \sum_{i=1}^{n} \left(c + e^{ax_i+b} - y_i\right)^2$$

  becomes minimal. Such numbers $a$, $b$, $c$ indeed do exist, but finding them is now much harder (in contrast to Task 3 from above). We now have no formula that gives us the values of the optimal $a$, $b$, $c$. How to find reasonably good approximations to the optimal $a$, $b$, $c$ will be subject of the various courses on *Numerical Analysis*.

Chapter 5

# Mathematics in Production Planning

## 5.1 The Objective

Consider a factory that produces two products ($P_1$ and $P_2$), and the customers demand the following amount of products over the period of 8 weeks:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $P_1$ | 200 | 400 | 200 | 1800 | 800 | 200 | 100 | 300 |
| $P_2$ | 100 | 400 | 600 | 400 | 300 | 300 | 1200 | 200 |

The products are to be shipped on each Friday. Observe the peak demand of $P_1$ in week 4 and of $P_2$ in week 7. So either we spend a lot of money buying machines which we need only rarely or we produce more in earlier weeks and keep the products on storage.

**Question:** how to run the machines in each week, keeping in mind that the storage space is limited and costs money, with the objective of minimizing the costs ?

What you will learn: matrices, vectors, linear optimisation, working carefully.

## 5.2 Matrices

What **is** a matrix ? This is easily answered: you write certain numbers in a tabular form and put a pair of parentheses around. Next we learn how to multiply matrices by other matrices or by vectors. We also discover some properties of this multiplication. What **does** a matrix ? We will give a geometrical answer and obtain a deeper understanding.

### 5.2.1 Cakes and Their Ingredients

Imagine a cake shop that is baking their own cakes every day. They have (*for simplicity*) two different types of cake (type 1 and type 2) with ingredients like this:

| | flour | butter | sugar | eggs |
|---|---|---|---|---|
| type 1 | 0.4 kg | 0.2 kg | 0.25 kg | 3 |
| type 2 | 0.5 kg | 0.2 kg | 0.2 kg | 4 |

Suppose they bake on a certain day 5 cakes of type 1 and 7 cakes of type 2. How much of the ingredients is being taken out of the storage room on that day ? Obviously

$$\begin{aligned}
\text{flour:} &\quad 0.4\,\text{kg} \cdot 5 + 0.5\,\text{kg} \cdot 7 = 5.5\,\text{kg} \\
\text{butter:} &\quad 0.2\,\text{kg} \cdot 5 + 0.2\,\text{kg} \cdot 7 = 2.4\,\text{kg} \\
\text{sugar:} &\quad 0.25\,\text{kg} \cdot 5 + 0.2\,\text{kg} \cdot 7 = 2.65\,\text{kg} \\
\text{eggs:} &\quad 3 \cdot 5 + 4 \cdot 7 = 43
\end{aligned}$$

In mathematics, we have a much more compact notation:

$$\begin{pmatrix} 0.4\,\mathrm{kg} & 0.5\,\mathrm{kg} \\ 0.2\,\mathrm{kg} & 0.2\,\mathrm{kg} \\ 0.25\,\mathrm{kg} & 0.2\,\mathrm{kg} \\ 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} 5 \\ 7 \end{pmatrix} = \begin{pmatrix} 5.5\,\mathrm{kg} \\ 2.4\,\mathrm{kg} \\ 2.65\,\mathrm{kg} \\ 43 \end{pmatrix}.$$

We see here three matrices (two on the left-hand side, one on the right-hand side). Let us neglect the kilogram units for a moment.

**Definition 5.1.** *A rectangular shaped table with at least one column and at least one row is called* matrix*. If a matrix $A$ has $p$ rows and $q$ columns, and each of the $pq$ entries of $A$ is a real number, then we write this as*

$$A \in \mathbb{R}^{p\times q},$$

*and we write $A$ as*

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1q} \\ a_{21} & a_{22} & \dots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pq} \end{pmatrix}.$$

*Matrices with exactly one column are called* vectors*, and instead of $A \in \mathbb{R}^{p\times 1}$ we then often write $\vec{a} \in \mathbb{R}^p$,*

$$\vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}.$$

$$\boxed{\text{Vectors are always written as columns.}}$$

A *survival rule* in maths is to find abbreviations that help us to de-clutter our text.

Therefore we define the recipe matrix

$$R = \begin{pmatrix} 0.4\,\mathrm{kg} & 0.5\,\mathrm{kg} \\ 0.2\,\mathrm{kg} & 0.2\,\mathrm{kg} \\ 0.25\,\mathrm{kg} & 0.2\,\mathrm{kg} \\ 3 & 4 \end{pmatrix},$$

the cake-order vector

$$\vec{c} = \begin{pmatrix} 5 \\ 7 \end{pmatrix},$$

and the ingredient-amount vector

$$\vec{i} = \begin{pmatrix} 5.5\,\mathrm{kg} \\ 2.4\,\mathrm{kg} \\ 2.65\,\mathrm{kg} \\ 43 \end{pmatrix}.$$

We then have the equation $R \cdot \vec{c} = \vec{i}$ with $R \in \mathbb{R}^{4\times 2}$, $\vec{c} \in \mathbb{R}^2$ and $\vec{i} \in \mathbb{R}^4$.

The multiplication of the matrix $R$ by the vector $\vec{c}$ is easiest to remember when you use your two index fingers: put the left index finger onto the upper left corner of $R$, the right index finger onto the the top entry of $\vec{c}$. Now move the left index finger rightwards and the right index finger downwards, both fingers at the same speed. Multiply the numbers onto which your two fingers point, and add all these two products up.

Repeat with second row of $R$. Repeat with third row of $R$. Repeat with fourth row of $R$. The result then will be the ingredient-amount vector $\vec{i}$.

That was the explanation how to multiply matrix times vector.

Now comes the explanation how to multiply matrix times matrix: basically in the same way. Just imagine that the right-factor matrix consists of several columns that are written next to each other, and each column of the right-factor matrix can be construed as a vector in its own right. Do the multiplication "matrix times vector" as before, and write the obtained result vectors from left to right next to each other.

**Let us practise this.**

**DIY:** *Choose a matrix $A \in \mathbb{R}^{2\times 3}$ and a matrix $B \in \mathbb{R}^{3\times 2}$. Form the product $A \cdot B$ and the product $B \cdot A$. These are matrices of the shape $\mathbb{R}^{2\times 2}$ and $\mathbb{R}^{3\times 3}$, respectively. Check with your neighbour.*

*For a funny digression: add up the two entries of the \ diagonal of $A \cdot B$. Now add up the three entries of the \ diagonal of $B \cdot A$. Enjoy the surprise. Can you prove it for all matrices $A$ and $B$ of that mentioned shape ?*

**DIY:** *Can you find a matrix $A \in \mathbb{R}^{2\times 2}$ and a matrix $B \in \mathbb{R}^{2\times 2}$ both having $2 \cdot 2 = 4$ entries, of which none is the number zero, such that the product $A \cdot B$ is then a matrix full of zeros ? For your chosen matrices, then also calculate the product $B \cdot A$.*

Let us now take a generic cake-order vector $\vec{c} = \binom{c_1}{c_2}$. We don't know the values of $c_1$ and $c_2$. Then the ingredient amount vector is

$$\vec{i} = R\vec{c} = \begin{pmatrix} 0.4\,\text{kg} & 0.5\,\text{kg} \\ 0.2\,\text{kg} & 0.2\,\text{kg} \\ 0.25\,\text{kg} & 0.2\,\text{kg} \\ 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0.4\,\text{kg} \cdot c_1 + 0.5\,\text{kg} \cdot c_2 \\ 0.2\,\text{kg} \cdot c_1 + 0.2\,\text{kg} \cdot c_2 \\ 0.25\,\text{kg} \cdot c_1 + 0.2\,\text{kg} \cdot c_2 \\ 3 \cdot c_1 + 4 \cdot c_2 \end{pmatrix},$$

and we can't simplify this any further because we don't know the values of $c_1$ and $c_2$. However, the right-hand side is a certain vector from $\mathbb{R}^4$ (we do not know which one because $c_1$ and $c_2$ are not known), and we remember from school that addition of vectors is done for each component separately. Therefore we may deduce that

$$\vec{i} = R\vec{c} = \begin{pmatrix} 0.4\,\text{kg} \cdot c_1 \\ 0.2\,\text{kg} \cdot c_1 \\ 0.25\,\text{kg} \cdot c_1 \\ 3 \cdot c_1 \end{pmatrix} + \begin{pmatrix} 0.5\,\text{kg} \cdot c_2 \\ 0.2\,\text{kg} \cdot c_2 \\ 0.2\,\text{kg} \cdot c_2 \\ 4 \cdot c_2 \end{pmatrix},$$

and we also remember that constants can be dragged out of a vector and written in front of it:

$$\vec{i} = R\vec{c} = \begin{pmatrix} 0.4\,\text{kg} & 0.5\,\text{kg} \\ 0.2\,\text{kg} & 0.2\,\text{kg} \\ 0.25\,\text{kg} & 0.2\,\text{kg} \\ 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = c_1 \cdot \begin{pmatrix} 0.4\,\text{kg} \\ 0.2\,\text{kg} \\ 0.25\,\text{kg} \\ 3 \end{pmatrix} + c_2 \cdot \begin{pmatrix} 0.5\,\text{kg} \\ 0.2\,\text{kg} \\ 0.2\,\text{kg} \\ 4 \end{pmatrix},$$

We stare at this line and recognize that we have multiplied the first column of the recipe matrix $R$ by $c_1$, and we have multiplied the second column of the recipe matrix $R$ by $c_2$, and then we have added these two vectors from $\mathbb{R}^4$.

Let us generalise this a bit. The matrix-vector product

$$A\vec{x} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1q} \\ a_{21} & a_{22} & \dots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pq} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{pmatrix}$$

means that we

- take the first column of $A$, multiplied by $x_1$,
- and then take the second column of $A$, multiplied by $x_2$,
- and then take the third column of $A$, multiplied by $x_3$,
- and so on,
- and then take the $q$-th column of $A$, multiplied by $x_q$,

and then, finally, we sum up these $q$ products (each product being a vector from $\mathbb{R}^p$).

### 5.2.2 Prices of Ingredients and of Cakes

Now let us bring prices into the description:

| | flour | butter | sugar | egg |
|---|---|---|---|---|
| price | $1.2\frac{£}{\mathrm{kg}}$ | $1.4\frac{£}{\mathrm{kg}}$ | $0.9\frac{£}{\mathrm{kg}}$ | $0.15£$ |

We arrange them into the price matrix:

$$P := \begin{pmatrix} 1.2\frac{£}{\mathrm{kg}} & 1.4\frac{£}{\mathrm{kg}} & 0.9\frac{£}{\mathrm{kg}} & 0.15£ \end{pmatrix}.$$

This matrix $P \in \mathbb{R}^{1\times 4}$ has one row and four columns.

The cake shop bakes 5 cakes of type 1 and 7 cakes of type 2. What do all the ingredients cost together ?

**Method 1:** We know already the needed ingredients for these 5 and 7 cakes together:

$$\vec{i} = R \cdot \vec{c} = \begin{pmatrix} 5.5\,\mathrm{kg} \\ 2.4\,\mathrm{kg} \\ 2.65\,\mathrm{kg} \\ 43 \end{pmatrix}.$$

Therefore, the total price $\mathrm{Sum}_{\mathrm{total}}$ is

$$\begin{aligned}\mathrm{Sum}_{\mathrm{total}} &= 1.2\frac{£}{\mathrm{kg}} \cdot 5.5\,\mathrm{kg} + 1.4\frac{£}{\mathrm{kg}} \cdot 2.4\,\mathrm{kg} + 0.9\frac{£}{\mathrm{kg}} \cdot 2.65\,\mathrm{kg} + 0.15£ \cdot 43 \\ &= 6.6£ + 3.36£ + 2.385£ + 6.45£ \\ &= 18.795£.\end{aligned}$$

We can write this as

$$\mathrm{Sum}_{\mathrm{total}} = P \cdot \vec{i} = P \cdot \left(R \cdot \vec{c}\right).$$

**Method 2:** The price for one cake of type 1 is found by evaluating the matrix-matrix product

$$\begin{aligned}&\begin{pmatrix} 1.2\frac{£}{\mathrm{kg}} & 1.4\frac{£}{\mathrm{kg}} & 0.9\frac{£}{\mathrm{kg}} & 0.15£ \end{pmatrix} \cdot \begin{pmatrix} 0.4\,\mathrm{kg} \\ 0.2\,\mathrm{kg} \\ 0.25\,\mathrm{kg} \\ 3 \end{pmatrix} \\ &= \left(1.2\frac{£}{\mathrm{kg}} \cdot 0.4\,\mathrm{kg} + 1.4\frac{£}{\mathrm{kg}} \cdot 0.2\,\mathrm{kg} + 0.9\frac{£}{\mathrm{kg}} \cdot 0.25\,\mathrm{kg} + 0.15£ \cdot 3\right) \\ &= \left(0.48£ + 0.28£ + 0.225£ + 0.45£\right) \\ &= 1.435£.\end{aligned}$$

Now we perform a dance of joy and happiness because all the kilogram units have nicely cancelled, and all the four items which we added have the same unit £.

And the price for one cake of type 2 is found like this:

$$\begin{aligned}&\begin{pmatrix} 1.2\frac{£}{\mathrm{kg}} & 1.4\frac{£}{\mathrm{kg}} & 0.9\frac{£}{\mathrm{kg}} & 0.15£ \end{pmatrix} \cdot \begin{pmatrix} 0.5\,\mathrm{kg} \\ 0.2\,\mathrm{kg} \\ 0.2\,\mathrm{kg} \\ 4 \end{pmatrix} \\ &= \left(1.2\frac{£}{\mathrm{kg}} \cdot 0.5\,\mathrm{kg} + 1.4\frac{£}{\mathrm{kg}} \cdot 0.2\,\mathrm{kg} + 0.9\frac{£}{\mathrm{kg}} \cdot 0.2\,\mathrm{kg} + 0.15£ \cdot 4\right) \\ &= \left(0.6£ + 0.28£ + 0.18£ + 0.6£\right) \\ &= 1.66£.\end{aligned}$$

We can write this as

$$P \cdot R = \begin{pmatrix} 1.2\frac{£}{\mathrm{kg}} & 1.4\frac{£}{\mathrm{kg}} & 0.9\frac{£}{\mathrm{kg}} & 0.15£ \end{pmatrix} \cdot \begin{pmatrix} 0.4\,\mathrm{kg} & 0.5\,\mathrm{kg} \\ 0.2\,\mathrm{kg} & 0.2\,\mathrm{kg} \\ 0.25\,\mathrm{kg} & 0.2\,\mathrm{kg} \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1.435£ & 1.66£ \end{pmatrix}.$$

The cake shop bakes 5 cakes of type 1 and 7 cakes of type 2, which then amounts to the total price

$$\begin{pmatrix} 1.435£ & 1.66£ \end{pmatrix} \cdot \begin{pmatrix} 5 \\ 7 \end{pmatrix} = 1.435£ \cdot 5 + 1.66£ \cdot 7 = 18.795£,$$

and our calculation can be written in compressed form as

$$\mathrm{Sum}_{\mathrm{total}} = \Big(P \cdot R\Big) \cdot \begin{pmatrix} 5 \\ 7 \end{pmatrix} = \Big(P \cdot R\Big) \cdot \vec{c}.$$

**Comparison of both methods:** we obtained the same total price 18.795£ with both methods, and we can express this observation as

$$P \cdot \Big(R \cdot \vec{c}\Big) = \Big(P \cdot R\Big) \cdot \vec{c}.$$

$$\boxed{\text{The matrix-matrix-product is associative.}}$$

$$\boxed{\text{If necessary, we read column vectors as matrices with just one column.}}$$

### 5.2.3 What is the Meaning of all this ?

$$\boxed{\text{A matrix } A \in \mathbb{R}^{r\times q} \text{ with } q \text{ columns and } r \text{ rows induces a mapping from } \mathbb{R}^q \text{ into } \mathbb{R}^r.}$$

We give three examples of this general principle:

**The matrix $R$:** the ingredients-per-cake matrix $R \in \mathbb{R}^{4\times 2}$ enables us to calculate the ingredients $\vec{i} \in \mathbb{R}^4$ from the known cake-order vector $\vec{c} \in \mathbb{R}^2$.

**The matrix $P$:** the price-per-ingredient matrix $P \in \mathbb{R}^{1\times 4}$ enables us to calculate the price (which is a vector from $\mathbb{R}^1$, AKA a real number) from a known ingredient vector $\vec{i} \in \mathbb{R}^4$ that we have taken out of the storage room.

**The matrix $PR$:** the matrix $PR \in \mathbb{R}^{1\times 2}$ is obtained as product of the price-per-ingredient matrix $P$ by the ingredient-per-cake matrix $R$, hence it is a price-per-cake matrix, and it calculates the total price (which is a vector from $\mathbb{R}^1$, AKA a real number) from a known cake-order vector $\vec{c} \in \mathbb{R}^2$.

