Introduction to Mathematics and Optimization

Niels Lauritzen

Compressed static version (11.8.2025) of interactive book: https://edtech.dk/IMO25

Contents

1 The language of mathematics and prompting **6**

1.1 The art of prompting . . . 6

1.2 Black box warnings . . . 8

1.2.1 Interacting with chatbots . . . 8

1.3 Computer algebra (and python) . . . 8

1.4 Numbers . . . 10

1.4.1 The natural numbers $\mathbb{N}$ and the integers $\mathbb{Z}$ . . . 11

1.4.2 The rational numbers $\mathbb{Q}$ . . . 11

1.4.3 The real numbers $\mathbb{R}$ . . . 12

1.4.4 Arithmetic rules for numbers . . . 13

1.5 Propositional logic . . . 14

1.5.1 Propositional logic as a formal language . . . 16

1.5.2 Truth tables and equivalent propositions . . . 17

1.5.3 Using Sage to compute truth tables . . . 18

1.5.4 Variables, predicates and quantification . . . 18

1.5.5 Proofs and inference rules . . . 20

1.5.6 The use of implication ($\Longrightarrow$) and bi-implication ($\Longleftrightarrow$) . . . 20

1.5.7 More on mathematical proofs . . . 21

1.5.8 Proof by contradiction . . . 22

1.6 More on sets . . . 24

1.6.1 Objects and equality . . . 24

1.6.2 The symbols $\in$ and $\notin$ . . . 26

1.6.3 Subsets . . . 26

1.6.4 Set-builder notation . . . 28

1.6.5 Intersections, unions and the symbols $\cap$, $\cup$ and $\setminus$ . . . 30

1.6.6 Pairs, triples and tuples . . . 33

1.7 Ordering numbers . . . 34

1.7.1 Ordering $\mathbb{Z}$ . . . 35

1.7.2 Ordering $\mathbb{Q}$ . . . 36

1.7.3 Ordering $\mathbb{R}$ . . . 39

1.8 Proof by induction . . . 39

1.9 The concept of a function . . . 43
1.9.1 When are two functions the same? . . . 45
1.9.2 Notations for defining a function . . . 45
1.9.3 Composition of functions . . . 46
1.9.4 Functions from and into products . . . 47
1.9.5 Injective and surjective functions . . . 48
1.9.6 The inverse function . . . 49
1.9.7 The preimage . . . 49
1.9.8 Neural networks . . . 50

2 Linear equations **54**
2.1 One linear equation with one unknown . . . 54
2.2 Several linear equations with several unknowns . . . 56
2.2.1 Several equations . . . 56
2.3 Gauss elimination . . . 57
2.4 Polynomials . . . 61
2.4.1 Polynomial division . . . 62
2.4.2 Roots of polynomials . . . 64
2.5 Applications of linear equations to polynomials . . . 67
2.5.1 The magic of Lagrange polynomials . . . 70
2.6 Shamir secret sharing . . . 72
2.7 Fitting data . . . 72

3 Matrices **74**
3.1 Matrices . . . 74
3.1.1 Definitions . . . 74
3.2 Linear maps . . . 76
3.3 Matrix multiplication . . . 77
3.3.1 Matrix multiplication in `numpy` . . . 79
3.3.2 The identity matrix . . . 79
3.3.3 Examples of matrix multiplication . . . 80
3.4 Matrix arithmetic . . . 83
3.4.1 Matrix addition . . . 83
3.4.2 Multiplication of a number and a matrix . . . 84
3.4.3 The distributive law . . . 84
3.4.4 The miraculous associative law . . . 85
3.5 The inverse matrix . . . 87
3.5.1 Well, how do I find the inverse of a matrix? . . . 90
3.6 The transposed matrix . . . 91
3.7 Symmetric matrices . . . 92
3.7.1 Positive definite matrices . . . 92

3.7.2 Positive semi-definite matrices . . . 93
3.7.3 Symmetric reductions . . . 93

4 What is optimization? 95
4.1 What is an optimization problem? . . . 95
4.2 General definition . . . 97
4.3 Convex optimization . . . 99
4.4 Linear optimization . . . 105
4.5 Fourier-Motzkin elimination . . . 108
4.6 Application in machine learning and data science . . . 111
4.6.1 Formulation as a linear optimization problem . . . 113

5 Euclidean vector spaces 114
5.1 Vectors in the plane . . . 114
5.2 Higher dimensions . . . 115
5.2.1 Dot product, norm and cosine . . . 116
5.3 The unreasonable effectiveness of the dot product . . . 118
5.3.1 The dist formula from high school . . . 118
5.3.2 The perceptron algorithm . . . 119
5.3.3 Why does the perceptron algorithm work? . . . 123
5.4 Pythagoras and the least squares method . . . 123
5.5 The Cauchy-Schwarz inequality . . . 128
5.5.1 The triangle inequality . . . 129
5.5.2 Cosine similarity in machine learning . . . 130
5.6 Special subsets of euclidean spaces . . . 131
5.6.1 Bounded subsets . . . 132
5.6.2 Open, closed and compact subsets and boundaries and interiors of subsets . . . 134
5.7 Continuous functions . . . 138
5.7.1 An elegant way of characterizing a continuous function . . . 140
5.7.2 Working with continuous functions . . . 141
5.8 Important and special results for continuous functions . . . 144

6 Convex functions 146
6.1 Strictly convex functions . . . 146
6.2 Why are convex functions interesting? . . . 148
6.3 Differentiable functions . . . 151
6.3.1 Definition . . . 151
6.3.2 Formulas . . . 154
6.3.3 The derivative of a product . . . 155
6.3.4 The one variable chain rule . . . 156
6.3.5 The Newton-Raphson method for finding roots . . . 157

6.3.6 Critical points and extrema . . . 159
6.3.7 Increasing functions . . . 159
6.4 Taylor polynomials . . . 162
6.5 Differentiable convex functions . . . 164

7 Several variables **166**
7.1 Introduction . . . 166
7.2 Vector functions . . . 169
7.3 Differentiability . . . 170
7.3.1 Partial derivatives . . . 170
7.4 Newton-Raphson in several variables! . . . 174
7.5 Local extrema in several variables . . . 174
7.6 The chain rule . . . 177
7.6.1 Matrix multiplication graphically . . . 179
7.6.2 Unpacking the chain rule . . . 181
7.7 Logistic regression . . . 184
7.7.1 Estimating the parameters . . . 185
7.8 3Blue1Brown . . . 188
7.8.1 Introduction to neural networks . . . 188
7.8.2 Gradient descent . . . 188
7.8.3 Backpropagation and training . . . 188
7.8.4 The chain rule in action . . . 188
7.9 Lagrange multipliers . . . 189
7.10 Optimization using the interior and boundary of a subset . . . 192

8 The Hessian **195**
8.1 Introduction . . . 195
8.2 Several variables . . . 195
8.3 Newton's method for finding critical points . . . 197
8.3.1 Transforming data for better numerical performance . . . 198
8.4 The Hessian and critical points . . . 199
8.5 Differential convex functions of several variables . . . 203
8.6 How to decide the definiteness of a matrix . . . 206
8.7 A schematic procedure for transforming symmetric matrices . . . 209

9 Convex optimization **211**
9.1 Finding the optimal hyperplane separating data . . . 213
9.1.1 Separating by non-linear functions . . . 218
9.1.2 Kernel functions . . . 219
9.1.3 The kernel perceptron algorithm . . . 220
9.2 Logarithmic barrier functions . . . 221

9.2.1 Quadratic function with polyhedral constraints 222
9.3 A geometric optimality criterion 223
9.4 KKT 225
9.5 Computing with KKT 227
9.5.1 Strategy 227
9.5.2 Example 228
9.6 Optimization exercises 229

Chapter 1

The language of mathematics and prompting

1.1 The art of prompting

In the past year generative AI has evolved with explosive speed. So much so, that the use of it is in **no way** allowed during the written exam in this course. This also includes local models run on your own computer and code completions.

In May 2025, Google introduced the reasoning model Gemini 2.5 Pro and just recently (August 7) OpenAI launched GPT-5. These are incredibly powerful models, which do college level mathematics (and computer science) superbly. You can access Gemini 2.5 Pro using Google AI Studio for free with a gmail account. It seems that GPT-5 is also freely available through ChatGPT with a limited number of prompts per day. I suspect that GPT-5 will also be available via a student license at Microsoft Copilot.

In order to get a contextual and good response it seems important to work with a reasoning model, which means that you have to wait up to a few minutes for feedback. The quicker, non-thinking models sometimes mess up the context.

You communicate with chatbots (large language models) through natural language. This process is called prompting. The more precise your prompt is, the better the response. When learning new material, you can work from prompts instructing the chatbot not to give away the answers but emphasize guidance. Here is a good example of this.

(1.1) EXERCISE.

LLM

> I am a student following the course based on the attached notes. Please guide me through Exercise 1.15. Emphasize my learning and do not give me the answers but only hints. You must only use material from the attached notes in the solution. Be sure to reference what you use. Please make a serious effort to render the mathematics in your output using KaTeX so that I can read it!

A click on a chatbot link copies the prompt to the clipboard and takes you to the chatbot, where you can paste the prompt. Click on a chatbot of your choice. Then attach the pdf version imo25.pdf of the interactive notes. Submit and interact.
Try different scenarios and chatbots. Browse through the beginning of the notes and change the prompt to suit you. Was the mathematics presented nicely in the browser (in Google AI Studio it helps to set the temperature to zero)?›

♠

Gemini has a mode called guided learning and ChatGPT has something called study mode that you may also use. In any case, precise prompting is a very valuable skill.

In this chapter several examples of prompts will be given. In the following chapters less so. Here you are expected to prompt the chatbots on your own. Sometimes a prompt related to the context pops up as below.

LLM

> Please solve the equation $x^2 - x - 1 = 0$. Guide me through the steps. Make sure that the underlying logic in your arguments is correct.

(1.2) EXERCISE.

Come up with prompts that make a chatbot act like a mathematics tutor for you. Here is a small example that you may extend.

LLM

> Please act like a friendly tutor and teach me about the derivatives of simple functions. Test my understanding after each concept you explain.

Try out the features guided learning in Gemini 2.5 Pro and study mode in ChatGPT. on the example above. ♠

(1.3) EXERCISE.

Start your LaTeXjourney using the prompt below.

LLM

> I am doing weekly exercises in mathematics at the college level. Please suggest a very simple template in LaTeX for hand in of these exercises. Also, show me how to typeset an equation in LaTeX.
> There seems to be a web interface to LaTeX called Overleaf. Please tell me how to access this so that I can enter a weekly exercise.

Come up with your own prompt for a question related to software. ♠

(1.4) EXERCISE.

Below I ask for feedback from the chatbot on some dubious chunk of mathematics.

LLM

> Please give feedback on the mathematics contained in the LaTeX below in triple quotes. Emphasize logic and precision. """
>
> $$x^2 = 1 \implies x = 1$$
>
> From this it follows that $1 + 1 = 3$. """

Insert your own mathematics in LaTeXnotation and ask for feedback in a prompt. ♠

1.2 Black box warnings

Modern mathematics is perhaps not like anything you have encountered so far. It calls for a lot of focus and precision, especially when writing down solutions to problems. It is a bit like programming a computer. There is no room for imprecision and half-baked sentences.

This course amounts to 10 ECTS or approximately 280 hours. Suppose that you spend a week studying for the exam, say 40 hours. Lectures, exercise classes, and MatLab amount to $14 \cdot (4+2+3)$ hours $= 126$ hours. This leaves around 114 hours for your own study and immersion. Put in other terms, you are supposed to work around 8 hours per week outside classes for this course. With classes, each week calls for 17 hours of work. There is a very close relationship between the amount of hours you log each week and your result at the exam. To state the obvious: numbers don't lie. If you put in the time, you will almost certainly do well. Try to allocate time for IMO in your weekly schedule and please (ab)use all the help that is provided.