This can be understood also geometrically: a vector from $\mathbb{R}^q$ is some geometrical thing. If you take your matrix $A \in \mathbb{R}^{r\times q}$ and multiply it by that vector, then you get another vector that lives in $\mathbb{R}^r$. In some sense, you have a machine, where you put a geometric vector from $\mathbb{R}^q$ in, and where you get a geometric vector from $\mathbb{R}^r$ out.

## 5.3 Production Planning

What do we intend to include into our calculations for the production planning problem from the introduction ?

We make a list:

- amount of ingredients needed for production
- shipping costs for ingredients
- low reliability of supplier company
- *The wrong kind of snow* on the roads
- supplier delivers ingredients too early and then we don't have enough space in storage
- storage costs for ingredients

- storage restrictions for ingredients (e.g., limited space)
- maybe some ingredients perish fast and cannot be stored for long
- capacity restrictions for machines
- working time restrictions for employees
- extra salary for overtime hours of employees
- vacation plans of workers
- shipping costs for products
- storage costs for products
- storage limits for products
- unpredictable cancellations of orders

Perhaps this list is incomplete. Obviously we cannot include all these effects into our considerations, because this gets quickly too complicated for a first year course. But it is crucial to clearly specify what we include into our modelling and what we keep out.

*Survival rule:* if we do not get our thinking and our writing straight at the beginning, then all further calculations will never lead us to the correct solution. That is why it is essential to write your homeworks in complete sentences.

Our assumptions are:

- All suppliers, customers, machines, employees are reliable (otherwise we would need to use methods from probability theory, which we don't have at our disposal).
- The salaries of the workers are arranged in such a way that however we schedule the machines, they do not change. Hence we can neglect the salary costs.
- Before the start of week 1, we have enough ingredients stored to produce everything in the next 8 weeks. The storage cost for the ingredients are zero. Somehow the storage will be replenished during the 8 weeks, so that we get back to the initial value on Saturday of week 8.

  This implies that we can neglect the cost of the ingredients and their shipment. We don't even care *what* the ingredients are.
- Machines run only from Monday till Friday.
- Shipments to customers are being done every Friday evening after work is finished.
- There is a limit on how much the machines can produce per week.
- Storage space for products have a limit, and storing them has a positive cost per week and per item.

That was *Step 1* of the *Modelling Cycle.* Now we turn these modelling assumptions into mathematical expressions, which is *Step 2.*

We know the demands of the products:

$$D = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \\ \vdots & \vdots \\ d_{81} & d_{82} \end{pmatrix} = \begin{pmatrix} 200 & 100 \\ 400 & 400 \\ 200 & 600 \\ 1800 & 400 \\ 800 & 300 \\ 200 & 300 \\ 100 & 1200 \\ 300 & 200 \end{pmatrix}, \tag{5.1}$$

and $d_{ij}$ denotes the amount of product $j$ to be shipped on Friday evening of week $i$.

We want to find the production plan $X$, with

$$X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{81} & x_{82} \end{pmatrix},$$

and $x_{ij}$ denotes the number of items of product $j$ produced during week $i$.

We are also interested in the storage plan $S$, with

$$S = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \\ \vdots & \vdots \\ s_{81} & s_{82} \end{pmatrix},$$

and $s_{ij}$ denotes the number of items of product $j$ in storage on Monday morning of week $i$. It holds $s_{11} = s_{12} = 0$, and we have the obvious relations

$$s_{i+1,j} = s_{ij} + x_{ij} - d_{ij}, \qquad \forall i \in \{1, \dots, 7\}, \qquad \forall j \in \{1, 2\}.$$

These are 14 equations. The matrix $S$ is mostly a tool; we could avoid introducing this matrix, but then various equations and inequalities further down would become more cumbersome to write.

We have also the obvious restrictions

$$x_{ij} \geq 0, \qquad s_{ij} \geq 0, \qquad \forall i \in \{1, \dots, 8\}, \qquad \forall j \in \{1, 2\}.$$

These are 32 inequalities.

The demand of the last week has to be satisfied as well, which gives two more inequalities

$$s_{8j} + x_{8j} - d_{8j} \geq 0, \qquad \forall j \in \{1, 2\}.$$

There is a limit on how much the machines can produce per week:

$$x_{ij} \leq b_j^p, \qquad \forall i \in \{1, \dots, 8\}, \qquad \forall j \in \{1, 2\}.$$

The superscript $p$ at $b_j^p$ means *production*; it is not an exponent. We choose $b_1^p = 700$ and $b_2^p = 600$, and we have 16 inequalities here.

And there is a bound on the available storage space:

$$s_{i1} + x_{i1} + s_{i2} + x_{i2} \leq b^s, \qquad i \in \{1, \dots, 8\},$$

where the superscript $s$ means *storage*. We have to add $x_{i1}$ and $x_{i2}$ because we need enough storage also on Friday early afternoon. Let us take $b^s = 3000$.

The given quantities are

$$d_{ij}, \quad b_j^p, \quad b^s,$$

and the wanted quantities are

$$x_{ij}, \qquad s_{ij}.$$

These wanted quantities form some point in $\mathbb{R}^{32}$, because we have 32 unknown quantities. But not every point in $\mathbb{R}^{32}$ describes a valid situation. We have various constraints on the $x_{ij}$ and $s_{ij}$; and if these constraints are not being met, then that point from $\mathbb{R}^{32}$ is forbidden. These constraints are expressed partly as equations (such as $s_{12} = 0$ or $s_{i+1,j} = s_{ij} + x_{ij} - d_{ij}$) and partly as inequalities (such as $x_{ij} \leq b_j^p$). All those points from $\mathbb{R}^{32}$ that satisfy all the constraints form the *feasible region*, by definition.

And our **objective** is now: to find a production plan $X$ (and subsequently a storage plan $S$), such that we are in the feasible region, in such a way that costs are minimized. The only costs are related to the storage of products. As explained above, salary costs etc. do not matter.

In week $i$ (with $i \in \{1, \dots, 8\}$), the storage of product $j$ (with $j \in \{1, 2\}$) grows in an affine linear way from $s_{ij}$ to $s_{ij} + x_{ij}$. In a diagram of storage-over-time, it looks like a trapezoid. Then the storage costs of that product in that week are

$$c_j \cdot \frac{s_{ij} + (s_{ij} + x_{ij})}{2} = c_j \cdot \left( s_{ij} + \frac{x_{ij}}{2} \right),$$

with some given $c_j$. Then the total costs are

$$C = \sum_{i=1}^{8} \sum_{j=1}^{2} c_j \cdot \left( s_{ij} + \frac{x_{ij}}{2} \right).$$

Let us choose $c_1 = 3$ and $c_2 = 4$.

We summarise:

**Given** are the matrix $D$ from (5.1), and the numbers $b_1^p = 700$, $b_2^p = 600$, $b^s = 3000$, $c_1 = 3$, $c_2 = 4$.

**Objective:** to minimize the objective function $\sum_{i=1}^{8} \sum_{j=1}^{2} c_j (s_{ij} + \frac{x_{ij}}{2})$

**for the variables** $\{x_{ij}, s_{ij} \colon 1 \le i \le 8, \quad 1 \le j \le 2\} \in \mathbb{R}^{32}$

**living in the feasible region** that is described by the 16 equality constraints

$$\begin{aligned} &s_{11} = 0, \quad s_{12} = 0, \\ &s_{ij} - s_{i+1,j} + x_{ij} = d_{ij}, \qquad 1 \le i \le 7, \quad 1 \le j \le 2 \end{aligned}$$

and the 32 nonnegativity constraints

$$x_{ij} \ge 0, \qquad s_{ij} \ge 0, \qquad 1 \le i \le 8, \quad 1 \le j \le 2,$$

and the 26 inequality constraints

$$-x_{8j} - s_{8j} \le -d_{8j}, \qquad x_{ij} \le b_j^p, \qquad s_{i1} + s_{i2} + x_{i1} + x_{i2} \le b^s, \qquad 1 \le i \le 8, \quad 1 \le j \le 2.$$

Now the crucial observation is that the unknown numbers $x_{ij}$ and$s_{ij}$ only appear *linearly* — they are not squared or put into nonlinear functions such as logarithm or sine, and we never multiply two unknown quantities.

How does the feasible region look like ? It is a subset of $\mathbb{R}^{32}$, and it is obtained like this: each inequality $x_{ij} \ge 0$ cuts the $\mathbb{R}^{32}$ into two equal pieces, of which one is thrown away (namely that one where $x_{ij}$ is negative, for that choice of $ij$). Similarly, each of the 8 inequalities $s_{i1} + s_{i2} \le b^2$ (where $1 \le i \le 8$) cuts the $\mathbb{R}^{32}$ into two equal pieces, of which one is thrown away. The feasible region is what remains in the end, and it is a so-called *polyedron*. Crucial for us is that the feasible region is *convex*[1] which is defined mathematically like this: a set in $\mathbb{R}^n$ is convex if and only if the following holds: if two points belong to that set, then also their connecting line segment. A horseshoe is not convex.

The key advantage of the convexity of the feasible region is: the minimum of the objective function then is attained at a vertex of the feasible region. Therefore, we only have to visit all the vertices of the feasible region, and always calculate the value of the objective function there, and then pick the smallest value.

The bad news is that our feasible region has perhaps hundreds of vertices, and now we definitely need a computer. The method is called *Linear Programming*, and you will learn details about it in the courses of *Optimization*.

We are now entering *Step 3* of the modelling cycle. We build a long column vector of the 32 unknowns, and we call it $\vec{z} \in \mathbb{R}^{32}$. First come the $x_{j1}$, below that the $x_{j2}$, below that the $s_{j1}$, and then the $s_{j2}$. Consequently, the translation rules are

$$x_{ij} = z_{i+(j-1)\cdot 8}, \qquad s_{ij} = z_{16+i+(j-1)\cdot 8}, \qquad 1 \le i \le 8, \quad 1 \le j \le 2.$$

[1] from the Latin adjective *convexus* which means arched, vaulted, which comes from the Latin verb *conveho, convehere, convexi, convectus*, meaning *to bring (veho) together (cum), to convey.* The idea is: to form an arch, you have to bring its ends together.

If we do not work carefully, then we will mix up the indices in what follows.
That is just one reason for the recommendation to always write your homeworks
in complete sentences.

We also build a vector $\vec{b}^{eq} \in \mathbb{R}^{16}$ as follows:

$$d_{ij} =: b^{eq}_{i+(j-1)\cdot 7}, \quad 1 \le i \le 7, \quad 1 \le j \le 2, \quad b^{eq}_{15} := 0, \quad b^{eq}_{16} := 0.$$

Then we can express the 16 equality restraints as an equation $A^{eq}\vec{z} = \vec{b}^{eq}$, with $A^{eq}$ being a matrix with 32 columns and 16 rows:

The top 7 rows of $A^{eq}$ correspond to $x_{i1} + s_{i1} - s_{i+1,1} = d_{i1}$, with $1 \le i \le 7$. Hence we put

$$1 \le k \le 7\colon \quad a^{eq}_{kk} = 1, \quad a^{eq}_{k,k+16} = 1, \quad a^{eq}_{k,k+17} = -1.$$

The next 7 rows of $A^{eq}$ correspond to $x_{i2} + s_{i2} - s_{i+1,2} = d_{i2}$, with $1 \le i \le 7$. Therefore

$$8 \le k \le 14\colon \quad a^{eq}_{k,k+1} = 1, \quad a^{eq}_{k,k+17} = 1, \quad a^{eq}_{k,k+18} = -1.$$

And the bottom 2 rows of $A^{eq}$ are related to $s_{11} = 0$ and $s_{12} = 0$, hence

$$a^{eq}_{15,17} = 1, \qquad a^{eq}_{16,25} = 1.$$

All other entries $a^{eq}_{ij}$ that have not been mentioned here are zero.

Now we come to the 26 inequality constraints, which we wish to write as $A^{ineq}\vec{z} \lll \vec{b}^{ineq}$, where the symbol $\lll$ means "componentwise less than or equal to". The vector $\vec{b}^{ineq}$ is a column with 26 entries like this:

$$\vec{b}^{ineq} := \left( \underbrace{b^p_1, \dots, b^p_1}_{8}, \underbrace{b^p_2, \dots, b^p_2}_{8}, \underbrace{b^s, \dots, b^s}_{8}, -d_{81}, -d_{82} \right)^\top,$$

and the $\top$ means "transposition", which turns it into a column.

The top $8 + 8$ rows of $A^{ineq}$ relate to $x_{ij} \le b^p_j$, therefore

$$1 \le k \le 16\colon \quad a^{ineq}_{kk} = 1.$$

The next 8 rows of $A^{ineq}$ relate to $x_{i1} + x_{i2} + s_{i1} + s_{i2} \le b^s$, hence

$$17 \le k \le 24: \quad a^{ineq}_{k,k-16} = 1, \quad a^{ineq}_{k,k-8} = 1, \quad a^{ineq}_{kk} = 1, \quad a^{ineq}_{k,k+8} = 1.$$

And the bottom 2 rows of $A^{ineq}$ correspond to $-x_{8j} - s_{8j} \le -d_{8j}$, consequently

$$a^{ineq}_{25,8} = -1, \quad a^{ineq}_{25,24} = -1, \quad a^{ineq}_{26,16} = -1, \quad a^{ineq}_{26,32} = -1.$$

Again, all other entries $a^{ineq}_{ij}$ are zero.

The feasible region is then

$$\Omega := \left\{ \vec{z} \in \mathbb{R}^{32} : \quad A^{eq}\vec{z} = \vec{b}^{eq}, \quad A^{ineq}\vec{z} \lll \vec{b}^{ineq}, \quad 0 \lll \vec{z} \right\}.$$

And the objective function is

$$\vec{z} \mapsto \vec{\gamma} \cdot \vec{z}$$

with a vector $\vec{\gamma} \in \mathbb{R}^{32}$ as follows:

$$\vec{\gamma} := \left( \underbrace{\frac{c_1}{2}, \dots \frac{c_1}{2}}_{8}, \underbrace{\frac{c_2}{2}, \dots \frac{c_2}{2}}_{8}, \underbrace{c_1, \dots, c_1}_{8}, \underbrace{c_2, \dots, c_2}_{8}, \right)^\top.$$

The problem to be solved is then

to minimize $\quad \vec{\gamma} \cdot \vec{z} \quad$ subject to the constraint $\quad \vec{z} \in \Omega.$

It is time for some `python` program. We remark that the indices that enumerate the entries in a vector or in a matrix start with zero in python and start with one in the usual mathematics. Therefore we have to work even more careful than we did up to now:

```
import scipy as SP
import scipy.optimize as SPO

demand = [[200.0, 100], [400,400], [200, 600], [1800,400],
          [800,300],[200,300],[100,1200],[300,200]]
demand = SP.mat(demand) # turns it into a scipy matrix
print "The demand matrix is "
print demand
print " "

##################################################################

b1p = 700.0
b2p = 600.0
bs = 3000.0


beq = SP.zeros((16,1))      # constructs a zero matrix with 16 rows and 1 column
beq = SP.mat(beq)           # turns it into a scipy matrix
                            # rows and columns are numbered from zero
beq[0:7,0] = demand[0:7,0]  # 0:7 means "0 1 2 3 4 5 6"
beq[7:14,0] = demand[0:7,1] # 7:14 means "7 8 9 10 11 12 13"
print "The right-hand side of the equality constraints is"
print beq
print " "

bineq = SP.zeros((26,1))
bineq = SP.mat(bineq)
bineq[0:8,0] = b1p * SP.ones((8,1))  # 0:8 means "0 1 ... 7"
bineq[8:16,0] = b2p * SP.ones((8,1)) # 8:16 means "8 9 ... 15"
bineq[16:24,0] = bs * SP.ones((8,1))
bineq[24,0] = -demand[7,0]
bineq[25,0] = -demand[7,1]
print "The right-hand side of the inequality constraints is"
print bineq
print " "

##################################################################

Aeq = SP.zeros((16,32))
Aeq = SP.mat(Aeq)
for k in range(7):    # range(7) means "0 1 ... 6"
    Aeq[k,k] = 1.0
    Aeq[k,k+16] = 1.0
    Aeq[k,k+17] = -1.0
for k in range(7,14):  # range(7,14) means "7 8 ... 13"
    Aeq[k,k+1] = 1.0
    Aeq[k,k+17] = 1.0
    Aeq[k,k+18] = -1.0
print "The matrix Aeq is "
print Aeq
print " "

Aineq = SP.zeros((26,32))
Aineq = SP.mat(Aineq)
for k in range(16):
    Aineq[k,k] = 1.0
for k in range(16,24):
    Aineq[k,k-16] = 1.0
    Aineq[k,k-8] = 1.0
    Aineq[k,k] = 1.0
    Aineq[k,k+8] = 1.0
Aineq[24,7] = -1.0
Aineq[24,23] = -1.0
Aineq[25,15] = -1.0
Aineq[25,31] = -1.0

##################################################################

c1 = 3.0
c2 = 4.0

gammavec = SP.zeros(32)
gammavec[0:8] = c1 / 2.0 * SP.ones(8)
gammavec[8:16] = c2 / 2.0 * SP.ones(8)
gammavec[16:24] = c1 * SP.ones(8)
gammavec[24:32] = c2 * SP.ones(8)

##################################################################


res = SPO.linprog(gammavec, A_ub = Aineq, b_ub = bineq, A_eq = Aeq, b_eq = beq)

print "The output of the linear programming procedure is "
print res
```