LLM

> I am taking a first semester college level mathematics course spanning 14 weeks and 10 ECTS. One ECTS amounts to 28 hours. The teaching activities every week amount to 4 hours of lectures, 2 hours of exercise sessions and 3 hours of study cafe. I expect that final exam will take 40 hours of the 10 ECTS. Please schedule a weekly study plan for me along with a plan for the final exam.

1.2.1 Interacting with chatbots

Using chatbots is strongly encouraged, but it takes a while to pick up how to engage them to boost learning. The most useless prompt of them all is given below

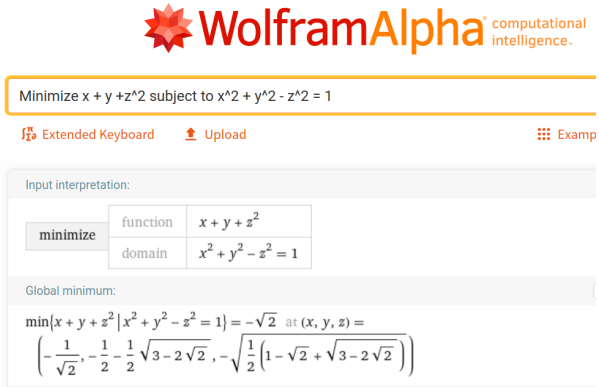
LLM

> Please give me the complete solution to Exercise 3.42 in the attached pdf file. Be sure to use only mathematics from this attached file and referencing precisely the proper definitions, propositions, theorems, etc. Give your answer as the source code for perfectly formatted LaTeX.

Here every inch of the cognitive effort is outsourced. Once in a while we all crawl down this rabbit hole. Personally, I get depressed using such mindless interaction. More importantly, it is arguably the worst way of preparing for the exam in this course.

1.3 Computer algebra (and python)

Computers are exceptionally fun, but be careful! Nothing really beats a clear thinking human mind. To wit, I asked WolframAlpha to solve a certain optimization problem and it came up with the answer



(1.5) EXERCISE.

Prompt a chatbot with

LLM

> What is
> $$-\sqrt{\frac{1}{2}(1-\sqrt{2}+\sqrt{3-2\sqrt{2}})}$$

and explain why the output from WolframAlpha is weird. Use prompting to make it explain the mathematics input notation.

Finally use your own mental powers (and feedback from the chatbot) to explain what the proper output should have been.

Hint:

$$3-2\sqrt{2}=(\sqrt{2}-1)^2.$$

♠

We will use the computer algebra system Sage in exploring and experimenting with mathematics. This means that you will have to write small commands and code snippets.

Sage is built on top of the very wide spread language python and you can in fact enter Python code[1] in the Sage input windows in the interactive notes. First adjust the prompt below according to your needs and get feedback from an LLM.

LLM

> I am taking a mathematics course that uses the computer algebra system Sage. The course uses the browser, where I can enter and run small snippets of code in Sage and python. I have no/some/extensive prior programming experience. Give me a brief introduction to Sage and explain how it relates to python. Finish your reply with a small exercise I can do. If no/some/extensive is present above in this prompt, remark this and only reply with Please select your programming experience.

Below is an example of a basic graphics command in Sage. Push the Compute button to evaluate.

> Interactive code not included in static version.

You can install Sage on your own computer following the instructions on https://www.sagemath.org/.

(1.6) EXERCISE.

Did you notice that you can edit and enter new commands in the Sage window? Do the following problems using Sage based on the Sage guided tour or asking a chatbot.

(i) Consider $f(x)=x\sin(1/x)$. Plot the graph of f from 0 to 0.1. Computing $f(0)$ does not make sense. Do you see a way of assigning a natural value to $f(0)$ using the graph?

[1] One may also enter code in several other languages

(ii) Find an approximate solution with four decimals to the equation $\cos(x) = x$.

Hint: This is an example of an equation, that can only be solved numerically. Try first plotting the graph of $f(x) = x - \cos(x)$ from 0 to 1. Then use a suitable function from the Sage guide.

(iii) Compute π with 100 decimals.

LLM

> Give me the sage code to compute pi with 100 decimals. I want a one line command.

♠

(1.7) EXERCISE.

Compute the sum

$$\frac{1}{\sqrt{1}+\sqrt{2}} + \frac{1}{\sqrt{2}+\sqrt{3}} + \frac{1}{\sqrt{3}+\sqrt{4}}.$$

What is the elegant answer? Explain!

Bonus question: Generalize your answer/method to computing the sum

$$\frac{1}{\sqrt{1}+\sqrt{2}} + \frac{1}{\sqrt{2}+\sqrt{3}} + \frac{1}{\sqrt{3}+\sqrt{4}} + \cdots + \frac{1}{\sqrt{N-1}+\sqrt{N}},$$

for $N = 5, 6, 7, \ldots$. ♠

We need a precise setup for communicating mathematics. This involves the introduction of propositional logic, predicates and sets. Let me emphasize, that this is an introductory course and not a rigorous introduction to mathematics. As such it is an organic approach, where I hope that you will return and fill out the gaps instead of getting overwhelmed by formal details already from the beginning.

However, the underlying goal is to show that mathematical precision and proofs are similar to constructing correct computer programs.

In fact, this whole first chapter may be viewed as the beginning of a computer program, where we state the exact definitions for use in the following chapters. We begin by introducing sets and numbers.

1.4 Numbers

A set is a collection of (mathematical) objects or elements. When defining a set we use the symbols $\{$ and $\}$ to denote the beginning and end of its definition. For example, $\{$N, i, e, l, s$\}$ is the set of characters in my first name and $\{8,0\}$ are the digits in the postal code for Aarhus C. The ordering in the listing of the elements is unimportant so that

$$\{\text{N, i, e, l, s}\} = \{\text{l, e, i, s, N}\}$$
$$\{8,0\} = \{0,8\}$$

are identical sets. If S is a set, we will use the notation $x \in S$ to denote that x is an element in S. For example, e$\in \{$N, i, e, l, s$\}$. The set with $\{\}$ with no elements is called the empty set. It is denoted $\emptyset$.

Later, we will introduce much more detail about sets. For now, we just need the basic notation for defining them.

Our fundamental mathematical objects in this course are numbers and we introduce them right away.

1.4.1 The natural numbers $\mathbb{N}$ and the integers $\mathbb{Z}$

The set of natural numbers is

$$\mathbb{N} = \{1, 2, 3, \dots\}. \tag{1.1}$$

The set of integers is

$$\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}. \tag{1.2}$$

These are *infinite* sets, since they contain infinitely many elements as indicated by the dots $\dots$

It makes sense to add and multiply two integers a and b. We will denote their addition or sum as $a+b$ and their multiplication or product as ab. A fundamental fact is that addition and multiplication are commutative i.e., $a+b=b+a$ and $ab=ba$.

Please notice right away that expressions like $a+b+c$ and abc are complete nonsense for three integers a, b and c. We only know how to add and multiply two integers, not three. A wonderful fact comes to our rescue:

$$\begin{aligned} (a+b)+c &= a+(b+c) \\ (ab)c &= a(bc). \end{aligned} \tag{1.3}$$

You get the same result no matter if you start adding (multiplying) a and b or b and c and then adding (multiplying) c or a. So we may write $a+b+c$ and abc as a placeholder for one of the two ways of computing this expression in (1.3).

As you can see we use the symbol $+$ for addition, but no symbol for multiplication. This is the convention in (clean) mathematics as opposed to the $a*b$ coming from computer algebra (except perhaps in Mathematica or Wolfram language, where space is allowed for multiplication). However, when one of the factors is an actual number, we will use $\cdot$, so that for example 7 times 9 is written as $7\cdot 9$ and a times 3 is written $a\cdot 3$.

1.4.2 The rational numbers $\mathbb{Q}$

The natural numbers $\mathbb{N}$ and the integers $\mathbb{Z}$ are well defined by their representations in (1.1) and (1.2).

(1.8) DEFINITION.

A rational number $a/b \in \mathbb{Q}$ consists of a numerator $a \in \mathbb{Z}$ and a denominator $b \in \mathbb{N}$.
If $a, c \in \mathbb{Z}$ and $b, d \in \mathbb{N}$. Then a/b and c/d are considered equal i.e.,

$$\frac{a}{b} = \frac{c}{d}$$

if and only if $ad = bc$.

(1.9) EXAMPLE.

So there are many different ways of representing a rational number, such as

$$\frac{3}{7} = \frac{6}{14} = \frac{9}{21}.$$

Here

$$\frac{3}{7} = \frac{9}{21} \qquad \text{since} \qquad 3\cdot 21 = 7\cdot 9.$$

In fact, a fraction stays the same when its numerator and denominator are multiplied by the same natural number. ♠

I will assume that you know how to add and multiply fractions, and that you **do not make mistakes like**

$$\frac{1}{2}+\frac{2}{3}=\frac{1+2}{2+3}=\frac{3}{5}.$$

In fact, if you temporarily forgot how to add fractions, you can use the wisdom in Definition 1.8. You can replace $\frac{1}{2}$ by $\frac{3}{6}$ and $\frac{2}{3}$ by $\frac{4}{6}$ and then add the numerators as in

$$\frac{1}{2}+\frac{2}{3}=\frac{3}{6}+\frac{4}{6}=\frac{3+4}{6}=\frac{7}{6}.$$

The computation above says that it is straightforward to add pizza slices of the same size (one sixth), but that you need to think a bit when adding one half pizza slice and two pizza slices of size one third. In general,

$$\frac{a}{b}+\frac{c}{d}=\frac{ad+bc}{bd} \qquad \text{and} \qquad \frac{a}{b}\frac{c}{d}=\frac{ac}{bd}.$$

(1.10) QUIZ.

Quiz not included in static version. ♠

1.4.3 The real numbers $\mathbb{R}$

(1.11) DEFINITION.

A real number is defined by

$$d_0.d_1d_2\ldots, \tag{1.4}$$

*where $d_0 \in \mathbb{Z}$ and $d_1, d_2, \ldots$ is an **infinite** sequence of integers (digits) in $\{0,1,2,3,4,5,6,7,8,9\}$.*

Informally (1.4) represents the real number

$$d_0+\frac{d_1}{10}+\frac{d_2}{100}+\cdots = d_0+d_1\cdot 10^{-1}+d_2\cdot 10^{-2}+\cdots$$

given by an infinite set of digits as opposed to a rational number, which is given finitely by an integer (numerator) and a natural number (denominator). The format in (1.4) is the usual (floating point) output from your pocket calculator or computer algebra system. For example,

$$\begin{aligned}
\frac{1}{3} &= 0.333\ldots \\
\frac{3}{7} &= 0.42857142857\ldots \\
\frac{3}{14} &= 0.21428571428571428571\ldots \\
\frac{22}{7} &= 3.14285714285714285714285714\ldots \\
\frac{355}{113} &= 3.14159292035398230088495575\ldots \\
\pi &= 3.14159265358979323846264338\ldots
\end{aligned}$$

The decimal expansion of $355/113$ above looks chaotic, but it eventually repeats itself after 112 digits. In fact the decimal expansion of every rational number is periodic i.e., it repeats itself from a certain point.

In the definition (1.4) of a real number, we are forced to make identifications as in the definition of a rational number. I will not go into details here, but just notice that the identification

$$0.99999\cdots = 1.000\ldots$$

is forced upon us: if

$$x = 9\cdot 10^{-1} + 9\cdot 10^{-2} + \cdots .$$

Then

$$10x = 9 + 9\cdot 10^{-1} + 9\cdot 10^{-2} + \cdots = 9 + x.$$

Therefore we must have $x = 1$.

A real number that is not rational is called irrational. There are many more irrational numbers than rational ones. Famous ones are $\sqrt{2}, \pi$ and e. The irrational number $\sqrt{2}$ is a root in the polynomial $x^2 - 2$ with integer coefficients (it is an algebraic number). The numbers e and π are not even algebraic (they are transcendental).

(1.12) EXERCISE.

Explore using prompting the history of numbers in mathematics. Also make a chatbot explain why a rational number must have a periodic decimal expansion. ♠

1.4.4 Arithmetic rules for numbers

(1.13) PROPOSITION.

Suppose that A is one of the sets $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ or $\mathbb{R}$. Then for numbers x, y, z all in A we have

(i) $1 \cdot x = x$

(ii) $xy = yx$

(iii) $x + y = y + x$

(iv) $(x + y) + z = x + (y + z)$

(v) $(xy)z = x(yz)$

(vi) $x(y + z) = xy + xz$

If A is one of $\mathbb{Z}, \mathbb{Q}$ or $\mathbb{R}$, then $0 \in A$ and for every $x \in A$,

resume $x + 0 = x$

resume $x + y = 0$ *for some number $y \in A$.*

Finally if A is $\mathbb{Q}$ or $\mathbb{R}$ and $x \in A$ is not 0,then

resume $xz = 1$ *for some number $z \in A$.*

The number z above is called the inverse *of x and is usually denoted x^{-1}. The number y above is called the* negative *of x and is usually denoted $-x$. We will also the symbol $-$ (minus) defined as an operation on two numbers $x, y \in A$ as*

$$x - y := x + (-y).$$

(1.14) EXERCISE.

Argue precisely that $-(x-y) = y-x$ for $x, y \in A$ using Proposition 1.13. ♠

The long list of arithmetic rules above may seem complicated at first, but they are just a formal version of what you already know, such as for example, $\pi + 0 = \pi$, $2 + y = 0$ if $y = -2$ and $3z = 1$ for $z = \frac{1}{3} = 3^{-1}$. Beware however, that the precision in Proposition 1.13 is necessary when programming a computer.

The rules iv and v are called the *associative* laws for addition and multiplication respectively. The rule vi is called the *distributive* law. It connects multiplication with addition.

(1.15) EXERCISE.

We know that zero times any number is zero. Deduce this from the rules in Proposition 1.13 starting with $0 + 0 = 0$. ♠

(1.16) EXERCISE.

Verify that vi is true for some specific non-zero numbers. Also convince yourself that WolframAlpha actually accepts space (between numbers and variables) as multiplication. ♠

(1.17) EXERCISE.

Suppose that $x, y, z \in \mathbb{Q}$ and $w = xy + xz$. It seems that computing w involves two multiplications and one addition. Multiplications are expensive operations on a computer. Is there a way of computing w with only one multiplication and one addition? ♠

(1.18) EXERCISE.

Suppose that $n \in \mathbb{N}$. Use the distributive law to show that

$$n^2 + n = n(n+1).$$

♠

We now move on to the (formal) language involved in reasoning about mathematics in general.

1.5 Propositional logic

A proposition is a (mathematical) statement that is true (t) or false (f). This could be a boolean expression in a computer program, like $1 < 2$.

Sage:

Interactive code not included in static version.

Later we will see propositions with variables in them like $x < 2$. These are called predicates.

Propositions can be combined into new (compound) propositions. Take for example the propositions

$$p : \text{it rains}$$
$$q : \text{it is cloudy.}$$

Then (p and q) is a perfectly good new proposition reading *it rains and it is cloudy*. The same goes for (if p then q), which reads *if it rains then it is cloudy*. The proposition (if q then p) reads *if it is cloudy then it rains*. This proposition is (clearly) false.

We need some notation to describe these compound propositions:

$p \wedge q$	p and q
$p \vee q$	p or q
$p \implies q$	if p then q
$\neg p$	not p

The compound propositions are either true(t) or false (f) depending on p and q. The dependencies are displayed in the *truth tables* below.

(1.19) DEFINITION.

p	q	$p \wedge q$
t	t	t
t	f	f
f	t	f
f	f	f

p	q	$p \vee q$
t	t	t
t	f	t
f	t	t
f	f	f

p	q	$p \implies q$
t	t	t
t	f	f
f	t	t
f	f	t

p	$\neg p$
t	f
f	t

The tables for the compound propositions $p \wedge q, p \vee q$ and also $\neg p$ are not too hard to grasp. The table for $p \implies q$ raises a few more questions. Why is $f \implies t$ true? I will not go into this at this point (see Example 1.32), but just point out that there are many explanations available online and, perhaps more importantly, refer you to Exercise 1.20.

(1.20) EXERCISE.

Suppose that we are presented with four cards

(1.5)

with a (natural) number on the front and the color blue or red on the back. In (1.5), the first and third cards are shown with their fronts facing up and the second and fourth cards are shown with their backs facing up.

A claim (proposition) is made that if a card has an even number on the front, then it must have the color blue on the back.

Your task is to verify this for the cards above. Of course you can do this by turning all four cards, but is there a way of checking this by turning less than four cards?

What if we add the claim, that if a card has the color blue on the back, then it must have an even number on the front?

Hint: Find two propositions p and q so that the claim reads $p \implies q$.

♠

(1.21) EXERCISE.

A prosecutor says to the defendant: "If you committed this crime you did not act alone". Explain why the defendant should not answer "no, that is not true" here. ♠

(1.22) EXERCISE.

Explain why Python/Sage thinks that the value[2] of

```
Interactive code not included in static version.
```

is False! Notice that you are dividing one by zero in the last "integer" above. ♠

1.5.1 Propositional logic as a formal language

The entities p and q above may have real world interpretations like *it rains* or *it is cloudy*, but we will view them as variables that can be assigned the values true or false. Independent of this assignment we define a proposition as follows.

(1.23) DEFINITION.

A proposition in the variables $x_1, \ldots, x_r$ is an expression involving the symbols $x_1, \ldots, x_r, (,), \neg, \wedge, \vee, \implies$ that can be generated using the rules below

(i) The variables $x_1, \ldots, x_r$ are (atomic) propositions.

(ii) If P is a proposition, then $(\neg P)$ is a proposition.

(iii) If P and Q are propositions, then $(P \wedge Q)$, $(P \vee Q)$ and $(P \implies Q)$ are propositions.

(1.24) EXAMPLE.

The expression $(x_1 \implies ((\neg x_2) \vee x_3))$ is a proposition. Let us see how it is generated using the rules in Definition 1.23.

(1) First, x_2 is a proposition using i.

(2) Then $(\neg x_2)$ is a proposition by using ii with $P = x_2$, since we know by 1 that P is a proposition.

(3) Since x_3 is a proposition by i, it follows that $((\neg x_2) \vee x_3)$ is a proposition using iii with $P = (\neg x_2)$ and $Q = x_3$, since we know by 2 that P is a proposition.

[2] Thanks to Gerth Brodal for pointing this out to me

(4) Finally, since x_1 is a proposition by i it follows by iii with $P = x_1$ and $Q = ((\neg x_2) \vee x_3)$ that

$$(x_1 \implies ((\neg x_2) \vee x_3))$$

is a proposition, since we know by 3 that Q is a proposition.

♠

(1.25) QUIZ.

Quiz not included in static version. ♠

1.5.2 Truth tables and equivalent propositions

Given a proposition, it makes sense to substitute values (t or f) for the variables and evaluate it using the rules in Definition 1.19, since the parentheses leave no ambiguity as to how the evaluation must take place. For a given proposition in the variables $x_1, \ldots, x_r$, there are 2^r ways of assignments to the set of variables. Each of these assignments results in the value true or false after evaluation. This is conveniently recorded in the *truth table* of the proposition as illustrated in the example below, where $r = 3$ so that the truth table has $2^3 = 8$ rows.

(1.26) EXAMPLE.

The truth tables corresponding to the propositions $(x_1 \wedge (x_2 \vee x_3))$ and $((x_1 \wedge x_2) \vee (x_1 \wedge x_3))$ are given below.

x_1	x_2	x_3	$(x_1 \wedge (x_2 \vee x_3))$
f	f	f	f
f	f	t	f
f	t	f	f
f	t	t	f
t	f	f	f
t	f	t	t
t	t	f	t
t	t	t	t

x_1	x_2	x_3	$((x_1 \wedge x_2) \vee (x_1 \wedge x_3))$
f	f	f	f
f	f	t	f
f	t	f	f
f	t	t	f
t	f	f	f
t	f	t	t
t	t	f	t
t	t	t	t

For example, if $x_1 = t, x_2 = f$ and $x_3 = t$, then

$$(x_1 \wedge (x_2 \vee x_3)) = (t \wedge (f \vee t)) = (t \wedge t) = t$$

and

$$((x_1 \wedge x_2) \vee (x_1 \wedge x_3)) = ((t \wedge f) \vee (t \wedge t)) = (f \vee t) = t.$$

♠

The two propositions in Example 1.26 have identical truth tables. In general, if two propositions P and Q have identical truth tables we call them *equivalent* and write

$$P \equiv Q.$$

In Example 1.26 we saw that

$$(x_1 \wedge (x_2 \vee x_3)) \equiv ((x_1 \wedge x_2) \vee (x_1 \wedge x_3)).$$

The definition below is very important too keep in mind.

(1.27) DEFINITION.

The notation $p \iff q$ *is used frequently. It means that both* $p \implies q$ *and* $q \implies p$ *are true i.e.,*

$$p \iff q \equiv (p \implies q) \wedge (q \implies p).$$

1.5.3 Using Sage to compute truth tables

Sage may be used to compute truth tables for propositions using the propositional calculus in Sage. Below is an example. Be sure to check how to enter $\wedge, \vee, \implies$ and $\neg$.

Interactive code not included in static version.

(1.28) EXERCISE.

Construct by hand the truth table for the proposition $(p \wedge q) \vee (\neg r)$. ♠

(1.29) EXERCISE.

Convince yourself either using Sage or by writing out truth tables that

(i) $(x_1 \implies x_2) \equiv ((\neg x_2) \implies (\neg x_1))$

(ii) $(\neg(x_1 \vee x_2)) \equiv ((\neg x_1) \wedge (\neg x_2))$

(iii) $\neg(x_1 \wedge x_2) \equiv (\neg x_1) \vee (\neg x_2)$

(iv) $x_1 \implies x_2 \equiv (\neg x_1) \vee x_2$

♠

1.5.4 Variables, predicates and quantification

(1.30) DEFINITION.

A predicate is a proposition depending on one or more variables.

Variables are fundamental in computer programs. In the predicate $x = 1$, x appears as a variable. Depending on what you substitute for x, the resulting proposition could be true or false or even meaningless. As an example, the latter case appears if x is replaced by the character 'a'. This is what computer scientists call a type error. You cannot compare a character with a digit.

(1.31) EXAMPLE.