Which produces the following output:

```
The demand matrix is
[[  200.   100.]
 [  400.   400.]
 [  200.   600.]
 [ 1800.   400.]
 [  800.   300.]
 [  200.   300.]
 [  100.  1200.]
 [  300.   200.]]

The right-hand side of the equality constraints is
```

```
[[  200.]
 [  400.]
 [  200.]
 [ 1800.]
 [  800.]
 [  200.]
 [  100.]
 [  100.]
 [  400.]
 [  600.]
 [  400.]
 [  300.]
 [  300.]
 [ 1200.]
 [    0.]
 [    0.]]

The right-hand side of the inequality constraints is
[[  700.]
 [  700.]
 [  700.]
 [  700.]
 [  700.]
 [  700.]
 [  700.]
 [  700.]
 [  600.]
 [  600.]
 [  600.]
 [  600.]
 [  600.]
 [  600.]
 [  600.]
 [  600.]
 [ 3000.]
 [ 3000.]
 [ 3000.]
 [ 3000.]
 [ 3000.]
 [ 3000.]
 [ 3000.]
 [ 3000.]
 [ -300.]
 [ -200.]]

The matrix Aeq is
[[ 1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1. -1.
   0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.
  -1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
   1. -1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
   0.  1. -1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
   0.  0.  1. -1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
   0.  0.  0.  1. -1.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
   0.  0.  0.  0.  1. -1.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.
   0.  0.  0.  0.  0.  0.  1. -1.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.  0.
   0.  0.  0.  0.  0.  0.  0.  1. -1.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.
   0.  0.  0.  0.  0.  0.  0.  0.  1. -1.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.  0.
   0.  0.  0.  0.  0.  0.  0.  0.  0.  1. -1.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.
   0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1. -1.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.
   0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1. -1.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.
   0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1. -1.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
   0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
   0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]]

The output of the linear programming procedure is
  status: 0
   slack: array([  100.,     0.,     0.,     0.,     0.,   500.,   600.,   400.,
          500.,   200.,     0.,   200.,     0.,     0.,     0.,   400.,
         2300.,  1500.,  1000.,   700.,  1600.,  1900.,  1700.,  2500.,
            0.,     0.])
 success: True
     fun: 23800.0
       x: array([  600.,   700.,   700.,   700.,   700.,   200.,   100.,   300.,
          100.,   400.,   600.,   400.,   600.,   600.,   600.,   200.,
            0.,   400.,   700.,  1200.,   100.,     0.,     0.,     0.,
            0.,     0.,     0.,     0.,     0.,   300.,   600.,     0.])
 message: 'Optimization terminated successfully.'
     nit: 22
```

---

Some explanation: the array of slack variables has 26 entries that tell us for each of the 26 inequality constraints the difference between RHS and LHS. The array of the $x$ variables contains the optimal vector $\vec{z}$ from which we can read off the desired production plan $X$. The `python` procedure `linprog` needed 22 iterations to find this solution, and the total cost is 23800.0.

# Chapter 6

# Mathematics in Your SatNav

The material in this chapter is taken from [1] and [2].

## 6.1 Physical Background and Description of the Problem

**The situation:** GPS satellites fly around the earth, users on earth have SatNav devices which determine the user's location. The devices never send any signal to the satellites, but they only receive signals from the satellites. We intend to answer the following **question:** how can the SatNav device determine its location with an accuracy of about 15 metres ?

We go more into the details: the earth radius is about 6370 kilometres, the satellites fly in an orbit of circular shape with a radius of 26.600 kilometres, the centre of these orbits is always the centre of the earth. Originally (1993) there were 24 satellites distributed on 6 different planes that are tilted against the equator plane by 55° (out of curiosity: why 55° ?). From most points on the earth, at least 6 satellites are always visible. Nowadays, there are approximately 30 satellites, and they are arranged a bit more irregularly.

Each satellite needs 11 hours and 58 minutes for a revolution around the earth. Recall that the earth needs 23 hours and 56 minutes for one revolution around its axis. (Out of curiosity: what has happened to the missing 4 minutes ?) Therefore, the satellites fly over the same place on earth twice per day.

The light speed is approximately $3 \cdot 10^8 \frac{m}{s}$. Each satellite has several atomic clocks with an accuracy of about 10 nanoseconds.

Each satellite sends a sequence of signals electromagnetically. Each such signal is either a zero or a one (we make no attempt at explaining what "signal zero" and "signal one" mean in terms of electromagnetic waves), and there are 1023 such signals per millisecond. Each signal is called a *bit*. Hence the satellites transmit $1.023 \cdot 10^6$ bits per second, and each of these bits propagate with light speed.

There is some complicated mathematical procedure that makes one of these 1023 bits *special* (expounding the details of this procedure is one of the points of this chapter). The SatNav device has to find that special bit by means of mathematics. Each bit of the $1.023 \cdot 10^6$ bits per second that propagates with light speed $3 \cdot 10^8 \frac{m}{s}$ hence occupies a spatial region of length

$$\frac{3 \cdot 10^8 \frac{m}{s}}{1.023 \cdot 10^6 \frac{1}{s}} \approx 300m.$$

Hence the SatNav device can determine its own location up to an error of about 300 metres (how to improve that to 15 metres would need one more chapter of explanations).

This description is very vague, we have glossed over many details.

There are **obstacles** to overcome in the engineering process:

- We have no clear meaning of "special bit". A naive idea would be to send a string of 1022 zeros and one one. But this does not really work because some bits will become damaged in the flight over a distance of more than 20.000 kilometres.

- The satellites are sending with a power of 27 Watts. When the signal arrives at the SatNav device, there is only a power of 100 attowatts remaining, which is $100 \cdot 10^{-18}$ Watts. Receiving such a weak signal is physically hard, considering that there a so many other signals flying around.
- The user's device has no atomic clock, for reasons of cost. Typically it contains a clock, but it has a much lower accuracy than an atomic clock.
- The user and its device do not know the location of the satellites up to an accuracy of 300 metres (moreover, the satellites travel with a velocity of about $3.8km/s$).
- The SatNav device cannot rely on previous knowledge of its own location because it is a realistic situation that a user switches their smart-phone on after an intercontinental flight.

The complete GPS system comprises

- all the users' devices
- the satellites
- the control centre with antennas on various continents.

The control centre observes where all the satellites are, and it tells the satellite about its orbit every two hours (these data are called *ephemeris*), valid for 4 hours. The control centre also tells each satellite about the orbits of the other satellites (called *almanac* data), valid for a day. All clocks of all satellites are synchronized via the control centre.

Hence all satellites know their own position and time, and each satellite will announce to the whole world the ephemeris data, the almanac data, and the time. These data are being transmitted with a speed of 50 bits per second. Sending the ephemeris data takes 30 seconds, 12.5 minutes are required for the almanac data. That is the main reason why your SatNav device needs about 30 seconds before it can tell you your own position.

In the sequel, we will attempt to get a first understanding how GPS works, and, in doing so, we will meet the following branches of mathematics:

- probability theory
- abstract algebra
- multi-variable calculus
- numerics.

We will also see that a little understanding of the special and the general theory of relativity will prove useful.

## 6.2 Elementary Ideas of Probability Theory

**Definition 6.1 (Unscientific).** *A* random variable *is a variable that depends on chance and can take only a finite number of values $v_1$, $v_2$, $\dots$, $v_n$ with probabilities $p_1$, $p_2$, $\dots$, $p_n$, respectively.*

This is not scientific because we can't specify what "depends on chance" actually means, and we also omitted several calculation rules for probabilities. You are expected to learn the details in the *Statistics* courses.

**Examples:**

**toss a fair coin:** *the values are $v_1 = 1$ and $v_2 = 2$, with probabilities $p_1 = p_2 = \frac{1}{2}$.*

**roll a fair dice:** *the values are $v_1 = 1$, $v_2 = 2$, $\dots$, $v_6 = 6$, and the probabilities are $p_j = \frac{1}{6}$ $\forall j \in \{1, 2, \dots, 6\}$.*

**roll a red dice and a green dice, both fair:** *the values are* $v_1 = (1,1)$, $v_2 = (1,2)$, …, $v_6 = (1,6)$, …, $v_{36} = (6,6)$*, where the first position in the parentheses stands for the value of the red dice. The probabilities are* $p_j = \frac{1}{36}$ *for all* $j$*. Observe in this example that the value* $v_j$ *of a random variable need not be a real number (here the values are ordered pairs of real numbers).*

To simplify notation, we will write $\mathbb{P}$(event) for "the probability of event".

**Definition 6.2 (statistically independent).** *Two random variables* $x$ *and* $y$ *with possible values* $x_1$, …, $x_m$ *and* $y_1$, …, $y_n$*, respectively, are called* statistically independent *if*

$$\forall k \in \{1,\ldots,m\}, \qquad \forall \ell \in \{1,\ldots,n\}: \qquad \mathbb{P}\Big((x = x_k) \text{ and } (y = y_\ell)\Big) = \mathbb{P}(x = x_k) \cdot \mathbb{P}(y = y_\ell).$$

**Examples:**

- *two distinguishable fair dice*
- *a monkey rolls a red dice (random variable* $r$*) and a green dice (random variable* $g$*). Then a human forms the sum* $s := r + g$ *which is again a random variable with values* 2, 3, …, 12*. The random variables* $r$ *and* $s$ *are* not *statistically independent.*

**DIY:** *Why ?*

**Definition 6.3 (expectation value).** *If* $x$ *is a random variable with real values* $v_1$, …, $v_n$ *and corresponding probabilities* $p_1$, …, $p_n$*, then the* expectation value $\mathbb{E}[x]$ *is defined as*

$$\mathbb{E}[x] := \sum_{j=1}^{n} p_j \cdot v_j.$$

**Example:** *Let* $g$ *be the random variable associated to a green fair dice, then*

$$\mathbb{E}[g] = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5.$$

*If the wooden dice has been rigged by means of a strategically places piece of lead, then some probabilities* $p_j$ *will grow at the expense of other probabilities, and* $\mathbb{E}[g]$ *will change accordingly.*

The meaning of the expectation value is that you take the average over all possible outcomes $v_j$, and this average is weighted with the probabilities $p_j$.

**Example:** ***(European Roulette)*** *We bet* 1£ *on the Roulette number 15. The numbers are* 0, 1, …, 36*. If we lose, then the* 1£ *is lost. If we win, then we get that* 1£ *back and 35 additional pounds.*

*The expectation value of the pay-out is*

$$p_1 \cdot v_1 + p_2 \cdot v_2 = \frac{1}{37} \cdot 36£ + \frac{36}{37} \cdot 0£ = \frac{36}{37}£.$$

*The expectation value of the pay-in is*

$$1 \cdot 1£ = 1£.$$

*Which is more than the expected pay-out. The casinos make their living from this difference.*

It is crucial to understand that the expectation value of a random variable only describes an average over all possible outcomes. But it makes no statement about a specific instance of that random variable. For example, if the life expectation of a newborn in a certain country is 78 years, then this result has been obtained by taking the average over the lifespans of many persons that have actually lived. But it gives no guarantee about how long one specific newborn person will live.

## 6.3 Random Binary Sequences

**Definition 6.4 (random binary sequence).** *A sequence* $\vec{c} = (c_0, c_1, \ldots, c_{n-1})$ *is called a* random binary sequence of length $n$ *if each* $c_j$ *is a random variable with values* 0 *and* 1*, with associated probabilities* $p_1 = p_2 = \frac{1}{2}$*, and each* $c_i$ *is statistically independent to each* $c_k$*, provided* $i \neq k$*.*

Imagine $n$ independent monkeys, who (being British) stand in an ordered queue, and each monkey is tossing a fair coin.

**Definition 6.5 (Hamming distance).** *Let $\vec{a} = (a_0, \dots, a_{n-1})$ and $\vec{c} = (c_0, \dots, c_{n-1})$ be two random binary sequences of length $n$. The* HAMMING[1]*-distance* $\operatorname{dist}(\vec{a}, \vec{c})$ *is the number of positions where $\vec{a}$ and $\vec{c}$ differ.*

**Example:** *We have* $\operatorname{dist}((010010), (011110)) = 2$.

Obviously, if $\vec{a}$ and $\vec{c}$ have length $n$, then $0 \le \operatorname{dist}(\vec{a}, \vec{c}) \le n$.

**Lemma 6.6.** *If $\vec{a}$ and $\vec{c}$ are both varying randomly, then* $\operatorname{dist}(\vec{a}, \vec{c})$ *is a random variable with expectation value*

$$\mathbb{E}\left[\operatorname{dist}(\vec{a}, \vec{c})\right] = \frac{n}{2}.$$

*The same expectation value is obtained if only one of the sequences is variable, and the other sequence has been fixed.*

[No proof given.]

Recall that such expectation values mean that we have to take the average over all imaginable random binary sequences $\vec{a}$ of length $n$ (DIY: how many such $\vec{a}$ do exist ?) and over all possible $\vec{c}$.

Consider now two given random binary sequences $\vec{a}$ and $\vec{c}$. We wish to describe how closely they are related to each other. Perhaps

- $\vec{c}$ always behaves as $\vec{a}$ (meaning $\operatorname{dist}(\vec{a}, \vec{c}) = 0$),
- $\vec{c}$ always behaves in the opposite way as $\vec{a}$ (like certain mother-in-laws do), meaning $\operatorname{dist}(\vec{a}, \vec{c}) = n$,
- $\vec{c}$ does not care what $\vec{a}$ does (then $\operatorname{dist}(\vec{a}, \vec{c}) \approx \frac{n}{2}$).

**Definition 6.7 (correlation).** *Given two random binary sequences $\vec{a}$ and $\vec{c}$, we define their* correlation[2] $\Phi(\vec{a}, \vec{c})$ *as*

$$\Phi(\vec{a}, \vec{c}) := \frac{(n - \operatorname{dist}(\vec{a}, \vec{c})) - \operatorname{dist}(\vec{a}, \vec{c})}{n} = 1 - \frac{2}{n}\operatorname{dist}(\vec{a}, \vec{c}).$$

**Remark 6.8.**

- *We read the difference $(n - \operatorname{dist}(\vec{a}, \vec{c})) - \operatorname{dist}(\vec{a}, \vec{c})$ as difference between the number of matches and the number of mismatches between $\vec{c}$ and $\vec{c}$.*
- *The purpose of dividing by $n$ is to normalize the values of $\Phi$ between $-1$ and $+1$. Compare Remark 4.1.*

**Lemma 6.9.** *If $\vec{a}$ if a fixed binary sequence, and the random binary sequence $\vec{c}$ is allowed to depend on chance, then $\Phi(\vec{a}, \vec{c})$ is a random variable with expectation value $0$.*

*Sketch of proof.* We calculate

$$\mathbb{E}\left[\Phi(\operatorname{dist}(\vec{a}, \vec{c}))\right] = \mathbb{E}\left[1 - \frac{2}{n}\operatorname{dist}(\vec{a}, \vec{c})\right] = \mathbb{E}[1] - \frac{2}{n}\mathbb{E}\left[\operatorname{dist}(\vec{a}, \vec{c})\right] = 1 - \frac{2}{n} \cdot \frac{n}{2} = 0.$$

□

In the courses on *Statistics* you will learn to justify this calculation.

Having obtained a little understanding of the correlation of two random binary sequences, we next consider the correlation of a random binary sequence with a shifted copy of itself.

[1] Richard Hamming, 1915–1998

[2] The etymology here is: the prefix *cor-* means *with*, and *relative* comes from the Latin past participle *relatus* of the latin verb *refero* to produce the Latin adjective *relativus* with the meaning *having reference or relation.* It seems that we are running circles here, but you get the gist. . . The Latin verb *refero, referre, retuli, relatus* means *to bring (fero) back (re-)*, and therefore the word *relation* means *a bringing back, a report.*

**Definition 6.10** (**cyclical shift**). *If* $k \in \{0, 1, \dots, n-1\}$ *and* $\vec{c} = (c_0, c_1, \dots, c_{n-1})$, *then we define*

$$\vec{c}_k := (c_k, c_{k+1}, \dots, c_{n-1}, c_0, c_1, \dots, c_{k-1})$$

*and call it a* cyclical shift of $\vec{c}$ by $k$ positions.

**Definition 6.11** (**autocorrelation**). *Given a random binary sequence* $\vec{a}$ *and* $k \in \{0, 1, \dots, n-1\}$, *we define*

$$\phi(\vec{c}, k) := \Phi(\vec{c}, \vec{c}_k),$$

*and all these values (with* $k$ *running) are called the* autocorrelations[3] *of* $\vec{c}$.

**Lemma 6.12.** *If* $\vec{c}$ *is a variable random binary sequence, and* $k$ *is fixed, then*

$$\mathbb{E}\Big[\phi(\vec{c}, k)\Big] = \begin{cases} 0 & : k \neq 0, \\ 1 & : k = 0. \end{cases}$$

We can't give a proof because it requires more time than we have available, so we only give some pointers on what has to be done for a proof. First of all, we have to look up what $\phi$ and $\mathbb{E}$ actually mean: the function $\phi$ is a machine where you put two things in ($\vec{c}$ and $k$), and you get a real number out. Now $k$ has been fixed, and $\vec{c}$ is a random variable that can take $2^n$ distinct values. Therefore also $\phi(\vec{c}, k)$ is then a random variable, and $\mathbb{E}[\phi(\vec{c}, k)]$ is then the expectation value of that random variable. Expectation values are defined above, and we imagine them as weighted averages over all values of the underlying random variable. So, now we have determined *what* we have to calculate, and in the second step we actually have to *do* this calculation, which is a bit lengthy. The key assumption which enters the proof is that $\vec{c}$ is a *random* binary sequence, which means that all the components $c_j$ and $c_i$ are statistically independent provided that $i \neq j$, which leads to a lot of cancellations.

## 6.4 Pseudo–Noise Sequences

Now we have a practical understanding of random binary sequences, and next we look at what the satellite is actually sending over its antennas.

We accept some **input from information theory**[4] which is a scientific discipline in its own right: The sequence to be transmitted over the antenna should be a periodic sequence with a long period $n$, and each segment of length $n$ should be a binary sequence that looks as much as possible like a *random* binary sequence of length $n$, although it is not a random binary sequence since not even one of its bits is actually random. The sequence instead is completely deterministic.

Sequences of that kind are called *pseudo–noise sequences.*

The deeper reason why the sequence to be transmitted should look like random without being random is given by information theory; we will have to skip the details here. We only remark that, in some sense, precious information transport capacity will go wasted if the sequence is "too much non-random".

In the case of GPS, the period of the sequence is 1023. A famous information theorist, Solomon Golomb (1932—) determined properties of a sequence $\vec{g} = (g_0, g_1, \dots, g_{n-1})$ to be admissible as a pseudo–noise sequence. These are as follows (for odd $n$):

- $\frac{n+1}{2}$ of these bits are one, the others are zeros,
- half of the runs must be runs of length one, a quarter of the runs must be runs of length two, an eighth of the runs must be runs of length three, etc. Here a *run of length* $\ell$ is a sequence of $\ell$ consecutive identical bits, surrounded by the opposite bits. For instance, 100001 is a run of length four.
- The autocorrelation values are

$$\phi(\vec{g}, k) = \begin{cases} 1 & : k = 0, \\ -\frac{1}{n} & : k \neq 0. \end{cases}$$

[3] The etymology is a bit hodgepodgy here, because the Greek αὐτο (which means *self, same*) has been mixed with the Latin *correlation*, but the word *correlation* is not even proper Latin; it was only invented by the French around 1600AD.

[4] in the same way we accepted some input from sports when we figured how to throw a ball, compare page 72

The second condition is important because it means that you can observe the sequence for as long as you want, you will never be able to predict the next bit to come with a probability of more than $\frac{1}{2}$.

**DIY:** *Take $n = 7$ and put $\vec{g} = (0100111)$. Determine all the autocorrelation values. Write several copies of $\vec{g}$ one after each other, and count how many runs of the various lengths of the two bits there are.*

Let us call the three conditions listed above *Golomb postulates*[5].

## 6.5 Advanced Algebraic Methods

We need to find a pseudo–noise sequence of length $n = 1023$ that satisfies the three Golomb postulates. In case of $n = 7$, it was quite easy: out of 7 digits, we need four ones, which gives $\binom{7}{4} = 35$ possibilities for a seven-digit sequence that satisfies the first Golomb postulate. Now you only have to check each of these 35 sequences, to find those that satisfy also the other two postulates. You can do this in an afternoon.

But for $n = 1023$, we have $\binom{1023}{512} \approx 2.2406 \cdot 10^{306}$ and a problem. No human or computer can ever work through that many cases. Remember that all the universe contains only about $10^{80}$ particles.

**Advanced algebra will now solve all our problems.**

Let us have a look at human languages. Their vocabulary consists (at least) of

**nouns,** which denote objects,

**adjectives,** which describe nouns more in detail,

**verbs,** which are actions that you can perform upon objects described by the nouns,

**adverbs,** which describe the actions more in detail.

Now look at $\mathbb{R}$. We have

**objects:** these are all the real numbers,

**properties of the real numbers:** a real number could be positive, or it could be irrational. Or one real number could be greater than another real number.

**actions:** you can add two numbers and get again a real number. Similarly for subtraction, multiplication, division. We intentionally omit other operations.

**The four operations $+$, $-$, $\times$, $\div$ have rules:**

**$+$ and $\times$ are commutative:**

$$\forall a \in \mathbb{R}, \quad \forall b \in \mathbb{R}: \quad a + b = b + a$$
$$\forall a \in \mathbb{R}, \quad \forall b \in \mathbb{R}: \quad a \times b = b \times a$$

**$+$ and $\times$ are associative:**

$$\forall a \in \mathbb{R}, \quad \forall b \in \mathbb{R}, \quad \forall c \in \mathbb{R}: \quad (a + b) + c = a + (b + c)$$
$$\forall a \in \mathbb{R}, \quad \forall b \in \mathbb{R}, \quad \forall c \in \mathbb{R}: \quad (a \times b) \times c = a \times (b \times c)$$

**$+$ and $\times$ have a distributive property:**

$$\forall a \in \mathbb{R}, \quad \forall b \in \mathbb{R}, \quad \forall c \in \mathbb{R}: \quad a \times (b + c) = a \times b + a \times c$$

**equations with $+$ can be solved uniquely:** [6]

$$\forall a \in \mathbb{R}, \quad \forall b \in \mathbb{R} \quad \exists! \ x \in \mathbb{R}: \quad a + x = b$$

[5] The Latin verb *postulo, postulare, postulavi, postulatus* means *to request, claim, demand.*

[6] the exclamation mark after the $\exists$ means that there is only one such $x$

**equations with × can be solved uniquely, provided the given factor is non-zero:**

$$\forall a \in \mathbb{R} \setminus \{0\}, \quad \forall b \in \mathbb{R} \quad \exists! \ y \in \mathbb{R}: \quad a \times y = b$$

We introduce the notation $x = b - a$ and $y = b \div a$, for $x$ and $y$ introduced above.

In abstract[7] algebra, this long list will be compressed into the sentence "$(\mathbb{R}, +\times)$ is a field".

The interesting step is now that we obtain another field if we replace the set $\mathbb{R}$ by the set $\{0, 1\}$ consisting of only two elements, and change the addition slightly. Namely, we define addition and multiplication as follows:

| + | 0 | 1 |
|---|---|---|
| 0 | 0 | 1 |
| 1 | 1 | 0 |

| × | 0 | 1 |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 0 | 1 |

We calculate almost as usual, and the only change is that now $1 + 1 = 0$ instead of $= 2$.

Subtraction and division are then defined as inverse operations to addition and multiplication, as it is always done when some algebraic structure wants to become an algebraic field.

**Lemma 6.13.** *The set $\{0, 1\}$ together with these two operations $+$ and $\times$ is also a field.*

*Sketch of proof.* Just check all the rules (commutativity etc.) for all possible values of $a$ and of $b$ and of $c$. □

This field is called $\mathbb{F}_2$ or $GF(2)$, which stands for Galois field, named after Évariste Galois (1811–1832).

Next we work with polynomials, and all the coefficients are now from $\mathbb{F}_2$. Examples of such polynomials are

$$x^2 + 1, \quad x^2 + x + 1, \quad x^2 - 1, \quad x^3 + x - 1.$$

Note that $1 + 1 = 0$ implies (subtract one from both sides) that $1 = -1$, and therefore $x^2 + 1$ and $x^2 - 1$ are in fact the same polynomial. We can also multiply polynomials:

$$(x + 1)^2 = (x + 1) \times (x + 1) = x^2 + x + x + 1 = x^2 + (1 + 1) \times x + 1 = x^2 + 1.$$

Observe that $\mathbb{F}_2$ is a field which makes children happy because now the equation $(a+b)^2 = a^2 + b^2$ becomes true, called *binomial formula*, named after *Ernesto Binomi (121–1331)*.

We can also divide one polynomial by another polynomial, perhaps with a remainder. Keep in mind that the coefficients of all polynomials here are either 0 or 1, and all the calculations with numbers are to be done according to the two tables above. What you need here is that $(\mathbb{F}_2, +, \times)$ is a field in the same right as $(\mathbb{R}, +, \times)$ is a field, which means that the "adverbs in the language $\mathbb{F}_2$" are the same as the "adverbs in the language $\mathbb{R}$".

[7] the word *abstract* comes from the Latin verb *traho, trahere, traxi, tractus* which means *to drag*. You know the noun *tractor* as *that guy who pulls*. And the prefix *ab-* here means *away*, and altogether *abstract* therefore means *having been pulled away (from everyday experience)*. That is the reason why abstract stuff is hard stuff.

**Here comes something magical.**

Take the polynomial $Q(x) = x^{10} + x^9 + x^8 + x^6 + x^3 + x^2 + 1$. Next take a sequence of polynomials

$$A_0(x) = 1, \quad A_1(x) = x, \quad A_2(x) = x^2, \quad \ldots, \quad A_{1022}(x) = x^{1022}$$

and then perform (for each $j$) a long polynomial division $A_j(x) \div Q(x)$, which will give you a remainder $R_j(x)$ that is a polynomial of degree at most 9, and all the coefficients of all polynomials here are from $\mathbb{F}_2$. All the calculations are to be done in $\mathbb{F}_2$.

Next we define

$$\begin{array}{rl} g_0 & \text{as the lowest order coefficient of the remainder } R_0(x), \\ g_1 & \text{as the lowest order coefficient of the remainder } R_1(x), \\ g_2 & \text{as the lowest order coefficient of the remainder } R_2(x), \\ \ldots & \\ g_{1022} & \text{as the lowest order coefficient of the remainder } R_{1022}(x). \end{array}$$

Then $(g_0, g_1, \ldots, g_{1022})$ is a sequence that satisfies the three Golomb postulates.

Take a break and let that sink in. You have just now observed the magical power of abstract algebra.

To give a definite statement:

**Theorem 6.14.** *Let $Q = Q(x)$ be a polynomial of degree $N$ with coefficients in $\mathbb{F}_2$, and assume that it is impossible to write $Q$ as a product $Q(x) = U(x) \cdot V(x)$ of polynomials $U$ and $V$ with degrees strictly less than $N$ and coefficients in $\mathbb{F}_2$.*

*Then the sequence $(g_0, g_1, \ldots, g_n)$ obtained as the lowest order coefficients $g_j$ of the remainders by polynomial long division of $x^j$ by $Q(x)$ satisfies the three Golomb postulates with $n = 2^N - 1$.*

The proof requires elaborate techniques that will be developed in the courses on *Abstract Algebra*.

## 6.6 Navigation in 1D

Now we have found a pseudo–noise sequence $g$ of length 1023. We also have two satellites (called $A$ and $B$), and each of them gets a cyclically shifted copy $g_A$, $g_B$, respectively, of $g$. The shift length is known.

Now we intend to determine the position of the user $U$ relative to the satellites $A$ and $B$, and the positions of $A$ and of $B$ are known, and we know that $U$ is between $A$ and $B$.

We need to distinguish two different procedures of adding numbers:

- in information science and abstract algebra, we only possess the numbers 0 and 1, and we add them according to the rule $1 + 1 = 0$.
- in physics and electronics, where you are dealing with voltages and electrical currents, any real number is admissible for the strength of an electrical current, and of course you add them as numbers in $\mathbb{R}$.

We translate from the abstract algebra world into the electrical world by the map

$$T_{aa\to e} \colon s \mapsto (-1)^s, \qquad s \in \{0, 1\}.$$

Here "$aa \to e$" abbreviates "abstract algebra → electrical". How does this translation affect the correlations ?

**Lemma 6.15.** *If $\vec{a}$ and $\vec{c}$ are binary sequences of length $n$ (with digits 0 and 1), then their correlation is*

$$\Phi(\vec{a}, \vec{c}) = \frac{1}{n} \sum_{j=1}^{n} T_{aa\to e}(a_j) \cdot T_{aa\to e}(c_j).$$

**DIY:** *Prove this.*

In what follows, we will perform all the calculations in the electrical world, add as in $\mathbb{R}$, and tacitly omit the translation map $T_{aa\to e}$. The correlation of two sequences $\vec{a}$ and $\vec{c}$ (with digits as real numbers) then is (according to the lemma proven by you)

$$\Psi(\vec{a}, \vec{c}) := \frac{1}{n} \sum_{j=1}^{n} a_j c_j. \tag{6.1}$$

Let us now look at two satellites $A$, $B$, and a user $U$ between them. $A$, $U$, $B$ are (in this order) on the same straight line.

- $U$ receives a sequence (periodically repeated) from $A$,
- $U$ receives another sequence (periodically repeated) from $B$,
- $U$ receives noise coming from wherever.

Both satellite sequences start from $A$, $B$ at the same time, and they arrive at $U$ at different times since $U$ is typically not in the middle of the line segment $\overline{AB}$.

As an example, $U$ receives the sequence

$$\vec{r} = \Big(r_0, r_1, \ldots, r_{1023}, r_{1024}, \ldots\Big)$$

with $r_i = a_i + b_i + n_i$ being a real number, where

- $a_i$ is an unknown bit coming from $A$,
- $b_i$ is an unknown bit coming from $B$,
- $n_i$ is an unknown noise bit.

The user knows $r_i$, but neither $a_i$, $b_i$, $n_i$. The bits $a_i$ and $b_i$ arrive at the same time at the user $U$, but they started from the satellites at different times because the distances $AU$ and $BU$ are different.

Now the user's SatNav device does the following:

- For each $i$, the device correlates $\vec{g}_A$ (a known 1023 bit sequence) with $(r_i, r_{i+1}, \ldots, r_{i+1022})$, and remembers all those $i$ for which a huge correlation occurs.
- For each $i$, the device correlates $\vec{g}_B$ (a known 1023 bit sequence) with $(r_i, r_{i+1}, \ldots, r_{i+1022})$, and remembers all those $i$ for which a huge correlation occurs.
- Then the device determines how many digit positions are between the peaks of the correlation with $\vec{g}_A$ and with $\vec{g}_B$.
- This way, the user can determine the run-time difference between the $A$-signal stream and the $B$-signal stream.