If $p(n) = n$ is a prime number, then $p(3)$ is true, whereas $p(6)$ is false.

$$q(m,n) = p(m) \wedge (\neg p(n))$$

is a predicate in two variables m and n. Here $q(3,4)$ is true, whereas $q(5,7)$ is false. ♠

For every $\forall$ and there exists $\exists$

Suppose that we have a predicate $p(x)$, such that $p(x)$ is a proposition for $x \in S$, where S is some set. Then we define the proposition

$$\exists x \in S : p(x) \tag{1.6}$$

to be true if there exists $x \in S$, such that $p(x)$ is true. We let

$$\forall x \in S : p(x) \tag{1.7}$$

be the proposition defined by

$$\neg(\exists x \in S : \neg p(x)).$$

In other words, (1.7) says that $p(x)$ is true for every $x \in S$, since there does not exist $x \in S$ making $p(x)$ false. Le me be absolutely clear. To show that $\forall x \in S : p(x)$ is false, it is enough to find just a single $x \in S$ so that $p(x)$ is false.

(1.32) EXAMPLE.

Here is a statement about real numbers

$$x^2 = 1 \qquad \Longrightarrow \ (x-1)(x+1)(x-2) = 0 \tag{1.8}$$

This statement reads: no matter which real number x you pick, if $x^2 = 1$, then $(x-1)(x+1)(x-2) = 0$. We definitely want this to be true. Being true means that (1.8) must hold for all numbers x, also $x = 2$, which reads

$$2^2 = 4 = 1 \qquad \Longrightarrow \qquad (2-1)(2+1)(2-2) = 0 = 0$$

The above statement is an example of a false implies true statement, which we want to be true.

In general terms, in proving the statement that $p(x) \implies q(x)$ holds for every x in some set S, we are really only interested in $x \in S$ for which $p(x)$ is true, since $p(x)$ is our assumption. We still need $p(x) \implies q(x)$ to be true for $x \in S$ for which $p(x)$ is false. This is assured by the truth table for $\implies$, since $f \implies t$ and $f \implies f$ are both true. ♠

The following is an excerpt from the infamous *Beredskabsprøve Datalogi*.

(1.33) QUIZ.

Quiz not included in static version. ♠

1.5.5 Proofs and inference rules

A proof begins with an assumption P and proceeds with a sequence of logical steps called inference rules leading to a conclusion Q.

You have been taught how to solve equations in steps leading to a solution. Each step turns out to be an inference rule and the conclusion is the solution. Let us see how for the simple equation $x+1=2$. Formally we want to prove

$$\forall x \in \mathbb{R} : x+1=2 \implies x=1.$$

The first inference rule is $a=b \implies a+c=b+c$ i.e., we are allowed to add the same number to both sides of an equality. This implies

$$x+1=2 \implies (x+1)-1=2-1=1.$$

To be very precise we now use Proposition 1.13 iv as an inference rule i.e.,

$$(x+1)-1=1 \implies x+(1-1)=1.$$

Then we use Proposition 1.13? to conclude

$$x+(1-1)=1 \implies x+0=1$$

and then finally, we get by Proposition 1.13? that

$$x+0=1 \implies x=1.$$

So to solve the equation $x+1=2$, we are actually using four (!) inference rules along the way.

1.5.6 The use of implication ($\implies$) and bi-implication ($\iff$)

As you have seen, $\implies$ and $\iff$ are applied to link propositions in a logical argument. For example,

$$x+1=2 \iff x=1 \qquad \text{or} \qquad \forall x \in \mathbb{Z} : x+1=2 \iff x=1.$$

However, for $x^2=1 \implies x=1$ we cannot link the two propositions by $\iff$, simply because $\forall x \in \mathbb{Z} : x^2=1 \implies x=1$ is false (for $x=-1$).

(1.34) EXERCISE.

Prove that

$$(x<y) \wedge ((y+3)<(z+10)) \implies (x+37)<(z+44)$$

for every $x,y,z \in \mathbb{Z}$ (see Definition 1.70 with $A=\mathbb{Z}$ for the precise definition of $<$). Write out every inference rule!

Hint: You need to be very precise here. What does $x<y$ mean precisely if $x,y \in \mathbb{Z}$? In Definition 1.70 you will see that it means that $y-x \in \mathbb{N}$. From this you need to deduce

(i) $(x<y) \wedge (y<z) \implies x<z$

(ii) $x<y \implies x+z<y+z$

for every $x,y,z \in \mathbb{Z}$ ♠

(1.35) EXERCISE.

Try out

LLM

> Is
>
> $$x \geq 0 \iff x^2 \geq 0$$
>
> a true statement? Give me 3 carefully crafted exercises training me in distinguishing $\implies$ and $\iff$. Only use basic mathematics involving numbers and arithmetic operations. After each exercise stop, ask for the answer and give valueable feedback and guidance.

and go through the exercises given to you. ♠

1.5.7 More on mathematical proofs

Most professional mathematicians rarely think about the precise definition of a proof and would probably feel uncomfortable defining a proof precisely. During many years of training they have assimilated knowledge by experience. Therefore many proofs seem born out of witchcraft containing several magical devices.

However, many proofs appearing in respected mathematical journals, submitted by respected mathematicians, have turned out to contain errors. Recent developments in automated proof systems like Coq and LEAN show promise in checking proofs like for example the famous four color theorem. These automated proof systems build on dependent type theory, which we will not go into.

Informally a proof of a proposition q, consists in arguing that an implication $p \implies q$ is true by first assuming p. Usually this is done not only through one implication $p \implies q$, but through a series of intermediate implications

$$p \implies q_1 \implies q_2 \implies q_3 \implies \cdots \implies q_N,$$

where the last proposition q_N is q. If p is true, this will constitute a proof that $q_N = q$ is true. Just like in (1.14), there is an imprecision here. Can you tell what it is?

(1.36) EXAMPLE.

An integer is called even if it is divisible by 2. So the even integers are

$$\{\ldots,-4,-2,0,2,4,\ldots\}.$$

An integer is called odd if it is not even. So the odd integers are

$$\{\ldots,-5,-3,-1,1,3,\ldots\}.$$

Consider the proposition:

$$\forall n \in \mathbb{Z} : p(n) \implies p(n^2), \tag{1.9}$$

where $p(n) = (n \text{ is odd})$ i.e., the square of an odd integer is odd. This seems true for a first selection of examples: $3^2 = 9, 5^2 = 25, \ldots$.

What does it mean exactly for a number to be odd? This means that it is not divisible by 2 or that there exists another integer a, such that $n = 2a+1$. So

$$p(n) = \exists a \in \mathbb{Z} : n = 2a+1.$$

Therefore we need to show that

$$(\exists a \in \mathbb{Z} : n = 2a+1) \implies \left(\exists b \in \mathbb{Z} : n^2 = 2b+1\right).$$

Notice that I had to change a into b in the second proposition above. The two variables are not the same: a is associated with n and b is associated with n^2.

Let us assume that $n = 2a+1$. Now we need to argue that $n^2 = 2b+1$ for some $b \in \mathbb{Z}$. You stare at this for a while and notice that we should use the assumption $n = 2a+1$ in computing n^2:

$$n^2 = (2a+1)^2 = (2a)^2 + 2(2a) + 1^2 = 4a^2 + 4a + 1 = 2(2a^2 + 2a) + 1.$$

Thus, using our assumption we may conclude that if $n = 2a+1$, then

$$n^2 = 2b+1,$$

where $b = 2a^2 + 2a$. This completes the proof. ♠

The beauty here is that we have verified for all odd natural numbers that their square is odd. Not just a finite selection like $3, 7, 11, 13$.

In many ways a proof is like a detailed argument in a court case, except that the rules of mathematics are universal. You need the absolute truth in the court of mathematics (or science).

(1.37) EXERCISE.

Consider the proposition $q(n) = n$ is even. Prove that

$$\forall n \in \mathbb{Z} : q(n^2) \implies q(n).$$

Hint: Use that $q(n) = \neg p(n)$, where $p(n)$ is defined in Example 1.36. ♠

1.5.8 Proof by contradiction

A proposition p is either true or false. This seemingly obvious statement goes by the name of the law of excluded middle and dates back to the writings of Aristotle. The law of excluded middle is key in the example below, where you are left with the feeling that you have been deprived of a fair and genuine proof.

(1.38) EXAMPLE.

An irrational number is a (real) number that is not rational. It is a startling fact that such numbers exist, but they do! The square root $\sqrt{2}$ of two is an example. We will prove that there exists two irrational numbers α, β, such that α^β is rational.

Consider the proposition p given by

$$\gamma = \sqrt{2}^{\sqrt{2}} \text{is rational.}$$

Either p is true or false. If p is true we are done putting $\alpha = \beta = \sqrt{2}$. If not, then p must be false and γ is irrational. But then

$$\gamma^{\sqrt{2}} = \left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}} = (\sqrt{2})^{\sqrt{2}\cdot\sqrt{2}} = \sqrt{2}^2 = 2$$

and we are done putting $\alpha = \gamma$ and $\beta = \sqrt{2}$.

So which one is it? Is

$$\sqrt{2}^{\sqrt{2}}$$

rational or irrational? This is really advanced mathematics based on the Gelfond-Schneider theorem. ♠

The law of excluded middle can be turned into a powerful proof technique called *proof by contradiction*.

Suppose we wish to establish that p is true. Then we turn things upside down by assuming that p is false i.e., that $\neg p$ is true. If we then by logical deduction can show that

$$\neg p \implies q,$$

for some proposition q, which is demonstrably false, then $\neg p$ cannot be true (since true $\implies$ false is false). Therefore $\neg p$ must be false and p must be true by the law of the excluded middle. This technique is used all the time!

(1.39) EXAMPLE.

Let us use proof by contradiction to show that the proposition p : ($\sqrt{2}$ is an irrational number) is true. Assuming that p is false, we must have that $\neg p$ is true. But $\neg p$ is the proposition p_1 (that $\sqrt{2}$ is a rational number)

$$\exists m,n \in \mathbb{N} : \sqrt{2} = \frac{m}{n}.$$

Here $p_1 \iff p_2$, where p_2 is the proposition

$$\exists m,n \in \mathbb{N} : \left(\sqrt{2} = \frac{m}{n}\right) \wedge ((m \text{ is odd}) \vee (n \text{ is odd})),$$

since we can assume that 2 is not a common divisor of m and n by Definition 1.8. However,

$$\sqrt{2} = \frac{m}{n} \iff \sqrt{2}n = m \iff 2n^2 = m^2 \implies m \text{ is even.}$$

The last implication above follows from Exercise 1.37. If m is even, then $m = 2k$ for some $k \in \mathbb{N}$. Therefore $m^2 = 4k^2$ and

$$2n^2 = 4k^2 \iff n^2 = 2k^2$$

so that n is also even. We have proved that p_2 implies the proposition p_3 given by

$$(m \text{ is even}) \wedge (n \text{ is even}).$$

Since we are assuming that p_2 is true, we must have that p_3 is false. However we have shown that $p_2 \implies p_3$ is true. But $t \implies f$ is a false. Therefore we must have that p_2 is false and therefore that p_1 is false. But then according to the law of the excluded middle, we must have that $p = \neg p_1$ is true. ♠

(1.40) EXERCISE.

Consider the first n prime numbers

$$p_1 = 2, p_2 = 3, p_3 = 5, \ldots, p_n.$$

Check that

$$\begin{aligned} &p_1 \\ &p_1\,p_2 + 1 \\ &p_1\,p_2\,p_3 + 1 \\ &p_1\,p_2\,p_3\,p_4 + 1 \end{aligned}$$

are prime numbers by using the Sage window below (`factor` gives the prime factorization of a natural number).

Interactive code not included in static version.

Is it true in general that

$$p_1\,p_2 \cdots p_n + 1$$

is a prime number?

Assume that we know that every natural number must be divisible by a prime number. Prove that there are infinitely many prime numbers using proof by contradiction.

Hint: Show how the assumption that there are only finitely many prime numbers say

$$p_1, p_2, \ldots, p_n$$

leads to a contradiction by using that the natural number

$$p_1\, p_2 \ldots p_n + 1$$

must be divisible by a prime number. ♠

1.6 More on sets

Propositions are important, but are confined by the binary values of true and false. We would like to work mathematically with objects like integers, floating point numbers, neural networks, computer programs and so on.

1.6.1 Objects and equality

One of the cornerstones of modern mathematics is deciding when two objects are the same i.e., given two objects A and B, deciding whether the proposition $A = B$ is true of false. Oftentimes an algorithm for evaluating $A = B$ is needed.

You may laugh here, but this is not always that easy. Even though objects appear different they are the same as in, for example the propositions

$$\frac{105}{189} = \frac{35}{63} \qquad \text{and} \qquad \sin\left(\frac{\pi}{2}\right) = 1.$$

The first proposition above is an identity of fractions (rational numbers). The second is an identity, which calls for knowledge of the sine function and real numbers. Each of these identities calls for some rather advanced mathematics. The first proposition is true in a very precise way, since $105 \cdot 63 = 189 \cdot 35$.

(1.41) EXERCISE.

Interactive code not included in static version.

Use the Sage window above to reason about equality in the quiz below. In each case describe the objects i.e., are they numbers, symbols, etc.? Also, please check your computations by hand with the old fashioned paper and pencil, especially $(a+b)(a-b)$.

Quiz not included in static version.

♠

(1.42) EXERCISE.

You know that $(a+b)^2 = a^2+2ab+b^2$. Use Sage to find a similar identities for $(a+b)^3$ and $(a+b)^4$.

Hint: Go back and look at (the beginning of) Exercise 1.41. ♠

For two objects A and B we will use the notation $A \neq B$ for the proposition $\neg(A = B)$.

We have already defined a set (informally) as a collection of distinct objects or *elements*. We introduce some more set theory here. A set is also an object as described in section 1.6.1 and it makes sense to ask when two sets are equal.

(1.43) DEFINITION.

Two sets A and B are equal i.e., $A = B$ if they contain the same elements.

An example of a set could be the set $\{1,2,3\}$ of natural numbers between 0 and 4. Notice again that we use the symbol "{" to start the listing of elements in a set and the symbol "}" to denote the end of the listing. Notice also that (by our definition of equality between sets), the order of the elements in the listing does not matter i.e.,

$$\{1,2,3\} = \{2,3,1\}.$$

We are also not allowing duplicates like for example in the listing $\{1,2,2,3,3,3\}$ (such a thing is called a multiset).

An example of a set not involving numbers could be the set of letters

$$S = \{A,n,e,x,a,m,p,l,c,o,u,d,b,t,h,s,r,i\}$$

used in this sentence. The number of elements in a set S is called the *cardinality* of the set. We will denote it by $|S|$.

To convince someone beyond a doubt (we will talk about this formally later in this chapter) that two sets A and B are equal, one needs to argue that if x is an element of A, then x is an element of B and the other way round, if y is an element of B, then y is an element of A. If this is true, then A and B must contain the same elements.

(1.44) EXERCISE.

Give a precise reason as to why the two sets $\{1,2,3\}$ and $\{1,2,4\}$ are not equal. Is it possible for a set with 5 elements to be equal to a set with 7 elements? ♠

Sets may be explored using (only) python. This is illustrated in the snippet below.

Interactive code not included in static version.

(1.45) EXERCISE.

Come up with three lines of Sage code that verifies $\{1,2,3\} \neq \{1,2,4\}$. Try it out. ♠

The empty set

There is a unique set containing no or zero elements. This set is called the empty set and is denoted $\emptyset$ i.e.,

$$\emptyset = \{\} \qquad \text{and} \qquad |\emptyset| = 0.$$

Interactive code not included in static version.

(1.46) EXERCISE.

For some reason (perhaps a good one) python does not accept $\{\}$ as input for the empty set. Why is this? Evaluate the python snippet below and explain.

Interactive code not included in static version.

♠

1.6.2 The symbols $\in$ and $\notin$

The symbol $\in$ is ubiquitous in set theory (and mathematics). If A is a set, then

$$x \in A \tag{1.10}$$

is a proposition. It is true if *x is an element of* or *belongs to A*. The notation

$$x \notin A$$

is defined by the proposition $\neg(x \in A)$. Also, as a bit of short hand notation, we will write

$$a_1, a_2, \ldots, a_n \in A \qquad \text{for the proposition} \qquad (a_1 \in A) \wedge (a_2 \in A) \wedge \cdots \wedge (a_n \in A).$$

Belongs to ($\in$) is straightforward in python.

Interactive code not included in static version.

(1.47) QUIZ.

Quiz not included in static version. ♠

1.6.3 Subsets

If A and B are sets, then $A \subseteq B$[3] means that every element of A is an element of B. So $A \subseteq B$ is a placeholder for the proposition

$$\forall x \in A : x \in B$$

[3] At times, the symbol $\subset$ is used instead of $\subseteq$. In our context these two symbols mean the same. However, the notation $A \subsetneq B$ means that $A \subseteq B$ and $A \neq B$. For example, $\{1,2,3\} \subseteq \{1,2,3\}$ and $\{1,2,3\} \subset \{1,2,3\}$.

In this case we say that *A is a subset of B*. We also use the notation $A \subsetneq B$ to indicate that $A \subseteq B$ and $A \neq B$. In this case we say that A is a *strict* subset of B.

(1.48) EXERCISE.

List the subsets of $\{1,2\}$. How many are there? ♠

(1.49) EXERCISE.

It turns out that the empty set $\emptyset$ is a subset of any set.

Interactive code not included in static version.

Explain why this is so using the definition of $\subseteq$.

LLM

Explain precisely in terms of propositions and logic why the empty set is a subset of any given set.

♠

(1.50) EXERCISE.

Below Sage (not python) will list all subsets of the set $\{1,2,3\}$. Before pressing the Compute button, try to write them down on your own.

Interactive code not included in static version.

List all the subsets of a set with five elements. In general, how many subsets does a set with n elements have? ♠

(1.51) QUIZ.

Quiz not included in static version. ♠

(1.52) QUIZ.

Quiz not included in static version. ♠

1.6.4 Set-builder notation

If S is a set and $p(x)$ a predicate for $x \in S$, then we build the subset

$$\{x \in S \mid p(x)\} \subseteq S \tag{1.11}$$

of $x \in S$ such that $p(x)$ is true.

(1.53) EXAMPLE.

Suppose that $S = \{-2,-1,2,3\}$ and

$$p(x) = x \text{ is positive.}$$

Then

$$\{x \in S \mid p(x)\} = \{2,3\} \subseteq S.$$

♠

Python: This notation has found its way to several programming languages like list comprehension in python.

Interactive code not included in static version.

Suppose that $p_1(x), \ldots, p_n(x)$ are predicates with a variable x taking values in S, then we often use the notation (using , instead of $\wedge$)

$$\{x \in S \mid p_1(x), \ldots, p_n(x)\} \quad \text{for} \quad \{x \in S \mid p_1(x) \wedge \cdots \wedge p_n(x)\}.$$

(1.54) EXERCISE.

List the elements in the following subsets.

(i)

$$\{x \in \mathbb{Z} \mid x^2 - 5x + 6 = 0\}.$$

(ii)

$$\{(x,y) \mid x \in \mathbb{Z}, y \in \mathbb{Z}, x^2 + y^2 < 5\}.$$

♠

(1.55) EXERCISE.

Consider the predicate $q(n)$

$$n \text{ and } n+2 \text{ are both prime numbers.}$$

Write down the elements in

$$\{n \in \mathbb{N} \mid q(n) \wedge n \leq 50\}.$$

Is

$$\{n \in \mathbb{N} \mid q(n)\}$$

an infinite set?

Hint: Explore the fascinating world of prime numbers and learn about twin primes. ♠

(1.56) EXERCISE.

You have previously encountered systems of linear equations like

$$\begin{aligned} x+y &= 3 \\ 3x-y &= 5. \end{aligned} \tag{1.12}$$

The solutions to (1.12) can be identified with a subset of $\mathbb{R}^2$. Define this subset precisely i.e., write the subset as

$$\{(x,y) \in \mathbb{R}^2 \mid p(x,y)\},$$

where $p(x,y)$ is a predicate in the variables $x, y \in \mathbb{R}$. ♠

(1.57) EXERCISE.

Suppose that $X = \mathbb{R}$ and

$$Y = \{x \in \mathbb{R} \mid x > 0, x < 2\}.$$

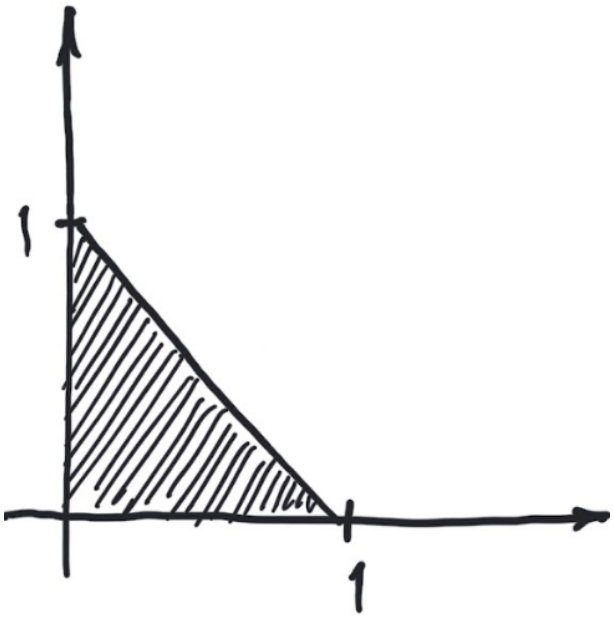
Then write down precisely what $X \setminus Y = \{x \in X \mid x \notin Y\}$ is i.e., find suitable predicates q_1, q_2 in the variable x, such that $q(x) = q_1(x) \vee q_2(x)$ and

$$X \setminus Y = \{x \in \mathbb{R} \mid q(x)\}.$$

♠

(1.58) EXERCISE.

Consider the subset S of $\mathbb{R}^2$ pictured in the drawing below



Express S as

$$S = \{(x,y) \in \mathbb{R}^2 \mid p_1(x,y), p_2(x,y), p_3(x,y)\},$$

where p_1, p_2, p_3 are predicates in the variables x, y.

Hint: A predicate in the variables x, y could be something like

$$x - y \geq 17.$$

Express $\mathbb{R}^2 \setminus S$ as

$$\{(x,y) \in \mathbb{R}^2 \mid p(x,y)\},$$

where

$$p(x,y) = q_1(x,y) \vee q_2(x,y) \vee q_3(x,y)$$

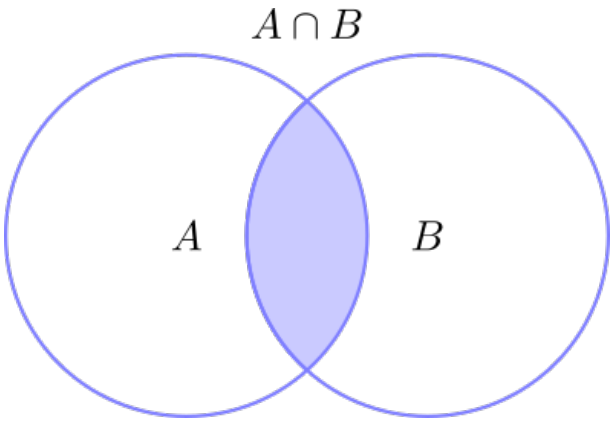
and q_1, q_2 and q_2 are suitable predicates in the variables x, y. ♠

1.6.5 Intersections, unions and the symbols $\cap$, $\cup$ and $\setminus$

Suppose that we have two sets A and B. Then the *intersection* $A \cap B$ is the set consisting of the elements in both A and B i.e.,

$$A \cap B = \{x \mid (x \in A) \wedge (x \in B)\}.$$

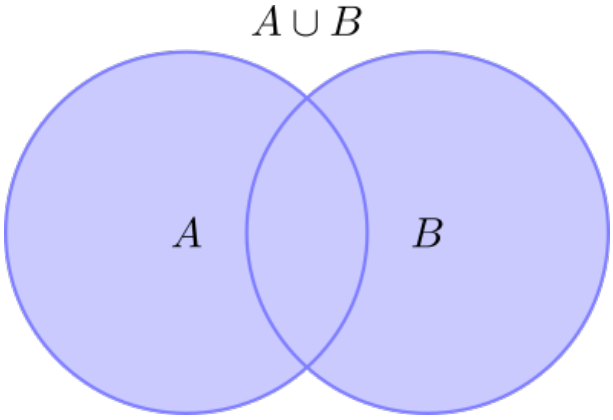
This is illustrated in the socalled Venn diagram below.