We practise this with some `python` code: instead of $n = 1023$, we take $n = 127$. The sequence $\vec{g}_B$ has been shifted by 37 positions from $\vec{g}_A$. The $B$-signals arrive 10 clock ticks later than the $A$-signal. The noise signal consists of 85% zeros, 10% added ones, 5% subtracted ones, which means that we are damaging 15% of the bits. The python code is now the following. We should remark that the line `lfsr('0000001',(7,1))` calls the procedure `lfsr` declared at the top, and this procedure `lfsr` then produces the sequence $\vec{g}_A$ (in an absolutely marvellous by means of repeated divisions of polynomials with coefficients from $\mathbb{F}_2$). In the python implementation of the correlation function $\Psi$ (called here `scalarproduct`), we neglected the division by $n = 127$ because otherwise the program output would become much longer than it already is. We define “peak correlation” as a correlation value bigger than 100.

```
###########################################################

def lfsr(seed, taps): # lfsr means linear feedback shift register
    sr = seed
    lfsrlist = [ ]
    while 1:
        temp = 0
        for t in taps:
            temp += int(sr[t-1])
        xor = temp % 2
        lfsrlist.append(1 - 2 * xor) # this (1-2*xor) translates into electrics
        sr = str(xor) + sr[:-1]
        if sr == seed:
            break
    return lfsrlist
############################################################

def scalarproduct(ulist, vlist): # ulist and vlist must have same length
    sum = 0.0
    for i in range(len(ulist)):
        sum += ulist[i] * vlist[i]
    return sum

############################################################

# create sequence for satellite a

lfsrlista = lfsr(’0000001’, (7,1))      # polynomial is x^7 + x + 1
print " "
print "The transmitted list of satellite A is ", lfsrlista
print "and it has length ", len(lfsrlista)
print " "

#############################################################

# create sequence for satellite b by a shift of 37 positions

lfsrlistb = lfsrlista[37:] + lfsrlista[:37]

print "The transmitted list of satellite B is ", lfsrlistb
print "and it has length ", len(lfsrlistb)
print " "


#############################################################

# concatenate four copies of the satellite lists

transmitlista = 4 * lfsrlista
transmitlistb = 4 * lfsrlistb

# now shift the B list by 10 digits, since the B signal comes
# 10 clock ticks later

transmitlistb = transmitlistb[10:] + transmitlistb[:10]

##############################################################

# create noise list. We are damaging 15 percent of all bits

import random

samplelist = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, -1]

noiselist = [0] * len(transmitlista)
for i in range(len(transmitlista)):
    noiselist[i] = random.choice(samplelist)
print "The noise list is ", noiselist
print " "

################################################################

# now add it all up
```

```
receiverlist = [0] * len(transmitlista)
for i in range(len(transmitlista)):
    receiverlist[i] = transmitlista[i] + transmitlistb[i] + noiselist[i]

print "The SatNav receives this list: ", receiverlist
print " "

################################################################

# now calculate all the correlations

correlationlista = [0] * (len(transmitlista) - len(lfsrlista))
for i in range(len(correlationlista)):
    slicelist = receiverlist[i:(i+len(lfsrlista))]
    correlationlista[i] = scalarproduct(slicelist, lfsrlista)
print "The correlation with the satellite A is ", correlationlista
print " "
maxalist = []
for i in range(len(correlationlista)):
    if correlationlista[i] > 100.0:
       maxalist.append(i)
print "The peaks of the correlations with A are at ", maxalist
print " "

correlationlistb = [0] * (len(transmitlistb) - len(lfsrlistb))
for i in range(len(correlationlistb)):
    slicelist = receiverlist[i:(i+len(lfsrlistb))]
    correlationlistb[i] = scalarproduct(slicelist, lfsrlistb)
print "The correlation with the satellite B is ", correlationlistb
print " "
maxblist = []
for i in range(len(correlationlistb)):
    if correlationlistb[i] > 100.0:
       maxblist.append(i)
print "The peaks of the correlations with B are at ", maxblist
print " "
```

---

We call this program `satnavdemo.py`. We run it on a command-line as `python satnavdemo.py` and obtain the output as follows:

---

```
The transmitted list of satellite A is
[-1, -1, -1, -1, -1, -1, -1, 1, -1, 1, -1, 1, -1, 1, 1, -1, -1, 1, 1, -1,
-1, -1, 1, -1, -1, -1, 1, -1, 1, 1, -1, 1, -1, -1, 1, 1, 1, -1, -1, 1, -1,
-1, -1, -1, 1, -1, -1, 1, -1, 1, -1, -1, 1, -1, -1, 1, 1, -1, 1, 1, -1, 1,
1, 1, -1, -1, -1, 1, 1, 1, 1, -1, 1, -1, -1, -1, -1, -1, 1, 1, -1, 1, -1,
1, -1, -1, -1, 1, 1, -1, -1, 1, -1, 1, 1, 1, -1, 1, 1, -1, -1, -1, -1, 1,
1, 1, -1, 1, -1, 1, 1, 1, 1, -1, -1, 1, 1, 1, 1, 1, -1, 1, 1, 1, 1, 1, 1]
and it has length  127

The transmitted list of satellite B is
[-1, -1, 1, -1, -1, -1, -1, 1, -1, -1, 1, -1, 1, -1, -1, 1, -1, -1, 1, 1,
-1, 1, 1, -1, 1, 1, 1, -1, -1, -1, 1, 1, 1, 1, -1, 1, -1, -1, -1, -1, -1,
1, 1, -1, 1, -1, 1, -1, -1, -1, 1, 1, -1, -1, 1, -1, 1, 1, 1, -1, 1, 1, -1,
-1, -1, -1, 1, 1, 1, -1, 1, -1, 1, 1, 1, 1, -1, -1, 1, 1, 1, 1, 1, -1, 1,
1, 1, 1, 1, 1, -1, -1, -1, -1, -1, -1, -1, 1, -1, 1, -1, 1, -1, 1, 1, -1,
-1, 1, 1, -1, -1, -1, 1, -1, -1, -1, 1, -1, 1, 1, -1, 1, -1, -1, 1, 1, 1]
and it has length  127

The noise list is  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0,
0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, -1,
0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, -1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1, 1,
-1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 1, 0, 1, 0, -1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0,
0, 1, 0, -1, -1, 0, 0, 1, 0, 0, 1, -1, -1, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
0, 0, -1, 0, -1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, -1, 0, 0, 0, 0, 0, 0, -1, 0,
0, -1, 0, 0, 0, 0, -1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1, 0,
0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, -1, 0, 1, 0, 1, -1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1, 0]

The SatNav receives this list:  [0, -2, 0, -2, -2, 0, -2, 0, 0, 2, -2, 2,
0, 0, 2, 0, 0, 0, 0, -2, 0, 0, 2, 0, -2, 0, 0, -2, 0, 0, -2, 2, 0, -2, 3,
1, 2, -2, -1, 0, 0, 0, -2, -2, 2, -2, 1, 2, 0, 1, 0, 0, 0, -1, -1, 0, 3, 0,
2, 1, 0, 0, 3, 2, 0, 0, -2, 0, 2, 2, 2, 0, 2, -2, 0, 0, 0, 0, 2, 2, -2, 0,
-2, 0, -2, -2, -2, 2, 0, 0, -2, 2, -2, 2, 2, 0, -2, 2, 2, -2, -2, -1, 0, 0,
0, 0, 0, 0, 0, 2, 0, 2, 0, -2, 0, 2, 3, 0, 0, 2, -2, 1, 0, 0, 2, 0, 0, 0,
-1, 0, -2, -2, 0, -1, 0, 0, 2, -1, 2, 0, 0, 2, -1, 0, 0, 0, -2, 0, -1, 2,
0, -2, 0, 1, -1, 0, 0, -2, 2, 0, -2, 2, 0, 2, -2, -2, 0, 0, 0, -2, -2, 1,
-2, 1, 2, 1, 0, 0, 0, 0, -2, -2, 0, 2, 0, 2, 0, 0, 0, 2, 1, 1, -1, -1, 0,
2, 2, 2, 0, 2, -2, 0, 0, 1, 0, 2, 2, -2, 0, -2, 0, -2, -1, -2, 2, 1, 0, -2,
2, -2, 2, 2, 0, -2, 2, 2, -2, -2, -2, 0, 0, 0, 0, 0, 0, 0, 2, 0, 2, 0, -2,
0, 3, 3, 0, 1, 2, -3, 0, 1, 0, 2, 0, 0, 0, -2, 0, -2, -2, 0, -2, 0, 0, 2,
-2, 2, 0, 0, 2, 0, 0, 0, 1, -2, 0, 0, 2, 0, -2, 0, 0, -2, 0, 0, -2, 2, 0,
-2, 2, 0, 2, -2, -2, 0, 0, 0, -2, -2, 3, -2, 0, 2, 0, 0, 0, 0, 0, -2, -2,
0, 2, -1, 2, 0, 0, 0, 2, 2, 1, 0, -3, -1, 2, 2, 3, 0, 2, -1, -1, -1, 0, 0,
1, 2, -2, 0, -2, 0, -2, -2, -2, 2, 0, 0, -2, 2, -2, 2, 2, 0, -2, 2, 2, -2,
-2, -2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 2, 0, -3, 0, 1, 3, 0, 0, 2, -2, 1, 0, 0,
3, 0, 0, 0, -2, 0, -2, -2, 0, -2, 0, 0, 2, -3, 2, 0, 0, 2, 0, 0, 0, 0, -2,
0, 0, 2, 0, -1, 0, 0, -2, 0, 0, -2, 2, 0, -2, 2, 0, 2, -2, -2, 0, 1, 0, -1,
-2, 2, -1, 1, 2, -1, 0, 0, 0, 0, -2, -2, -1, 2, 0, 1, 0, 0, 0, 2, 1, 0, 0,
-2, 0, 3, 3, 2, 0, 2, -2, 0, 0, 0, 0, 2, 1, -2, 0, -2, 0, -2, -2, -2, 3, 0,
0, -2, 2, -2, 3, 2, 0, -3, 2, 3, -2, -1, -3, 0, 0, 0, 0, 0, 0, 0, 2, 0, 2,
0, -2, 0, 2, 2, 0, 0, 2, -2, 0, 0, 0, 2, -1, 0]

The correlation with the satellite A is  [129.0, -3.0, -4.0, -2.0, 4.0,
-4.0, 0.0, 5.0, -11.0, -1.0, 3.0, -6.0, 2.0, -4.0, -6.0, -4.0, -7.0, -3.0,
5.0, -9.0, -1.0, -1.0, -4.0, 0.0, -4.0, -2.0, -8.0, -3.0, 6.0, -6.0, -6.0,
-4.0, -8.0, -4.0, -4.0, -3.0, -4.0, -6.0, -4.0, -5.0, 3.0, -9.0, 5.0, -3.0,
-7.0, -4.0, -2.0, -6.0, -8.0, 3.0, -2.0, -8.0, -4.0, -10.0, -5.0, -8.0,
-2.0, -3.0, 3.0, -3.0, -2.0, -6.0, -4.0, -1.0, -6.0, 3.0, 4.0, -3.0, -5.0,
-1.0, 3.0, -5.0, 5.0, -9.0, -5.0, 1.0, -3.0, -2.0, -4.0, -4.0, 118.0, 0.0,
6.0, -12.0, 4.0, -2.0, 1.0, -3.0, 1.0, -6.0, 0.0, -2.0, 0.0, -4.0, 8.0,
-6.0, -2.0, -2.0, 2.0, -4.0, 0.0, -10.0, -1.0, -1.0, -5.0, 5.0, 3.0, -3.0,
-1.0, -9.0, 3.0, -7.0, -3.0, -15.0, -3.0, 3.0, -8.0, 2.0, 0.0, 5.0, -3.0,
2.0, -7.0, -2.0, 6.0, -2.0, 4.0, 124.0, -2.0, -11.0, -5.0, -5.0, -3.0,
-1.0, 0.0, -8.0, 0.0, -6.0, -1.0, 3.0, 1.0, -3.0, 1.0, -12.0, 6.0, 2.0,
1.0, 7.0, -1.0, -4.0, -6.0, 0.0, -4.0, -10.0, -1.0, 0.0, 0.0, 0.0, -6.0,
-4.0, -2.0, -6.0, 2.0, -2.0, -4.0, 0.0, -4.0, -2.0, -2.0, -6.0, -8.0, -2.0,
-8.0, -2.0, -1.0, -1.0, 0.0, 0.0, -6.0, -4.0, 2.0, 2.0, 2.0, -6.0, 4.0,
-9.0, 3.0, 1.0, -7.0, -3.0, -1.0, -4.0, -10.0, 1.0, 1.0, -8.0, -6.0, -6.0,
-5.0, -7.0, -5.0, 0.0, -5.0, -4.0, -7.0, -3.0, -4.0, 124.0, -4.0, -2.0,
-6.0, -6.0, -8.0, -7.0, 1.0, 3.0, -8.0, -4.0, -10.0, -4.0, 4.0, -2.0, -8.0,
-4.0, -6.0, -6.0, -4.0, -2.0, -2.0, 0.0, -2.0, 4.0, -2.0, -6.0, -8.0, -2.0,
6.0, -1.0, -5.0, -3.0, -1.0, -10.0, -8.0, -4.0, -2.0, -4.0, 7.0, -1.0,
-4.0, -3.0, -8.0, -4.0, -3.0, -5.0, 133.0, 5.0, 1.0, -1.0, -3.0, 1.0, -3.0,
-1.0, -9.0, 5.0, -7.0, -2.0, 2.0, -4.0, 2.0, 0.0, -6.0, -2.0, -2.0, -9.0,
-9.0, 3.0, -9.0, -3.0, 1.0, -2.0, 2.0, -2.0, -2.0, 4.0, 2.0, -4.0, -10.0,
6.0, -2.0, 0.0, 0.0, 4.0, 0.0, -6.0, -6.0, 1.0, 7.0, -8.0, -4.0, -3.0, 0.0,
5.0, 3.0, -6.0, -4.0, 8.0, -4.0, 2.0, 2.0, 6.0, -3.0, 3.0, -4.0, -7.0,
-1.0, -5.0, 9.0, 1.0, -4.0, -3.0, -5.0, -10.0, -7.0, 0.0, -1.0, -8.0, 6.0,
0.0, 3.0, 0.0, -3.0, 1.0, -5.0, -4.0, 127.0, -5.0, -5.0, -3.0, -5.0, -1.0,
3.0, -11.0, -2.0, 0.0, 4.0, -2.0, 2.0, 0.0, -7.0, 1.0, -5.0, -8.0, 2.0,
1.0, 1.0, -4.0, 1.0, -1.0, 1.0, 3.0, -9.0, -1.0, -1.0, 1.0, 8.0, 0.0, -6.0,
-6.0, -1.0, 1.0, -2.0, 7.0, -7.0, -17.0, -3.0, -7.0, 0.0, -6.0, -8.0, -1.0,
-6.0]

The peaks of the correlations with A are at  [0, 80, 127, 207, 254, 334]

The correlation with the satellite B is  [-1.0, -9.0, 4.0, 2.0, 0.0, 0.0,
-10.0, -7.0, -1.0, -1.0, -7.0, -8.0, 4.0, 2.0, -2.0, -4.0, -1.0, -3.0,
-7.0, -3.0, -1.0, -1.0, -6.0, -6.0, -4.0, -2.0, -8.0, 1.0, 4.0, 2.0, -6.0,
0.0, -2.0, -4.0, 2.0, -1.0, 0.0, 126.0, -4.0, -7.0, -5.0, -1.0, -3.0, -5.0,
1.0, -14.0, 0.0, 2.0, -2.0, 3.0, 0.0, -2.0, -2.0, -8.0, -1.0, 2.0, -4.0,
-1.0, -5.0, -5.0, -2.0, -6.0, -8.0, -7.0, -10.0, 5.0, 0.0, 1.0, -5.0, -3.0,
1.0, -1.0, 3.0, -5.0, -5.0, 3.0, -3.0, 4.0, -6.0, 6.0, -4.0, -4.0, -8.0,
-2.0, -4.0, -6.0, 5.0, 3.0, -5.0, -4.0, -4.0, -2.0, -6.0, -4.0, 4.0, 2.0,
4.0, -2.0, -8.0, -2.0, 0.0, -10.0, 1.0, 1.0, -3.0, -5.0, -1.0, 1.0, -3.0,
3.0, -13.0, -3.0, -1.0, -3.0, -1.0, -3.0, -4.0, 118.0, 2.0, 5.0, -15.0,
2.0, -7.0, 0.0, -2.0, 0.0, -4.0, -2.0, -2.0, -5.0, -3.0, 5.0, -7.0, 3.0,
-2.0, 2.0, -8.0, 4.0, -5.0, -5.0, -5.0, -1.0, 9.0, 2.0, -4.0, 2.0, -7.0,
1.0, -7.0, -4.0, -12.0, -2.0, -2.0, -6.0, 3.0, -2.0, 2.0, -2.0, 2.0, -8.0,
-2.0, 4.0, -2.0, 2.0, 126.0, 0.0, -8.0, -4.0, 0.0, -2.0, 2.0, 4.0, -4.0,
0.0, -5.0, -3.0, -2.0, -2.0, -6.0, 0.0, -10.0, 0.0, 2.0, -2.0, 2.0, -1.0,
-1.0, -5.0, 1.0, -1.0, -7.0, 6.0, -2.0, -1.0, 1.0, -8.0, -8.0, 0.0, -1.0,
1.0, 1.0, -2.0, -5.0, -8.0, -11.0, -3.0, -2.0, -8.0, -4.0, -4.0, 0.0, 2.0,
-2.0, -5.0, -3.0, 1.0, -2.0, 0.0, 4.0, 6.0, 0.0, 0.0, -8.0, 2.0, 0.0, -6.0,
-2.0, 0.0, 0.0, -10.0, 6.0, -2.0, -8.0, -6.0, -4.0, -6.0, -4.0, 1.0, 3.0,
-1.0, -3.0, -8.0, -2.0, -8.0, 126.0, -4.0, -1.0, -1.0, -8.0, -3.0, -6.0,
-2.0, 5.0, -9.0, -1.0, -9.0, 1.0, 3.0, -1.0, -3.0, -5.0, -9.0, -3.0, -3.0,
-5.0, -8.0, 0.0, 2.0, -2.0, -4.0, -8.0, -2.0, -6.0, 3.0, 5.0, -5.0, 3.0,
5.0, -7.0, -6.0, -4.0, 0.0, -4.0, 2.0, -2.0, -4.0, 0.0, -10.0, -6.0, 0.0,
-6.0, 132.0, 2.0, 4.0, -2.0, -3.0, 1.0, 2.0, 0.0, -9.0, 4.0, -5.0, 3.0,
2.0, -6.0, 4.0, 0.0, -6.0, 0.0, -2.0, -13.0, -11.0, 0.0, -13.0, -1.0, -1.0,
-1.0, -1.0, -6.0, 3.0, 5.0, -4.0, 3.0, -6.0, 9.0, -4.0, -2.0, -4.0, -1.0,
2.0, -3.0, 1.0, -5.0, 0.0, -3.0, -7.0, -9.0, 1.0, 5.0, 3.0, -3.0, -3.0,
4.0, 0.0, 2.0, 2.0, 6.0, -2.0, 3.0, 3.0, -7.0, 0.0, -4.0, 7.0, -1.0, -2.0,
-5.0, -13.0, -11.0, -13.0, 5.0, -5.0, -5.0, 3.0, -4.0, 0.0, -4.0, 2.0, 1.0,
-5.0, -4.0, 129.0, -7.0, -5.0, -9.0, -1.0, -4.0, 6.0, -14.0, -7.0, 2.0]

The peaks of the correlations with B are at  [37, 117, 164, 244, 291, 371]
```

We observe that we are indeed able to determine that the peaks of $A$ and $B$ are 10 digits apart, which can be seen from the following diagram.

$$\begin{tikzcd}
0 \arrow[d, "37"] & 80 \arrow[d, "37"] & 127 \arrow[d, "37"] & 207 \arrow[d, "37"] & 254 \arrow[d, "37"] & 334 \arrow[d, "37"] \\
37 & 117 \arrow[ur, "10"] & 164 & 244 \arrow[ur, "10"] & 291 & 371
\end{tikzcd}$$

You may wonder whether there is a bit of ambiguity in this diagram, and why do we have so many peaks in the correlation (we concatenated four copies of the sequence $\vec{g}_A$, but we observe 6 peaks for the $A$–correlations, which is more than we expected). There is some truth in this observation, and in reality, the GPS system overcomes this trouble by putting more abstract algebra in — they do not take just one polynomial of degree 10, but two such polynomials: $Q_1(x) = x^{10} + x^9 + x^8 + x^6 + x^3 + x^2 + 1$ and $Q_2(x) = x^{10} + x^3 + 1$. Let $\vec{g}_1$ be the sequence of zeros and ones associated to $Q_1$, and $\vec{g}_2$ be the sequence of zeros and ones associated to $Q_2$. Both have 1023 digits. Then you add (in $\mathbb{F}_2$) a shifted copy of $\vec{g}_1$ to an unshifted copy of $\vec{g}_2$, to obtain the sequence $\vec{g}_A$. And $\vec{g}_B$ is obtained similarly: an unshifted version of $\vec{g}_2$ is added (in $\mathbb{F}_2$) to a shifted copy of $\vec{g}_1$. For $\vec{g}_B$, you shift $\vec{g}_1$ by another length than for $\vec{g}_A$.