The *union* $A \cup B$ is the set consisting of the elements in A or B i.e.,

$$A \cup B = \{x \mid (x \in A) \vee (x \in B)\}.$$

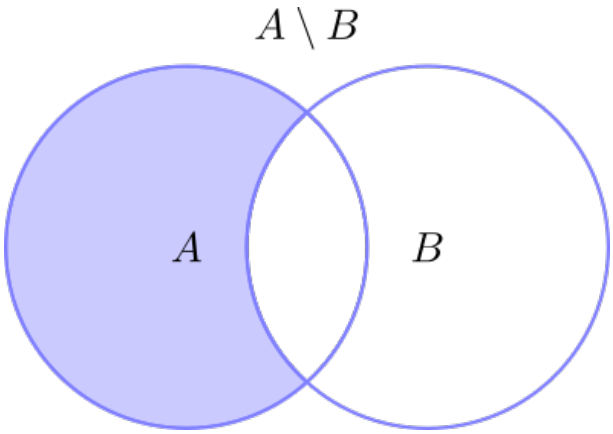
This is illustrated in the Venn diagram below.



Lastly, the difference $A \setminus B$ (between A and B) consists of the elements in A not contained in B i.e.,

$$A \setminus B = \{x \mid (x \in A) \wedge (x \notin B)\}.$$

This is illustrated in the Venn diagram below.



Python: You should experiment using the python window below to get a feeling for these three operations.

Interactive code not included in static version.

(1.59) EXERCISE.

Suppose that $A = \{1,2,3,4,5\}$, $B = \{-1,3,4,7\}$ and $C = \{2,3,8,9\}$. What is $((A \cup B) \setminus C) \setminus B$? ♠

(1.60) EXERCISE.

Let $A = \{1,2,3\}$, $B = \{3,4,5\}$ and $C = \{0,1,5\}$. Verify by hand (no computer) that

(i) $A \cup B = \{1,2,3,4,5\}$.

(ii) $A \cap B = \{3\}$.

(iii) $A \cap (B \cap C) = \emptyset$.

(iv) $B \setminus A = \{4,5\}$.

(v) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

♠

(1.61) EXERCISE.

Given two sets A and B, is it true that $A \cap B = B \cap A$ and $A \cup B = B \cup A$?
What about $A \setminus B = B \setminus A$?
Suppose that A and B are two finite sets. Is it true that

$$|A \setminus B| = |A| - |B|?$$

What about

$$|A \cup B| = |A| + |B|?$$

Seriously, both formulas are wrong. Can you come up with the correct version of the formula for $|A \cup B|$? Use your correct formula to find a formula for

$$|A \cup B \cup C|$$

viewing $A \cup B$ as the first set and C as the second set. Here you need the formula

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C).$$

Why is this formula true? Finally, explain why

$$C \setminus (A \cap B) = (C \setminus A) \cup (C \setminus B).$$

Hint: You may find it useful to notice that two sets S_1, S_2 are equal i.e, $S_1 = S_2$ if and only if

$$x \in S_1 \iff x \in S_2.$$

Also,

$$\begin{aligned}
x \in S_1 \cup S_2 &\iff x \in S_1 \vee x \in S_2 \\
x \in S_1 \cap S_2 &\iff x \in S_1 \wedge x \in S_2 \\
x \in S_1 \setminus S_2 &\iff x \in S_1 \wedge x \notin S_2 \\
x \notin S_1 &\iff \neg(x \in S_1).
\end{aligned}$$

♠

(1.62) EXERCISE.

There is one more operation called the symmetric difference between two sets A and B. It is denoted $A \Delta B$. Experiment in the python window below to find out exactly what it does. Is it true that $A \Delta B = B \Delta A$?

Interactive code not included in static version.

♠

The following is an excerpt from the infamous *Beredskabsprøve Datalogi*.

(1.63) QUIZ.

Quiz not included in static version. ♠

1.6.6 Pairs, triples and tuples

Oftentimes we want to consider more than one variable as input to a predicate. It is convenient to group the variables into one object consisting of the variables. This is done using tuples.

Given two sets A and B, we combine two elements $a \in A$ and $b \in B$ into a pair (a,b), which is an element of the Cartesian product

$$A \times B = \{(a,b) \mid a \in A, b \in B\}.$$

of A and B. This is a new set built from A and B.

(1.64) EXAMPLE.

If $A = \{1,2\}$ and $B = \{1,2,3\}$, then

$$A \times B = \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3)\}.$$

♠

(1.65) EXERCISE.

Consider two pairs (a,b) and (c,d) each from in $A \times B$. When is $(a,b) = (c,d)$? ♠

Python: The Cartesian product can be computed in python as shown below.

Interactive code not included in static version.

There is no need to restrict ourselves to pairs. We might as well consider triples $A \times B \times C$ i.e., the set of all (a,b,c), where A, B and C are sets, or for that matter general tuples

$$(a_1, a_2, \ldots, a_n) \in A_1 \times A_2 \times \cdots \times A_n \tag{1.13}$$

of any length $n \in \mathbb{N}$, where $a_1 \in A_1, a_2 \in A_2, \ldots, a_n \in A_n$. Based on the above example with tuples we have,

$$\{0\} \times \{1,2\} \times \{1,2,3\} =$$
$$\{(0,1,1),(0,1,2),(0,1,3),(0,2,1),(0,2,2),(0,2,3)\}.$$

You may check this using the python snippet below.

Interactive code not included in static version.

(1.66) DEFINITION.

For a given set A and $n \in \mathbb{N}$ we define the n-fold cartesian product of A as

$$A^n = \underbrace{A \times A \times \cdots \times A}_{n \text{ times}}.$$

(1.67) EXERCISE.

Formally $\mathbb{R}^2$ is the set of pairs (a,b), where $a,b \in \mathbb{R}$. Is there a natural way of drawing elements in $\mathbb{R}^2$? ♠

(1.68) EXERCISE.

Let A and B be two sets. Is $A \times B = B \times A$?

Let X be any set. What is $\emptyset \times X$?

Let A, B, C and D be four sets. Is

$$(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)?$$

Is

$$(A \times B) \setminus (C \times D) = (A \setminus C) \times (B \setminus D)?$$

Hint: See Exercise 1.69. ♠

(1.69) EXERCISE.

Use python to solve Exercise 1.68 by playing with (and extending) the code below.

Interactive code not included in static version.

♠

1.7 Ordering numbers

Let us be a little rigorous and introduce the (usual) ordering on our numbers with addition and multiplication using almost full blown mathematical formalities. The natural order $<$ on the natural numbers $\mathbb{N}$ is

$$1 < 2 < 3 < \cdots$$

For the numbers $\mathbb{Q}$ and $\mathbb{R}$ it is less obvious how to define an order. Mathematical simplicity comes to the rescue here. It is enough to define what (the) positive numbers are! We want (the) positive numbers to satisfy the conditions below.

(1.70) DEFINITION.

A subset A_+ of positive numbers in a set A of numbers must satisfy

(i) For every $x \in A$ one and only one of the following conditions must hold

(alpha) $-x \in A_+$

(blphb) $x = 0$

(clphc) $x \in A_+$

(ii) If $x, y \in A_+$, then $x + y \in A_+$ and $xy \in A_+$.

For a set A_+ of positive numbers in A, we define

$$x < y \iff y - x \in A_+$$

and

$$x \leq y \iff (x = y) \vee (x < y).$$

We will write $x > 0$ if $x \in A_+$ and $x < 0$ if $-x \in A_+$.

(1.71) REMARK.

Notice that we only use arithmetic operations to define orders on numbers in Definition 1.70. This is also how computers compare numbers algorithmically. Also if $A = \mathbb{Z}$, putting $A_+ = \mathbb{N}$ makes all of the conditions in Definition 1.70 hold. If you are given an integer, it is 0, positive or negative. This is the content of i in Definition 1.70. Also given two natural numbers, their product and sum are also natural numbers. This is the content of ii in Definition 1.70.

(1.72) EXERCISE.

Suppose that $a, x, y \in A$, where A is a set of numbers and $<$ given by a subset of positive numbers A_+ as in Definition 1.70. Prove that

(i) $(x < y) \wedge (y < z) \implies x < z$

(ii) $(a > 0) \wedge (x < y) \implies ax < ay$

(iii) $(a < 0) \wedge (x < y) \implies ay < ax$

♠

1.7.1 Ordering $\mathbb{Z}$

As we saw in Remark 1.71, the natural order on $\mathbb{Z}$ is defined by $\mathbb{Z}_+ = \mathbb{N}$, so that $x < y$ if $y - x \in \mathbb{N}$ for $x, y \in \mathbb{Z}$. This completely agrees with our preconception that

$$\cdots < -3 < -2 < -1 < 0 < 1 < 2 < \cdots \tag{1.14}$$

To be precise, writing $\cdots < -3 < -2 < -1 < 0 < 1 < 2 < \cdots$ is nonsense, since $<$ is only defined for two integers.

(1.73) EXERCISE.

How is one supposed to interpret $0 < 1 < 2$ for example? Go ahead and formulate (1.14) correctly comparing only two integers at a time. How does Python/Sage interpret $-3 < -2 < -1 < 0 < 1 < 2$? Find out using the Sage snippet below.

Interactive code not included in static version.

What about $1 < 5 > 3 < 4$? What about $0 < 1 > 2$? ♠

(1.74) QUIZ.

Quiz not included in static version. ♠

1.7.2 Ordering $\mathbb{Q}$

We define the positive rational numbers as

$$\mathbb{Q}_+ = \left\{ \frac{m}{n} \in \mathbb{Q} \middle| m > 0 \right\} = \left\{ 1, \frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \dots \right\}.$$

One can check that $\mathbb{Q}_+$ satisfies the conditions in Definition 1.70. So formally we get

(1.75) PROPOSITION.

For $\frac{a}{b}, \frac{c}{d} \in \mathbb{Q}$,

$$\frac{a}{b} < \frac{c}{d} \qquad \Longleftrightarrow \qquad ad < bc \qquad (\text{in } \mathbb{Z}).$$

Proof. We must check when

$$\frac{c}{d} - \frac{a}{b} = \frac{bc - ad}{bd} \in \mathbb{Q}_+.$$

This happens precisely when the numerator $bc - ad \in \mathbb{N}$ or $bc - ad > 0$. Therefore the condition in the proposition is satsified. □

(1.76) EXERCISE.

Use proof by contradiction (see section 1.5.8) to show precisely that there does not exist a smallest positive rational number. ♠

(1.77) EXERCISE.

Suppose that $a/b \in \mathbb{Q}$.

(i) Is it true in general that

$$\frac{a}{b} \leq \frac{a+1}{b+1}?$$

(ii) Is it true in some cases?

(iii) Suppose that $n \in \mathbb{N}$. Prove that

$$1 - \frac{a+n}{b+n} = \frac{b-a}{b+n}.$$

(iv) What happens to the rational number

$$\frac{a+n}{b+n}$$

when $n \in \mathbb{N}$ grows and becomes very big?

Interactive code not included in static version.

♠

Using Definition 1.75, you can check that $\frac{2}{3} < \frac{5}{7}$, since

$$2 \cdot 7 < 3 \cdot 5.$$

An easy, but surprising, way of finding a rational number strictly between these two is adding their numerators and denominators:

$$\frac{2}{3} < \frac{2+5}{3+7} < \frac{5}{7}.$$

We wil try to explain the first inequality in mathematical general terms going through a rather formal proof consisting of five steps. These steps are given in the quiz below. Your task is to drag from the left and drop them to the right in an order, so that the proof makes sense.

After that you are supposed, on your own, to write down a precise proof of the second inequality.

(1.78) QUIZ.

Quiz not included in static version. ♠

(1.79) EXERCISE.

Similarly to the quiz above, assume that

$$\frac{a}{b} < \frac{c}{d}.$$

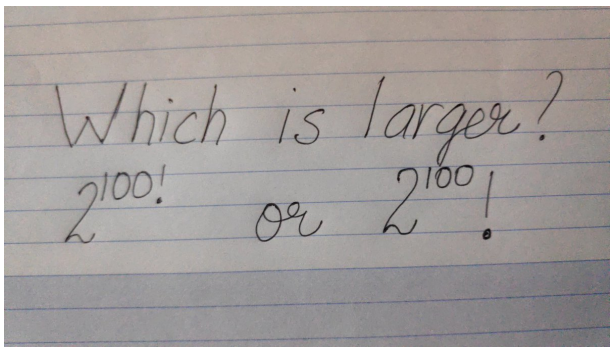
Write down a precise argument showing that

$$\frac{a+c}{b+d} < \frac{c}{d}.$$

♠

(1.80) EXERCISE.

On Twitter, Raman Gupta posted the note below



For a natural number $m \in \mathbb{N}$,

$$m! = m(m-1)(m-2)\cdot\cdots\cdot 2\cdot 1.$$

For example, $3! = 6$ and $5! = 120$. What is the answer for the question in the note?

Hint:

Experiment a bit with Sage: define a function $f(n)$, which computes

$$2^{n!} - 2^n!$$

Then look at

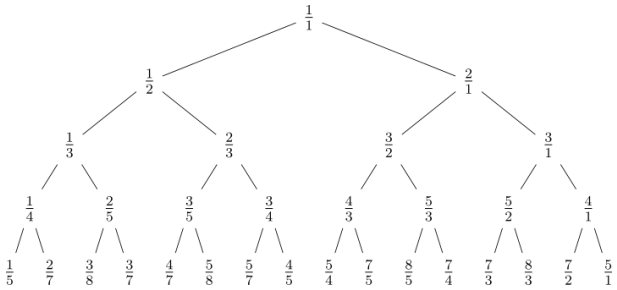
$$f(1), f(2), f(3), f(4), f(5), \ldots$$

♠

The exercise below shows that our trick for finding rational numbers in between two given rational numbers can be made into a machine for generating all positive rational numbers!

(1.81) EXERCISE.

Can you spot the system in the fractions in the diagram below?



Once you see the system, extend the diagram with the next level downwards. Is every positive fraction present in this diagram if one keeps adding levels?

Hint: Suppose that

$$\frac{p}{q} < \frac{r}{s}$$

and $qr - sp = 1$. Then for

$$\frac{p}{q} < \frac{p+r}{q+s} < \frac{r}{s},$$

we have $q(p+r)-(q+s)p=1$ and $(q+s)r-(p+r)s=1$. If $\frac{a}{b}$ is a positive fraction, such that

$$\frac{p}{q}<\frac{a}{b}<\frac{r}{s},$$

show that

$$a+b=(r+s)(qa-bp)+(p+q)(br-as)\geq p+q+r+s.$$

♠

1.7.3 Ordering $\mathbb{R}$

For the real numbers we define the positive numbers as

$$\mathbb{R}_+=\{x\in\mathbb{R}\mid x=d_0.d_1d_2\ldots,d_0\geq 0,x\neq 0\}.$$

Even though we have not precisely defined addition and multiplication of the real numbers, we claim that this definition of $\mathbb{R}_+$ satisfies the conditions of Definition 1.70.

One may prove that for every $x,y\in\mathbb{R}_+$, there exists $N\in\mathbb{N}$, such that $Nx>y$. This is the archimedean property of the real numbers.

(1.82) EXERCISE.

Given two distinct real numbers $\xi_1<\xi_2$. Prove that there exists a rational number $r\in\mathbb{Q}$, such that

$$\xi_1<r<\xi_2.$$

♠

One other crucial property is the completeness of $\mathbb{R}$. It says that a non-empty subset $S\subseteq\mathbb{R}$ with an upper bound B i.e., $\forall x\in S:x\leq B$, always has a smallest upper bound. The rational numbers do not share this property, since for example

$$S=\{x\in\mathbb{Q}\mid x^2<2\}$$

does not have a smallest upper bound inside $\mathbb{Q}$.

1.8 Proof by induction

A precocious Gauss[4] proved the formula

$$1+2+\cdots+n=\frac{n(n+1)}{2} \tag{1.15}$$

at the age of seven displaying remarkable ingenuity for his age. Lesser mortals usually use induction to prove this formula. Gauss was asked along with his classmates to compute the sum of all natural numbers $1,2,\ldots,100$. Using his formula he quickly came up with the correct answer 5050. His classmates had to work for the entire lesson.

Suppose that the formula in (1.15) is viewed as a proposition $p(n)$. To prove the formula we need to prove it for all natural numbers (you can easily see that $p(1)$ and $p(2)$ are true) i.e., we need to prove

$$\forall n\in\mathbb{N}:p(n).$$

An induction proof is a way of proving this statement by showing two things:

[4] See the article Gauss's Day of Reckoning for some history of this anecdote.

(i) $p(1)$

(ii) $\forall n \in \mathbb{N} : p(n) \implies p(n+1)$

These two statements ensure that $p(1) \implies p(2)$. Therefore $p(2)$ must be true, since we assumed $p(1)$ true from the beginning. Similarly $p(2) \implies p(3)$ ensures that $p(3)$ is true and so on. In fact we have proved $p(n)$ for every $n \in \mathbb{N}$ using this technique. One can prove this using proof by contradiction and that every non-empty subset of $\mathbb{N}$ has a first element. In general if S is a subset of set with an order $\leq$, then $s \in S$ is called a first element if

$$\forall x \in S : s \leq x.$$

A crucial rule (or axiom) is that every non-empty subset of $\mathbb{N}$ has a first element! Notice that this is false for $\mathbb{Z}$.

(1.83) THEOREM.

Suppose that $p(n)$ are infinitely many propositions given by $n \in \mathbb{N}$. Then

$$\forall n \in \mathbb{N} : p(n)$$

is true if

(i) $p(1)$ is true.

(ii) $(\forall n \in \mathbb{N} : p(n) \implies p(n+1))$ is true.

Proof. Suppose by contradiction that there exists $n \in \mathbb{N}$, such that $p(n)$ is false. Then the subset

$$S = \{n \in \mathbb{N} \mid \neg p(n)\} \subseteq \mathbb{N}$$

is non-empty. Therefore it has a first element $n_0 \in S$. Here $n_0 > 1$, since $p(1)$ is assumed to be true. So we know that $p(n_0 - 1)$ is true and that $p(n_0 - 1) \implies p(n_0)$ is true. But the latter implication is a contradiction, since true implies false is false. □

Let us see how an induction proof plays out in the above example with the statement $p(n)$ that

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}. \tag{1.16}$$

Clearly $p(1)$ is true. We need to prove $p(n) \implies p(n+1)$, so we assume that $p(n)$ holds i.e., that (1.16) is true. Then we may add $n+1$ to both sides of (1.16) to get

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

Here the right hand side can be rewritten as

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

which is exactly what we want. This is the conjectured formula for the sum of the numbers $1, 2, \ldots, n, n+1$. Therefore we have proved that $p(n) \implies p(n+1)$ and the induction proof is complete.

(1.84) EXAMPLE.

For a real number $r \neq 1$, the extremely useful formula

$$1 + r + \cdots + r^n = \frac{1 - r^{n+1}}{1 - r} \tag{1.17}$$

holds. Let us prove this formula by induction. For $n = 1$ this amounts to the identity

$$1 + r = \frac{1 - r^2}{1 - r},$$

which is true since $1 - r^2 = (1 + r)(1 - r)$. We let $p(n)$ denote the identity in (1.17). We have seen that $p(1)$ is true. The induction step consists in proving $p(n) \implies p(n+1)$. We can prove this by adding r^{n+1} to the right hand side in (1.17):

$$\frac{1 - r^{n+1}}{1 - r} + r^{n+1} = \frac{1 - r^{n+1} + (1 - r)r^{n+1}}{1 - r} = \frac{1 - r^{n+2}}{1 - r}. \tag{1.18}$$

Real life application: In order to pay for a house you borrow P DKK at an interest of r per year. You want to pay off your debt over N years by paying a fixed amount each year. How much is the fixed yearly amount you need to pay?

Let us analyze the setup: suppose that the fixed yearly amount is Y. We will find an equation giving us Y in terms of P, N and r. Put $q = 1 + r$.

After one year you owe

$$qP - Y.$$

After two years you owe

$$q(qP - Y) - Y.$$

After three years you owe

$$q(q(qP - Y) - Y) - Y.$$

In general after n years you owe

$$q^n P - Y(1 + q + \cdots + q^{n-1}).$$

Since we want to be debt free after N years, the yearly payment will have to satisfy

$$q^N P = Y(1 + q + \cdots + q^{N-1}).$$

By the formula (1.17), we get

$$q^N P = Y \frac{1 - q^N}{1 - q}.$$

Here Y can be isolated giving the formula

$$Y = \frac{rP}{1 - \left(\frac{1}{1+r}\right)^N}.$$

With the current (August 2025) interest rate around four percent, you pay a fixed monthly amount of around 4770 DKK for borrowing one million DKK over 30 years.

Interactive code not included in static version.

♠

(1.85) EXERCISE.

Verify the computation (induction step) in (1.18) i.e., explain the operations used to go from the left to the right of the two equalities. ♠

(1.86) EXERCISE.

Locate the mistake in the following fake induction proof of the curious fact that $2^n = 2$ for every $n \in \mathbb{N}$.
Let $p(n)$ be the proposition $2^n = 2$. Then $p(1)$ is true.
We wish to prove that $p(n) \implies p(n+1)$ assuming that $p(1), \ldots, p(n)$ are true:

$$\begin{aligned} 2^{n+1} &= 2^n \cdot 2 \\ &= 2^n \cdot \frac{2^n}{2^{n-1}} \\ &= 2 \cdot \frac{2}{2} \text{ (by } p(n) \text{ and } p(n-1)) \\ &= 2. \end{aligned}$$

This shows that $p(n) \implies p(n+1)$ and therefore that $2^n = 2$ for every $n \in \mathbb{N} \setminus \{0\}$. ♠

(1.87) EXERCISE.