The sequences $\vec{g}_A$ and $\vec{g}_B$ are the famous Gold codes, named after ROBERT GOLD[8].

Now the calculation is this: we know the satellites positions $x_A$ and $x_B$, and we know the signals' arrival times $t_A$ and $t_B$. We do not know the common start time $t_0$ of the signals at $A$ and at $B$. Now the signal from $A$ travels the distance $x_U - x_A$ in the time duration $t_A - t_0$, and the signal from $B$ travels the distance $x_B - x_U$ in the time duration $t_B - t_0$. With $c$ as light speed, we then have

$$x_U - x_A = c \cdot (t_A - t_0), \qquad x_B - x_U = c \cdot (t_B - t_0).$$

[8] and you can find the details in the journal article: Robert Gold, *Optimal binary sequences for spread spectrum multiplexing*, IEEE Transactions on Information Theory, Volume 13, number 4, pages 619–621, 1967.

We subtract and then quickly get

$$x_U = \frac{x_A + x_B}{2} + \frac{c}{2} \cdot (t_A - t_B).$$

Everything on the right-hand side is known, hence we have determined $x_U$.

We conclude the considerations of how to navigate in 1D with a remark about how the satellites tell the user's devices about the satellite's position (in our example $x_A$ and $x_B$). These data are called *ephemeris* and *almanac* data, and the GPS satellites transmit these data with a rate of 50 bits per second, and these 50 bits are hidden in the 1023 × 1000 Bits per second, roughly as follows: if the satellite wants to transmit a zero-bit of the ephemeris data, then it sends all the bits as usual, for 20 milliseconds. But if the satellite wishes to transmit a one-bit of the ephemeris (or almanac) data, then it flips all the bits for 20 milliseconds (this flip affects 20 × 1023 bits). The SatNav device recognizes during these 20 milliseconds not 20 correlation peaks of $+1$, but 18 correlation peaks of $-1$ (one peak is lost at the start of this 20 millisecond period, and another one at the end of the 20 milliseconds). Note that $\frac{1s}{20ms} = 50$, hence we can indeed fit 50 extra bits into one second, but keep in mind that we have glossed over a lot of details; the actual implementation is much more complicated.

## 6.7 Navigation in 2D

Given are three satellites $A$, $B$, $C$ in $\mathbb{R}^2$, with their known positions $(x_A, y_A) \in \mathbb{R}^2$, $(x_B, y_B) \in \mathbb{R}^2$, $(x_C, y_C) \in \mathbb{R}^2$.

Wanted is the position $(x, y) \in \mathbb{R}^2$ of the user $U$.

The user's device has a clock that is not in sync with the satellites' clocks. We know the arrival times $t_A$, $t_B$, $t_C$ of the satellites' signals at $U$. These numbers $t_A$, $t_B$, $t_C$ refer to the user's clock. The three signals started at $A$, $B$, $C$ simultaneously at an unknown time $t_0$. The light speed is $c$.

Since the travelled distance equals velocity times travel time, we have

$$\begin{aligned}
\sqrt{(x - x_A)^2 + (y - y_A)^2} &= c(t_A - t_0), \\
\sqrt{(x - x_B)^2 + (y - y_B)^2} &= c(t_B - t_0), \\
\sqrt{(x - x_C)^2 + (y - y_C)^2} &= c(t_C - t_0),
\end{aligned}$$

which are three equations for three unknowns $(x, y, t_0)$. The trouble now is that these equations are *non-linear*, due to the squares and the roots. Because of the non-linearity, there is no easy solution formula, and we can only hope for approximate solutions, and we have no guarantee that our solution method actually works in a concrete situation.

We begin with an obvious simplification by twice subtraction,

$$\begin{aligned}
\sqrt{(x - x_A)^2 + (y - y_A)^2} - \sqrt{(x - x_B)^2 + (y - y_B)^2} &= c(t_A - t_B), \\
\sqrt{(x - x_C)^2 + (y - y_C)^2} - \sqrt{(x - x_B)^2 + (y - y_B)^2} &= c(t_C - t_B),
\end{aligned}$$

which are two equations for two unknowns $(x, y)$. We define a function $\vec{f}(x, y)$ with

$$\vec{f} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

and

$$\begin{aligned}
f_1(x, y) &= \sqrt{(x - x_A)^2 + (y - y_A)^2} - \sqrt{(x - x_B)^2 + (y - y_B)^2} - c(t_A - t_B), \\
f_2(x, y) &= \sqrt{(x - x_C)^2 + (y - y_C)^2} - \sqrt{(x - x_B)^2 + (y - y_B)^2} - c(t_C - t_B);
\end{aligned}$$

and we want to find $(x, y) \in \mathbb{R}^2$ with the property that

$$\vec{f}(x, y) \overset{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

**Mathematical method for solving $\vec{f}(x,y) = \binom{0}{0}$ (Newton–Raphson method)**

**Given:** a function $\vec{f} = \binom{f_1(x,y)}{f_2(x,y)}$ and a starting point $\binom{x_{old}}{y_{old}}$

**Wanted:** two real numbers $x_*$ and $y_*$ with $\vec{f}(x_*, y_*) = \binom{0}{0}$

**Solution:** calculate a new point

$$\begin{pmatrix} x_{new} \\ y_{new} \end{pmatrix} = \begin{pmatrix} x_{old} \\ y_{old} \end{pmatrix} - J^{-1} \cdot \begin{pmatrix} f_1(x_{old}, y_{old}) \\ f_2(x_{old}, y_{old}) \end{pmatrix},$$

where $J^{-1}$ is the inverse matrix of the Jacobi matrix $J$, and $J$ is a $2 \times 2$ matrix containing the partial derivatives:

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial x}(x_{old}, y_{old}) & \frac{\partial f_1}{\partial y}(x_{old}, y_{old}) \\ \frac{\partial f_2}{\partial x}(x_{old}, y_{old}) & \frac{\partial f_2}{\partial y}(x_{old}, y_{old}) \end{pmatrix}.$$

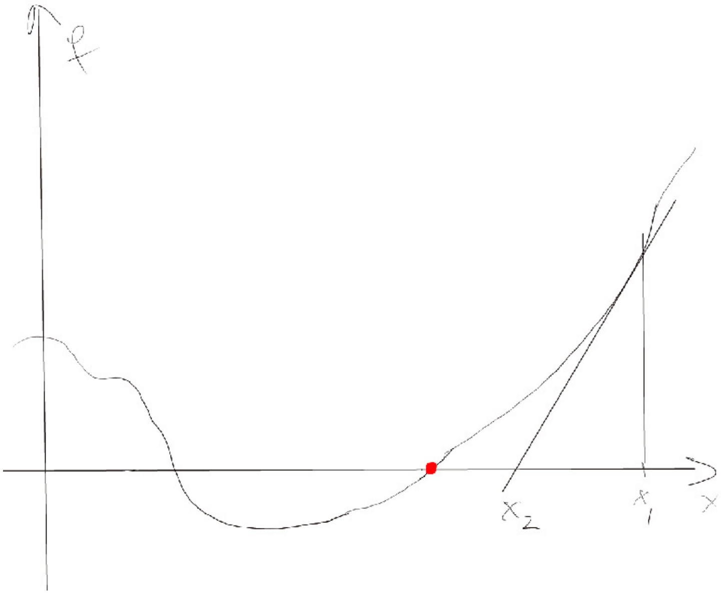
The new point $\binom{x_{new}}{y_{new}}$ should be more close to the solution $\binom{x_*}{y_*}$ than $\binom{x_{old}}{y_{old}}$, but this improvement is not guaranteed.

Lather, rinse, repeat.

**Remark 6.16.** *The matrix $J$ is called* Jacobi matrix *(after* Carl Gustav Jacob Jacobi*, 1804–1851). This matrix is the multi-dimensional analogue of the one-dimensional derivative known from school.*

**DIY:**

- *Please look at the picture below. Given is a complicated function $f = f(x)$, and a fairy godmother has given you a guessed value $x_1$. We want to find the red point. The geometrical idea is: from the point $(x_1, 0)$, you go upwards until you hit the graph of $f$. Then you go downwards along the tangent line, until you hit the axis $y = 0$. That is then your new guessed value $x_2$. We hope that $x_2$ is nearer to the red point than $x_1$. Repeat this and obtain an $x_3$. Etc. Etc.*
- *Now without geometry: give a formula that calculates $x_2$ from $f$, $f'$, and $x_1$. Then $x_3$ from $f$, $f'$, and $x_2$. Etc.*
- *Practise this with $f(x) = x^5 - 10x + 2$ and $x_1 = 2$.*



We practise now the 2D navigation using a `python` program with

$$A = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad B = \begin{pmatrix} 9 \\ 2 \end{pmatrix}, \quad C = \begin{pmatrix} 2 \\ 8 \end{pmatrix}.$$

We choose the light speed $c = 2$, and the runtime differences are $t_A - t_B = 0.1$ and $t_C - t_B = 0.2$. The start-point is $\binom{4}{3}$, and our python code is here:

```
import scipy as SP

##############################################################

def f(x,y):
    xa, ya = 1, 1 # position of satellite A
    xb, yb = 9, 2 # position of satellite B
    xc, yc = 2, 8 # position of satellite C
    c = 2.0       # velocity of light
    tb = 73       # time of arrival of signal from B
    ta = 73.1     # time of arrival of signal from A
    tc = 73.2     # time of arrival of signal from C

    f1 = SP.sqrt((x-xa)**2 + (y-ya)**2) - SP.sqrt((x-xb)**2+(y-yb)**2) - c * (ta-tb)
    f2 = SP.sqrt((x-xc)**2 + (y-yc)**2) - SP.sqrt((x-xb)**2+(y-yb)**2) - c * (tc-tb)
    return f1, f2

###############################################################

def Jacobi(x,y):
   j11 = (f(x+0.01,y)[0]- f(x,y)[0]) / 0.01
   j12 = (f(x,y+0.01)[0]- f(x,y)[0]) / 0.01
   j21 = (f(x+0.01,y)[1]- f(x,y)[1]) / 0.01
   j22 = (f(x,y+0.01)[1]- f(x,y)[1]) / 0.01
   return SP.mat([[j11, j12], [j21, j22]])

###############################################################


xold, yold = 4.0, 3.0   # this is the starting point of the Newton iteration

for i in range(10):     # do 10 iterations and hope that this will be good enough
    print "The function f has values ", f(xold, yold), "at the point (", xold, "," ,yold, ")"

    f1 = f(xold, yold)[0]
    f2 = f(xold, yold)[1]
    fold = SP.mat([[f1], [f2]])
    oldpoint = SP.mat([[xold], [yold]])
    newpoint = oldpoint - ((Jacobi(xold, yold)).I) * fold
    xold = newpoint[0,0]
    yold = newpoint[1,0]
```

Note that we implemented a simplified version of the Newton–Raphson scheme: Instead of calculating the correct Jacobi matrix

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial x}(x,y) & \frac{\partial f_1}{\partial y}(x,y) \\ \frac{\partial f_2}{\partial x}(x,y) & \frac{\partial f_2}{\partial y}(x,y) \end{pmatrix},$$

we approximate it instead by

$$J_{approx} = \begin{pmatrix} \frac{f_1(x+0.01,y)-f(x,y)}{0.01} & \frac{f_1(x,y+0.01)-f(x,y)}{0.01} \\ \frac{f_2(x+0.01,y)-f(x,y)}{0.01} & \frac{f_2(x,y+0.01)-f(x,y)}{0.01} \end{pmatrix},$$

since that is much easier to code in python, and then we calculate

$$\begin{pmatrix} x_{new} \\ y_{new} \end{pmatrix} = \begin{pmatrix} x_{old} \\ y_{old} \end{pmatrix} - J_{approx}^{-1} \cdot \vec{f}(x_{old}, y_{old}).$$

The output then is

```
The function f has values  (-1.693468238128784, -0.11385470645828644)
```

```
at the point ( 4.0 , 3.0 )
The function f has values  (-0.094431332191856399, 0.0042131315640920874)
at the point ( 4.77081328607 , 3.82534598984 )
The function f has values  (-0.00056081763845128307, -5.5150360863009951e-05)
at the point ( 4.81859773274 , 3.88594080586 )
The function f has values  (-3.013882476921026e-10, 3.2810530292692874e-08)
at the point ( 4.81889146485 , 3.88624708123 )
The function f has values  (6.1612936974597687e-12, 1.4998668973476015e-11)
at the point ( 4.81889146238 , 3.88624710482 )
The function f has values  (3.5527136788005009e-15, 7.1054273576010019e-15)
at the point ( 4.81889146238 , 3.88624710482 )
The function f has values  (-8.8817841970012523e-16, 0.0)
at the point ( 4.81889146238 , 3.88624710482 )
The function f has values  (0.0, 0.0) at the point ( 4.81889146238 , 3.88624710482 )
The function f has values  (0.0, 0.0) at the point ( 4.81889146238 , 3.88624710482 )
The function f has values  (0.0, 0.0) at the point ( 4.81889146238 , 3.88624710482 )
```

and we observe a nice convergence of a sequence of approximate solutions to some numerical limit

$$\begin{pmatrix} x_* \\ y_* \end{pmatrix} = \begin{pmatrix} 4.81889146238 \\ 3.88624710482 \end{pmatrix},$$

despite having chosen the “wrong” matrix $J_{approx}$ instead of the “correct” matrix $J$.

The reason for this fast convergence is that the initial approximation $\binom{4}{3}$ is quite close to the solution $\binom{x_*}{y_*}$. A badly chosen initial value $\binom{4}{30}$ will not lead to convergence:

```
The function f has values  (0.51183416757072919, -6.7522032722812675)
at the point ( 4.0 , 30.0 )
The function f has values  (0.50172874680521318, -6.6571956386473232)
at the point ( -158.96732764 , 4406.48575254 )
The function f has values  (0.50166048035026733, -6.6562871053814945)
at the point ( -1203688.87883 , 32329839.3762 )
The function f has values  (0.50434089750052635, -6.6540266811847744)
at the point ( -1877006.17783 , 50870991.3192 )
The function f has values  (0.6392880484461898, -6.5392390877008495)
at the point ( -1200091.28383 , 59801874.7512 )
The function f has values  (-0.62226093262432869, -7.5373634487390575)
at the point ( -10834947.3098 , 60659912.6452 )
Traceback (most recent call last):
  File "satnav2ddemo.py", line 39, in <module>
    newpoint = oldpoint - ((Jacobi(xold, yold)).I) * fold
  File "/usr/lib64/python2.6/site-packages/numpy/matrixlib/defmatrix.py",
line 870, in getI
    return asmatrix(func(self))
  File "/usr/lib64/python2.6/site-packages/numpy/linalg/linalg.py",
line 520, in inv
    ainv = _umath_linalg.inv(a, signature=signature, extobj=extobj)
  File "/usr/lib64/python2.6/site-packages/numpy/linalg/linalg.py",
line 90, in _raise_linalgerror_singular
    raise LinAlgError("Singular matrix")
numpy.linalg.linalg.LinAlgError: Singular matrix
```

We see that the sequence of approximate solutions runs away from $\binom{x_*}{y_*}$, and the error message at the end tells us that we try to calculate the inverse of a matrix which has no inverse matrix.

# Chapter 7

# Mathematics in Your Mobile Phone

The material of this chapter is taken from [1].

## 7.1 Historical Perspective

The second generation of mobile phones follow the GSM standard which was specified at a conference 1986 in Madeira. The participants at that conference were the post authorities of twelve countries, since at that time all services of post and telecommunication were run by governments. GSM abbreviates *Groupe Spécial Mobile*, and later its reading changed to *Global System for Mobile Communications*. The first GSM network was opened in 1992, and now the meaning changed to *God Send Mobiles* because the electronics engineers had a hard time making the phones small enough to make them actually portable. Remember that the first mobile phone (built for demonstration purposes in order to give evidence that the technical specifications are actually realistic) had the size of a refrigerator which was being driven around in a van. And that first mobile phone consumed so much power that its energy supply had to come from a 5000 Watt power supply system in a trailer attached to the van. The maximal data transfer rate in GSM is 55 kbit per second, and handling such speeds was a big technical challenge in the early Nineties. The third generation of mobile phones is called UMTS (Universal Mobile Telecommunications System) with a maximal data transfer rate of 384 kbit per second. The first UMTS network opened 2001 on the Isle of Man, and the second network then 2002 in Austria (again with very few phones). The UMTS standard was then extended to HSPA (generation 3G+) with speeds of 7 Mbit per second, sometimes up to 42 Mbit per second. HSPA became in widespread use around 2007/8. The next generations are LTE (generation 3.9) and LTE–Advanced (Generation 4; the marketing departments have their terminology wrong, as always) with data transfer rates up to 300 Mbit per second.

Before we come to the next chapter, I am sure you will join me in the admiration of the great works of the engineers who increased the data transfer rates by a factor of more than a thousand in less than 20 years. One may wish that managers could speed up their processes by a factor 1000 . . .

## 7.2 Mathematical Challenges

We will focus our considerations on GSM because that system is the easiest to understand, and the further generations contain much more mathematics, and much more *advanced* mathematics.

- How to squeeze a human talk and redundancy data and management data into 55 kbit per second ?
- Signals travel from the base station (which is typically an antenna with the shape of a $\approx$ 60cm cylinder) on a building to the receiver (which is your mobile phone). The base station does not know where you are, and therefore it can not send the signal to you like a laser beam. Instead, the signal is being sent in all directions with equal strength. A part of this stream arrives at your mobile phone directly (along the bee line), another part gets reflected at a high building and arrives at your mobile phone delayed and damped, and a further part of the signal is being reflected at a mountain far away, and hence it arrives at your phone even more delayed and damped.

Question: How to determine the strength of these echoes ?

- After having found out how strong the various echoes are, the next question comes up: How to subtract the echoes in order to get the original signal back ? This is hard because we only know the strength of the echo in comparison to the original signal, but we know neither of them individually.
- Signals get damaged during flight. How do we repair the damage ?
- How to protect paying customers against criminals impersonating as those customers ? "Remember" that in the early Nineties, a national mobile phone call had costs of about a pound per minute, and international phone calls were much more expensive. The technical answer to this question involves a lot of advanced mathematics, in particular *Abstract Algebra*, *Number Theory*, and *Cryptography.*
- How to do global roaming ? Which means a customer of a UK mobile phone company can travel to another continent, and you can still call them using their UK mobile phone number.

We will answer the second question and the third, both in the context of the GSM system, mainly because the higher data transfer rates of the more modern systems lead to considerable technical and mathematical complications. For instance, UMTS features the pseudo–noise sequences satisfying the Golomb postulates as we know them from Section 6.4.

## 7.3 How to Estimate the Channel

A *channel* consists of a sender (which is in our case the base station), a receiver (your mobile phone) and the connection (air) between them. Crucial for our understanding is that the signal received by your mobile phone is different from the signal that has been transmitted by the sender because of echoes, damping and fading, and atmospheric disturbances etc.

In this section we consider how to determine the strength of the echoes. The next section will then discuss how to remove the echoes from the received version of the signal.

The sender transmits a stream of bits such that each bit has a duration of $T_{bit} = 3.69\mu s = 3.69 \cdot 10^{-6}s$. Each bit in flight with light speed occupies a distance of

$$3 \cdot 10^8 \frac{m}{s} \cdot 3.69 \cdot 10^{-6} s \approx 10^3 m.$$

Electronics engineers have driven around with measuring equipment and observed that signal delays up to $16\mu s$ indeed *do* happen, which means that a part of the signal from the sender to the receiver did not take the bee line, but a detour of about 4 kilometres (because $16\mu s$ corresponds to 4 clock ticks, and each bit in flight occupies about one kilometre), for instance because of some reflection at a high mountain.

The following can also happen:

- The bee line is blocked by a big house.
- Only delayed (because of reflections) versions of the signal are being received.
- The user is sitting in a moving car, and after some time the user leaves the shadow of the building, and now the bee line version of the signal becomes available, but it arrives earlier than we expected.

We try to make sense of all this, and our scientific work technique (which you should get yourself accustomed to) is giving names to the quantities.

- The time ticks in slots of duration of $T_{bit} = 3.69 \cdot 10\mu s$, and the time variable is $k = 0, 1, 2, \ldots$ . Each number stands for one time slot.
- We consider only delays of integer multiples of $T_{bit}$. This is unrealistic, but easier for us. The maximal delay is four clock ticks (hence $4 \cdot 3.69\mu s = 14.76\mu s$ which is close enough to $16\mu s$).
- The sender sends a bit stream $s_0$, $s_1$, $s_2$, $\ldots$, with $s_k \in \{-1, 1\}$ for all $k$.
- The reflected signals are being damped with a factor $a_j \in [0, 1]$ with $j = 1, 2, 3, 4$.

- During the time slot $k$, the user receives the signal

  $$u_k = s_k + a_1 \cdot s_{k-1} + a_2 \cdot s_{k-2} + a_3 \cdot s_{k-3} + a_4 \cdot s_{k-4}.$$

  Here we assume (for simplicity) that user and sender have their clocks synchronized in a suitable manner which enables us to count their time variables using the same variable $k$. That is unrealistic, but otherwise we would have to use two different versions of $k$ (one for the user, one for the sender), leading to reader's confusion.

- The user's device can always measure all the $u_k$, and all these $u_k$ are real numbers.

- All the damping factors $a_1, \ldots, a_4$ are not known.

- All the $s_k$, $s_{k-1}$, ... are what the person on the other end of the conversation is telling us, and we don't know what they will tell us before the conversation has started (because if we knew what they are going to say, why would we dial their number in the first place ?). Therefore we do not know all the $s_k$. But we want to determine them.

As a summary: we know all the $u_k$ (by means of electrical measurement on the antenna of our phone), and we wish to calculate the $s_k$ from the $u_k$, but we can't because we do not know the $a_1$, $a_2$, $a_3$, $a_4$.

So we have a problem: **too many unknowns !**

Here comes the **solution:**

On each electromagnetic frequency, eight users are connected to the base station, but they are not communicating with the station all at the same time. This principle is called TDMA (time division multiple access). The time split into bigger time slots called TDMA frames, each such TDMA frame accommodates eight users and lasts $4.615ms$. The eight users are being served one after the other, which gives

$$\frac{4.615ms}{8} = 576.875\mu s \approx 577\mu s$$

available for each user. These $577\mu s$ are enough to accommodate 156.3 bits, each lasting $T_{bit} = 3.69\mu s$.

These 156.3 bits are arranged as follows:

- 3 tail bits (not very interesting to us)
- 57 data bits (containing the speech of the user)
- 1 stealing flag (we don't care what that means)
- 26 mid-amble bits (we will be very interested in them)
- 1 stealing flag
- 57 data bits
- 3 tail bits
- 8.25 guard bits (we don't care).

Basically, the tail-bits and the guard bits serve management purposes such as getting the clocks of the phone and the base station synchronized, etc. Management is boring.

Note that the user's speech data of $4.615ms$ is being compressed into $2 \times 57 = 114$ bits.

We are very interested in the 26 mid-amble bits because the phone uses them in order to calculate the $a_1$, $a_2$, $a_3$, $a_4$. Subsequently, these $a_j$ are being used for calculating the bits $s_k$ that have been transmitted at the sender from the bits $u_k$ which have been measured at the antenna of the user's phone (we will discuss this second step in the next section).

This is how to calculate the $a_j$. Take a *training sequence* with 16 digits as

$$\vec{m} := \Big( -1, +1, -1, -1, -1, +1, +1, +1, +1, -1, +1, +1, +1, -1, +1, +1 \Big).$$

**DIY:** *Calculate all its autocorrelation in the sense of equation (6.1) and display them in a diagram.*

Now we take the first 5 bits of $\overrightarrow{m}$ and append them at the end of $\overrightarrow{m}$; and we take the last 5 bits of $\overrightarrow{m}$ and prepend them at the front of $\overrightarrow{m}$. This way we obtain 26 bits, called the 26 **known** mid-amble bits.

We have

$$u_k = s_k + a_1 s_{k-1} + a_2 s_{k-2} + a_3 s_{k-3} + a_4 s_{k-4}, \qquad k = 5, 6, \ldots, 26, \tag{7.1}$$

with known numbers $s_1$, $s_2$, ..., $s_{26}$ (because they are being specified in official documents). And we know $u_1$, ..., $u_{26} \in \mathbb{R}$ from measurement. We neglect measurement errors and atmospheric noise etc.

Then the damping coefficients $a_1$, ..., $a_4$ are quickly being found by means of correlation coefficients.

Before we dive into the details of that calculation, it is a good moment to have a look at the *bigger picture* and at *Abstract Algebra*.

The correlation of two signals $\vec{v}$ and $\vec{w}$ from $\mathbb{R}^{16}$ is

$$\Psi(\vec{v}, \vec{w}) := \frac{1}{16} \sum_{j=1}^{16} v_j w_j,$$

which is a map $\mathbb{R}^{16} \times \mathbb{R}^{16} \to \mathbb{R}^1$ because we put two vectors from $\mathbb{R}^{16}$ in and get one real number out.

Now $\mathbb{R}^{16}$ is a *vector space*. What does this mean ?

Each human language has words, which are

**nouns,** which are names for existing things,

**adjectives,** which give a closer description of a thing denoted by a noun,

**verbs,** these are the actions that we can perform upon the nouns,

**adverbs,** which describe the verbs, hence they tell us *how* the actions are being done.

We have neglected some word types here.

Algebraic structures behave very similarly. They have

**objects**

**properties** of the objects (let us ignore them here),

**operations,** which tell us that you can do with the objects,

**rules** for these operations.

A first example of an algebraic structure is the *Algebraic Field* $\mathfrak{F}$ which has

**elements:** we imagine them as numbers

**two operations:** + and ×

**rules for these two operations:**

- + and × are commutative:

$$\forall a, b \in \mathfrak{F}: \quad a + b = b + a, \qquad a \times b = b \times a.$$

- + and × are associative:

$$\forall a, b, c \in \mathfrak{F}: \quad (a + b) + c = a + (b + c), \qquad (a \times b) \times c = a \times (b \times c).$$

- + and × together are distributive:

$$\forall a, b, c \in \mathfrak{F}: \quad (a + b) \times c = a \times c + b \times c.$$

- each equation $a + x = b$ with given $a, b \in \mathcal{F}$ has a unique solution $x \in \mathcal{F}$.
- there exists a unique element of $\mathcal{F}$ (typically called zero) that changes nothing when added to whatever number:

$$\exists\, 0 \in \mathcal{F}\colon \quad \forall a \in \mathcal{F}\colon \quad a + 0 = a.$$

- each equation $a \times y = b$ with given $a, b \in \mathcal{F}$ has a unique solution $y \in \mathcal{F}$ provided $a \neq 0$.

Examples of algebraic fields are $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{C}$, $\mathbb{F}_2$, compare page 101.

Another example of an algebraic structure is a *vector space* $\mathcal{V}$ *over the field* $\mathbb{R}$ which again has elements, operations, and rules for these operations:

**the elements of $\mathcal{V}$** are vectors which we imagine as arrows, and we imagine that two vectors are considered equivalent if they are parallel, have the same length, and point into the same direction[1]

**the operations:** there are two of them, namely *vector plus vector gives vector* and *real number times vector gives vector.*

**the rules for these two operations:**

- the first operation (called +) is commutative and associative:

$$\forall \vec{u}, \vec{v}, \vec{w} \in \mathcal{V}\colon \quad \vec{u} + \vec{v} = \vec{v} + \vec{u}, \qquad (\vec{u} + \vec{v}) + \vec{w} = \vec{u} + (\vec{v} + \vec{w}).$$

- We have two laws of distributivity:

$$\forall \vec{u}, \vec{v} \in \mathcal{V}, \quad \forall \lambda, \mu \in \mathbb{R}\colon \quad \lambda \cdot (\vec{u} + \vec{v}) = \lambda \cdot \vec{u} + \lambda \cdot \vec{v}, \quad (\lambda + \mu) \cdot \vec{u} = \lambda \cdot \vec{u} + \mu \cdot \vec{u}$$

- We have something resembling associativity of multiplication (but it is not actually a law of associativity because of the types of the objects being wrong):

$$\forall \lambda, \mu \in \mathbb{R}\colon \quad \forall \vec{u} \in \mathcal{V}\colon \quad (\lambda \cdot \mu) \cdot \vec{u} = \lambda \cdot (\mu \cdot \vec{u}).$$

- Each equation $\vec{u} + \vec{x} = \vec{v}$ with given $\vec{u}, \vec{v} \in \mathcal{V}$ has a unique solution $\vec{x} \in \mathcal{V}$.
- And then there is a rule which is needed in the abstract theory of vector spaces:

$$\forall \vec{u} \in \mathcal{V}\colon \quad 1 \cdot \vec{u} = \vec{u}.$$

Examples of vector spaces are $\mathbb{R}^1$, $\mathbb{R}^2$, ..., $\mathbb{R}^{16}$, ....

Another word for vector spaces are *linear spaces*. It is important to remember that linear spaces have *two operations*: adding two vectors, and multiplying a vector by a number.

Now take two linear spaces $\mathbb{R}^p$ and $\mathbb{R}^q$. We consider a function $f$ that maps from $\mathbb{R}^p$ into $\mathbb{R}^q$. We say that this mapping $f\colon \mathbb{R}^p \to \mathbb{R}^q$ is a *linear map* if it is compatible to the two operations of the two linear spaces $\mathbb{R}^p$ and $\mathbb{R}^q$. To be specific, this means the following two requirements:

$$\begin{aligned} \forall \vec{u}, \vec{v} \in \mathbb{R}^p\colon \quad & f\,(\vec{u} + \vec{v}) = f(\vec{u}) + f(\vec{v}), \\ \forall \vec{u} \in \mathbb{R}^p, \quad \forall \lambda \in \mathbb{R}\colon \quad & f\,(\lambda \cdot \vec{u}) = \lambda \cdot f(\vec{u}). \end{aligned}$$

By the way, it is an important theorem of *Linear Algebra* that each such a linear map $f\colon \mathbb{R}^p \to \mathbb{R}^q$ is being generated by a matrix $A \in \mathbb{R}^{q \times p}$ in the sense of $f(\vec{u}) = A \cdot \vec{u}$.

What can we do with these two operations in a vector space ? We can check whether three points are on a line, or whether two lines are parallel, but not much more: we can't determine angles, and we also can't measure lengths.

Therefore we need more algebraic structures.

A *real inner product space* is a vector space $\mathcal{V}$ over the field $\mathbb{R}$ that has one more operation $\Psi$ that takes two vectors from $\mathcal{V}$ and builds from them a real number. The properties of this additional operation are:

[1] keep in mind though that this imagination is not part of the official definition of a vector space. In abstract algebra you never specify what the objects (such as vectors) actually **are**, because the proofs of the theorems never need this information. Instead, you only declare how the operations **behave** (meaning: which rules they follow).

**$\Psi$ is linear in the first argument:**

$$\begin{aligned} \forall \vec{u},\vec{v},\vec{w}\in\mathcal{V}\colon &\quad \Psi(\vec{u}+\vec{v},\vec{w})=\Psi(\vec{u},\vec{w})+\Psi(\vec{v},\vec{w}),\\ \forall \vec{u},\vec{v}\in\mathcal{V},\quad \forall\lambda\in\mathbb{R}\colon &\quad \Psi(\lambda\cdot\vec{u},\vec{v})=\lambda\cdot\Psi(\vec{u},\vec{v}). \end{aligned}$$

**$\Psi$ is symmetric:**

$$\forall \vec{u},\vec{v}\in\mathcal{V}\colon\quad \Psi(\vec{u},\vec{v})=\Psi(\vec{v},\vec{u}).$$

**$\Psi$ is positive definite:**

$$\begin{aligned} &\forall \vec{u}\in\mathcal{V}\colon\quad \Psi(\vec{u},\vec{u})\ge 0,\\ &\Psi(\vec{u},\vec{u})=0\quad\text{if and only if}\quad \vec{u}=\vec{0}. \end{aligned}$$

Because of the symmetry property, $\Psi$ is also linear in the second argument:

$$\begin{aligned} \forall \vec{u},\vec{v},\vec{w}\in\mathcal{V}\colon &\quad \Psi(\vec{w},\vec{u}+\vec{v})=\Psi(\vec{w},\vec{u})+\Psi(\vec{w},\vec{v}),\\ \forall \vec{u},\vec{v}\in\mathcal{V}\quad,\forall\lambda\in\mathbb{R}\colon &\quad \Psi(\vec{v},\lambda\cdot\vec{u})=\lambda\cdot\Psi(\vec{v},\vec{u}). \end{aligned}$$

Hence we often say that $\Psi$ is *bilinear.*

An inner product space therefore has three operations (two of them inherited from the definition of abstract vector spaces), and the third operation $\Psi$ (often called *scalar product* or *inner product*) is as compatible as possible to the two operations that have been defined earlier. Compatibility here means bilinearity.

The big advantage of such a scalar product is that we can now calculate the angle between two vectors $\vec{a}$ and $\vec{b}$, based on a famous inequality:

**Proposition 7.1** (**Inequality of Cauchy and Schwarz**). *For all $\vec{a}$, $\vec{b}\in\mathcal{V}$, we have*

$$\Big(\Psi(\vec{a},\vec{b})\Big)^2\le\Psi(\vec{a},\vec{a})\cdot\Psi(\vec{b},\vec{b}).$$

*Proof.* Let $\vec{a}$, $\vec{b}\in\mathcal{V}$ be given, and let $\lambda\in\mathbb{R}$ be arbitrary. Then we have

$$0\le\Psi(\vec{a}+\lambda\cdot\vec{b},\vec{a}+\lambda\cdot\vec{b}) \tag{7.2}$$

because $\Psi$ is positive definite. Now let $\lambda$ run through all the set of real numbers: the term $\Psi(\vec{a}+\lambda\cdot\vec{b},\vec{a}+\lambda\cdot\vec{b})$ never becomes negative. But the bilinearity and the symmetry of $\Psi$ enable us to re-arrange (7.2) into

$$0\le\Psi(\vec{a},\vec{a})+2\lambda\cdot\Psi(\vec{a},\vec{b})+\lambda^2\cdot\Psi(\vec{b},\vec{b}).$$

Now we recall from school (if needed: DIY) the following fact: if the quadratic function $x\mapsto m+2nx+px^2$ never takes negative values, then $n^2-mp\le 0$.

Now we substitute $m:=\Psi(\vec{a},\vec{a})$, $n:=\Psi(\vec{a},\vec{b})$, $p:=\Psi(\vec{b},\vec{b})$, $\lambda:=x$, and the claim is proved. □

Then we obtain

$$-1\le\frac{\Psi(\vec{a},\vec{b})}{\sqrt{\Psi(\vec{a},\vec{a})}\cdot\sqrt{\Psi(\vec{b},\vec{b})}}\le 1,$$

and consequently there is an angle $\gamma\in[0,\pi]$ with

$$\cos(\gamma)=\frac{\Psi(\vec{a},\vec{b})}{\sqrt{\Psi(\vec{a},\vec{a})}\cdot\sqrt{\Psi(\vec{b},\vec{b})}}. \tag{7.3}$$

There is one more operation in an inner product space: the norm $\|\vec{a}\|$ of a vector $\vec{a}$ which is defined as

$$\|\vec{a}\|:=\sqrt{\Psi(\vec{a},\vec{a})}.$$

Then the formula (7.3) takes the more familiar form

$$\Psi(\vec{a},\vec{b})=\cos(\gamma)\cdot\|\vec{a}\|\cdot\left\|\vec{b}\right\|.$$

The geometrical meaning of the norm $\|\vec{a}\|$ is typically *the length of the vector $\vec{a}$.*

**Proposition 7.2.** *The norm in an inner product space $\mathcal{V}$ has the following properties:*

$$\begin{aligned}
&\|\vec{a}\| \geq 0 \quad \textit{for all } \vec{a} \in \mathcal{V},\\
&\|\vec{a}\| = 0 \quad \textit{if and only if } \vec{a} = \vec{0},\\
&\|\lambda \cdot \vec{a}\| = |\lambda| \cdot \|\vec{a}\| \quad \textit{for all } \vec{a} \in \mathcal{V} \textit{ and all } \lambda \in \mathbb{R},\\
&\left\|\vec{a} + \vec{b}\right\| \leq \|\vec{a}\| + \left\|\vec{b}\right\| \quad \textit{for all } \vec{a} \in \mathcal{V} \textit{ and all } \vec{b} \in \mathcal{V}.
\end{aligned}$$

**DIY:** *Prove it. For the last statement, you perhaps need the Cauchy Schwarz inequality.*

Now we have four operations in an inner product space. The statements $\|\lambda \cdot \vec{a}\| = |\lambda| \cdot \|\vec{a}\|$ and $\left\|\vec{a} + \vec{b}\right\| \leq \|\vec{a}\| + \left\|\vec{b}\right\|$ mean that the new (fourth) operation *norm of a vector* is as much as possible compatible to the two oldest operations *number times vector* and *vector plus vector*. In an ideal world, we would perhaps have hoped for equality $\left\|\vec{a} + \vec{b}\right\| = \|\vec{a}\| + \left\|\vec{b}\right\|$ instead of inequality, but in abstract algebra as well as in real life, you can't have everything.

Now let us get back to the question of determining the damping factors $a_1$, $a_2$, $a_3$, $a_4$ that relate the $u_k$ and the $s_k$ via

$$u_k = s_k + a_1 \cdot s_{k-1} + a_2 \cdot s_{k-2} + a_3 \cdot s_{k-3} + a_4 \cdot s_{k-4}, \qquad k = 5, 6, \ldots, 26. \tag{7.4}$$

A simplified version of the calculation of the $a_j$ is this: we define

$$\begin{aligned}
\vec{u} &:= (u_6, u_7, \ldots, u_{21}) \in \mathbb{R}^{16},\\
\vec{s} &:= (s_6, s_7, \ldots, s_{21}) \in \mathbb{R}^{16},\\
\vec{s_1} &:= (s_5, s_6, \ldots, s_{20}) \in \mathbb{R}^{16},\\
\vec{s_2} &:= (s_4, s_5, \ldots, s_{19}) \in \mathbb{R}^{16},\\
\vec{s_3} &:= (s_3, s_4, \ldots, s_{18}) \in \mathbb{R}^{16},\\
\vec{s_4} &:= (s_2, s_3, \ldots, s_{17}) \in \mathbb{R}^{16}.
\end{aligned}$$

Then the equation (7.4) implies

$$\vec{u} = \vec{s} + a_1 \cdot \vec{s_1} + a_2 \cdot \vec{s_2} + a_3 \cdot \vec{s_3} + a_4 \cdot \vec{s_4}.$$

The vector $\vec{u} \in \mathbb{R}^{16}$ is known from antenna measurement, and the $\vec{s}$, $\vec{s_1}$, $\ldots$, $\vec{s_4}$ are known from specification.

We also know from the DIY exercise on page 115 that

$$\begin{aligned}
&\Psi(\vec{s}, \vec{s_1}) = \Psi(\vec{s}, \vec{s_2}) = \Psi(\vec{s}, \vec{s_3}) = \Psi(\vec{s}, \vec{s_4}) = 0, \quad \Psi(\vec{s_j}, \vec{s_k}) = 0, \quad (j \neq k),\\
&\Psi(\vec{s}, \vec{s}) = \Psi(\vec{s_1}, \vec{s_1}) = \Psi(\vec{s_2}, \vec{s_2}) = \Psi(\vec{s_3}, \vec{s_3}) = \Psi(\vec{s_4}, \vec{s_4}) = 1.
\end{aligned}$$

Now $\Psi$ is linear in its first argument, hence

$$\begin{aligned}
\Psi(\vec{u}, \vec{s_1}) &= \Psi(\vec{s} + a_1 \cdot \vec{s_1} + a_2 \cdot \vec{s_2} + a_3 \cdot \vec{s_3} + a_4 \cdot \vec{s_4}, \vec{s_1})\\
&= \Psi(\vec{s}, \vec{s_1}) + a_1 \cdot \Psi(\vec{s_1}, \vec{s_1}) + a_2 \cdot \Psi(\vec{s_2}, \vec{s_1}) + a_3 \cdot \Psi(\vec{s_3}, \vec{s_1}) + a_4 \cdot \Psi(\vec{s_4}, \vec{s_1})\\
&= 0 + a_1 \cdot 1 + a_2 \cdot 0 + a_3 \cdot 0 + a_3 \cdot 0.
\end{aligned}$$

This gives us an easy formula for $a_1$:

$$a_1 = \Psi(\vec{u}, \vec{s_1}) = \frac{1}{16} \sum_{j=0}^{15} u_{6+j} s_{5+j},$$

and everything on the RHS is known. Similarly, we have

$$a_2 = \Psi(\vec{u}, \vec{s_2}), \quad a_3 = \Psi(\vec{u}, \vec{s_3}), \quad a_4 = \Psi(\vec{u}, \vec{s_4}).$$

The actual calculation is a bit more involved: we have to take care of measurement errors and damaged signals, which is where the full 26 bits enter the picture (we took only 16 bits of $u$ and 20 bits of $s$).

## 7.4 How to Reconstruct the Transmitted Signal

Now we know the damping factors $a_1, \ldots, a_4$, and we wish to reconstruct the 57 data bits that contain the speech of the other person. Let us list what we know and what we want:

**We know** 57 values $u_1, u_2, \ldots, u_{57} \in \mathbb{R}$ that have been measured at our antenna.

**We know** the four damping factors $a_1, \ldots, a_4$, and they are real numbers between 0 and 1.

**We don't know** 57 data bits $s_1, s_2, \ldots, s_{57} \in \{-1, 1\}$ that have been transmitted at the base station.

**We know** the relation $u_k = s_k + a_1 s_{k-1} + a_2 s_{k-2} + a_3 s_{k-3} + a_4 s_{k-4}$.

Our goal is to calculate the $s_k$. And here is how to do it, "engineering style": we clearly have

$$s_k = u_k - a_1 s_{k-1} - a_2 s_{k-2} - a_3 s_{k-3} - a_4 s_{k-4}.$$

And we also have

$$\begin{aligned}
s_{k-1} &= u_{k-1} - a_1 s_{k-2} - a_2 s_{k-3} - a_3 s_{k-4} - a_4 s_{k-5}, \\
s_{k-2} &= u_{k-2} - a_1 s_{k-3} - a_2 s_{k-4} - a_3 s_{k-5} - a_4 s_{k-6}, \\
s_{k-3} &= u_{k-3} - a_1 s_{k-4} - a_2 s_{k-5} - a_3 s_{k-6} - a_4 s_{k-7}, \\
s_{k-4} &= u_{k-4} - a_1 s_{k-5} - a_2 s_{k-6} - a_3 s_{k-7} - a_4 s_{k-8}.
\end{aligned}$$

We plug this into the above line and get

$$\begin{aligned}
s_k = u_k - a_1 u_{k-1} + a_1 \Big( a_1 s_{k-2} + a_2 s_{k-3} + a_3 s_{k-4} + a_4 s_{k-5} \Big) \\
- a_2 u_{k-2} + a_2 \Big( a_1 s_{k-3} + a_2 s_{k-4} + a_3 s_{k-5} + a_4 s_{k-6} \Big) \\
- a_3 u_{k-3} + a_3 \Big( a_1 s_{k-4} + a_2 s_{k-5} + a_3 s_{k-6} + a_4 s_{k-7} \Big) \\
- a_4 u_{k-4} + a_4 \Big( a_1 s_{k-5} + a_2 s_{k-6} + a_3 s_{k-7} + a_4 s_{k-8} \Big).
\end{aligned}$$

This looks perhaps a bit messy, so we may re-write it as

$$s_k = u_k - \sum_{j=1}^{4} a_j u_{k-j} + \sum_{\ell=2}^{8} \left( \sum_{\substack{i+j=\ell \\ 1 \le i,j \le 4}} a_i a_j \right) \cdot s_{k-\ell},$$

which is not that clear either. On the RHS, we know the $u_k$ and the $u_{k-j}$, but we don't know the $s_{k-\ell}$. However, experience tells the engineers that the coefficients $a_j$ are sometimes very small. The electromagnetic waves are being emanated like a spherical wave, hence their strength decays the longer they are travelling. Therefore $1 > a_1 > a_2 > a_3 > a_4 \ge 0$, and then the products $a_i a_j$ are even smaller than $a_4$ most of the time, so we may decide to just throw the terms $a_i a_j$ away. So the formula becomes

$$s_k = \text{Round} \left( u_k - \sum_{j=1}^{4} a_j u_{k-j} \right),$$

and the Round shall mean that we are rounding to the nearest of the numbers $+1$ and $-1$ (recall that we know $s_k \in \{+1, -1\}$).

This calculation style (just throw away those items that are believed to be small) might look scary, but in the real life it works quite well.

# Appendix A

# Some Program Codes

## A.1 The Python Code For Section 3.4

```
from scipy.optimize import fsolve
import numpy as NP

nwred = 1.330
nworange = 1.334
nwyellow = 1.335
nwgreen = 1.338
nwblue = 1.341
nwviolet = 1.349

def angle_from_y(y, nw):
    result = 180 * (4 * NP.arcsin(y / nw) - 2 * NP.arcsin(y) + NP.pi) / NP.pi
    return result

rainbowred = lambda x : 4 / (nwred * NP.sqrt(1-(x/nwred)**2)) - 2 / NP.sqrt(1-x**2)
yred = fsolve(rainbowred, 0.9)
alphared = angle_from_y(yred,nwred)
print("The angle of the red part is ", alphared[0] - 180.0)

rainboworange = lambda x : 4 / (nworange * NP.sqrt(1-(x/nworange)**2)) - 2 / NP.sqrt(1-x**2)
yorange = fsolve(rainboworange, 0.9)
alphaorange = angle_from_y(yorange,nworange)
print("The angle of the orange part is ", alphaorange[0] - 180.0)

rainbowyellow = lambda x : 4 / (nwyellow * NP.sqrt(1-(x/nwyellow)**2)) - 2 / NP.sqrt(1-x**2)
yyellow = fsolve(rainbowyellow, 0.9)
alphayellow = angle_from_y(yyellow,nwyellow)
print("The angle of the yellow part is ", alphayellow[0] - 180.0)

rainbowgreen = lambda x : 4 / (nwgreen * NP.sqrt(1-(x/nwgreen)**2)) - 2 / NP.sqrt(1-x**2)
ygreen = fsolve(rainbowgreen, 0.9)
alphagreen = angle_from_y(ygreen,nwgreen)
print("The angle of the green part is ", alphagreen[0] - 180.0)

rainbowblue = lambda x : 4 / (nwblue * NP.sqrt(1-(x/nwblue)**2)) - 2 / NP.sqrt(1-x**2)
yblue = fsolve(rainbowblue, 0.9)
alphablue = angle_from_y(yblue,nwblue)
print("The angle of the blue part is ", alphablue[0] - 180.0)

rainbowviolet = lambda x : 4 / (nwviolet * NP.sqrt(1-(x/nwviolet)**2)) - 2 / NP.sqrt(1-x**2)
yviolet = fsolve(rainbowviolet, 0.9)
alphaviolet = angle_from_y(yviolet,nwviolet)
print("The angle of the violet part is ", alphaviolet[0] - 180.0)
```

# Bibliography

[1] Martin Bossert and Sebastian Bossert. *Mathematik der digitalen Medien.* VDE Verlag, 2010.

[2] Global Positioning Systems Directorate Systems Engineering and Integration. *Interface Specification IS-GPS-800*, 2013.

[3] IAAF. *IAAF Scoring Tables for Combined Events*, 2001.

[4] Anthony Lo Bello. *Origins of mathematical words.* Johns Hopkins University Press, Baltimore, MD, 2013. A comprehensive dictionary of Latin, Greek, and Arabic roots.

[5] Matthias Ludwig. *Mathematik + Sport. Olympische Disziplinen im Blick.* Vieweg und Teubner, 2008.

[6] Marcus Vitruvius Pollio. *The Ten Books on Architecture.* Harvard University Press, 1914. Translated by Morris Hicky Morgan.

[7] James Pryde, editor. *Chambers seven-figure mathematical tables.* Chambers Edinburgh, 1974.

[8] David J. Segelstein. *The Complex Refractive Index of Water.* PhD thesis, University of Missouri-Kansas City, 1981.