Prove by induction that the sum of the first n odd numbers is given by the formula

$$1 + 3 + \cdots + (2n-1) = n^2,$$

i.e., for $n = 5$ we have

$$1 + 3 + 5 + 7 + 9 = 25.$$

♠

(1.88) EXERCISE.

Prove by induction that

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

i.e., for $n = 3$, we have

$$1^2 + 2^2 + 3^2 = 14 = \frac{3 \cdot 4 \cdot 7}{6}.$$

♠

(1.89) EXERCISE.

Prove by induction that

$$1^3 + 2^3 + 3^3 + \cdots + n^3 = \left(\frac{n(n+1)}{2}\right)^2$$

i.e., for $n = 3$, we have

$$1^3 + 2^3 + 3^3 = 36 = \left(\frac{3 \cdot 4}{2}\right)^2.$$

♠

(1.90) EXERCISE.

Prove using the idea of induction that

$$2^n < n!$$

for $n \geq 4$.

♠

The last exercise related to induction concerns the famous pigeonhole principle. The statement itself looks innocent, well almost ridiculous, but it is very powerful. Even the go-to website mathoverflow for research mathematicians has a quite nice thread about this.

(1.91) EXERCISE.

Prove the following by induction on m: if n items are put into m containers and $n > m$, then at least one container must contain more than one item. ♠

1.9 The concept of a function

A function is a crucial concept in mathematics. In Sage (actually python here) a simple function can be programmed like

Interactive code not included in static version.

The code above seems to take a number and returns the number plus one. This (f) is in fact a function taking as *input* a number and returning as *output* the number plus one. Notice that we do not even know which numbers we are talking about here. In mathematics we need to have a more precise notion of a function.

The above python function could more formally be denoted as $f : \mathbb{Z} \to \mathbb{Z}$ with $f(n) = n + 1$ if we are dealing with the integers, but we cannot tell from the code.

Well, to be fair ...: To be completely fair, it is possible from Python 3.5 to add type annotations to functions, so that we could write

Interactive code not included in static version.

in the Python code to state that the function should take values in the integers and return integers.

The precise mathematical definition of a function in terms of sets is the following. A function $f : S \to T$ is a subset $f \subseteq S \times T$, such that $(s,t_1) \in f \wedge (s,t_2) \in f \implies t_1 = t_2$. In words it states that a function $f : S \to T$ is a subset f of $S \times T$, containing pairs having only one second coordinate for every first coordinate.

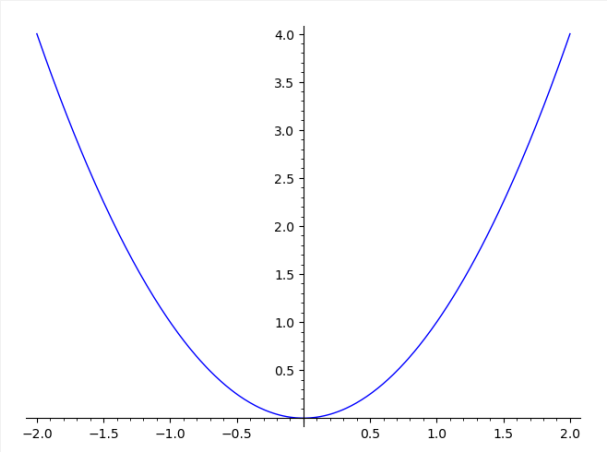
The everyday working definition of a function is more intuitive: a machine taking input from some set S and giving output in some set T. The uniqueness of the output is encoded in the mathematical definition of a function.

(1.92) DEFINITION.

Mathematically a function f takes values from a set S and returns values in a set T. In details, it is denoted $f : S \to T$ and the value associated with $s \in S$ is denoted $f(s) \in T$. Here S is called the domain *of f and T is called* the codomain *of f. Less, formally S is called the input set and T the output set for f.*

(1.93) REMARK.

Please notice that a function is a very, very general concept. It is not just something that you draw as a graph on a piece of paper. Of course, you can draw a function $f : \mathbb{R} \to \mathbb{R}$ like $f(x) = x^2$:



Generally, a function $f : S \to T$ is given by a machine, formula or algorithm that computes $f(x) \in T$ for every $x \in S$. Nothing more, nothing less. It really has nothing to do with a graph (even though graphs can sometimes be useful for visualizing certain functions like $f(x) = x^2$).

(1.94) EXAMPLE.

Good examples of functions can be found in the cryptographic hash functions. They are examples of complicated functions $f : S \to T$, where S is infinite and T finite. Here S could be data like plain text files and T could be a 256 bit number. This is the setup for the widely used `sha-256` cryptographic hash function. The whole point of a cryptographic hash function is that it must be humanly impossible to compute y with $f(y) = f(x)$ given $f(x)$[5]. In fact, `sha-256` is used in the Bitcoin block chain. The precise definition of `sha-256` can be found in FIPS PUB 180-4 approved by the Secretary of Commerce.

Other interesting functions output a bounded size digital footprint (checksum) of a file (like `md5`). This is very useful for checking data integrity of downloads over the internet. The `md5` hash is a 128 bit number.

Instead of listing 256 or 128 bits for the hash value one uses hexadecimal notation with digits in 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 , a, b, c, d, e, f. A pair of hexadecimal digits then represents a byte or 8 bits. Output from `sha-256` and `md5` consist of 64 and 32 hexadecimal digits respectively. You are welcome to experiment with these two hash functions in the Sage window below.

Interactive code not included in static version.

♠

(1.95) EXERCISE.

What is the `sha-256` hash of your name? Change a few letters and recompute. Do you see any system? What about the `md5` hash function? Can you find two different strings with the same `md5` hash using your computer?

[5] A pair $x \neq y$ with $f(x) = f(y)$ is called a collision

Hint: I have not answered the last question myself, but I am told that it is possible to find a collision for `md5` using a garden variety home computer. Browsing the internet, it seems that the two strings s_1 and s_2 given in hexadecimal notation[6] by

Interactive code not included in static version.

and

Interactive code not included in static version.

give a collision for `md5`. Verify that $s_1 \neq s_2$ and that they give the same `md5` hash. If you find a collision for `sha-256` you will become world famous.

Hint:

Interactive code not included in static version.

♠

1.9.1 When are two functions the same?

Suppose that $f(x) = x+1$. This way of defining a function is a bit sloppy. The domain and codomain is not defined. The correct way of defining a function also includes defining its domain and codomain as in Definition 1.92. Two functions f, g are the same when they have the same domain S and codomain T and $f(x) = g(x)$ for every $x \in S$.

(1.96) EXAMPLE.

The functions $f_1 : \mathbb{R} \to \mathbb{R}$ and $f_2 : \{x \in \mathbb{R} \mid x \geq 0\} \to \mathbb{R}$ given by $f_1(x) = x^2$ and $f_2(x) = x^2$ are not the same! Their domains are different. ♠

1.9.2 Notations for defining a function

If $f : S \to T$ is a function and S is a finite set, then you can define f using a simple table. This is best illustrated using an example. Suppose that $S = \{1,2,3\}, T = \mathbb{R}$ and

$$\begin{aligned} f(1) &= \sqrt{2} \\ f(2) &= \pi \\ f(3) &= -1. \end{aligned}$$

Then f is expressed in table form as

x	1	2	3
$f(x)$	$\sqrt{2}$	π	-1

[6] This notation represents a sequence of bytes given by pairs of hexadecimal digits

Very often the bracket (or *Tuborg* in Danish) notation is used. It is similar to `if-then-else` statements in programming:

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x^2 & \text{if } x > 0 \end{cases} \tag{1.19}$$

defines the function $f : \mathbb{R} \to \mathbb{R}$ that outputs 0 if the input $x \leq 0$ and x^2 if $x > 0$. In python we may express this as

```
Interactive code not included in static version.
```

(1.97) EXERCISE.

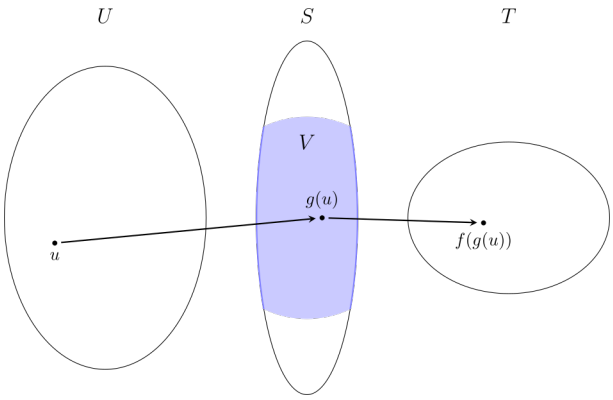
What is $f(-17)$ and $f(17)$ for the function defined in (1.19). Draw the graph of f. Come up with a function $f : S \to T$, where it does not make sense to draw a graph. ♠

1.9.3 Composition of functions

Given two functions $f : S \to T$ and $g : U \to V$, where $V \subseteq S$, we define a new function $f \circ g : U \to T$ by

$$(f \circ g)(u) = f(g(u)).$$

This notion calls for some reflection. We have a total of four sets in this definition: U, V, S and T and, not to forget, the condition that $V \subseteq S$. If this last condition was not satisfied it would be meaningless to apply the function f to $g(u)$. I hope the diagram below helps the understanding.



(1.98) REMARK.

The concept of a function is powerful and underlies functional programming in computer science: every computation can be realized as applying a composition of functions to an argument. This is exemplified in the computer language Haskell.

(1.99) EXERCISE.

Suppose that

$$U = \{1,2,3\}, \qquad S = \{1,2,3,4\} \qquad \text{and} \qquad T = \{7,8,9\}$$

and that $g : U \to S$ and $f : S \to T$ are given by the tables

x	1	2	3
$g(x)$	1	3	4

and

x	1	2	3	4
$f(x)$	7	8	9	7

Compute the table for $(f \circ g) : U \to T$. Show that $f \circ g$ is not injective. Adjust the table for f so that $f \circ g$ becomes bijective. ♠

(1.100) EXERCISE.

Consider $f : \mathbb{R} \to \mathbb{R}^2$ and $g : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(t) = (t^2, t^3)$$
$$g((x,y)) = \cos(xy) + x\sin(x+y).$$

What is $(g \circ f)(t)$ as a function from $\mathbb{R}$ to $\mathbb{R}$ in terms of t? ♠

1.9.4 Functions from and into products

(1.101) DEFINITION.

Suppose that

$$B = A_1 \times A_2 \times \cdots \times A_n$$

as in (1.13)*. Then the function* $\pi_i : B \to A_i$ *given by*

$$\pi_i(a_1, \ldots, a_i, \ldots, a_n) = a_i$$

is called the projection on the i-th coordinate.
If $f : A \to B$ *is a function, then*

$$f(a) = (f_1(a), \ldots, f_n(a)),$$

where $f_i = \pi_i \circ f$.

(1.102) EXAMPLE.

Suppose that $B = \mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ and $(x,y) \in B$. Then

$$\pi_1((1,2)) = 1$$
$$\pi_2((1,2)) = 2.$$

Now suppose that $A = \mathbb{R}^3$ and that $f : A \to B$ is given by

$$f((x,y,z)) = (x^2 + y, xy, z^3).$$

Then

$$f_1((1,2,3)) = 3$$
$$f_2((1,2,3)) = 2$$
$$f_3((1,2,3)) = 27.$$

♠

1.9.5 Injective and surjective functions

We now define three very important notions related to functions.

(1.103) DEFINITION.

Let $f : S \to T$ be a function. Then f is called

(i) injective, *if $f(x) = f(y) \implies x = y$ for every $x, y \in S$.*

(ii) surjective, *if for every $y \in T$, there exists $x \in S$, such that $f(x) = y$.*

(iii) bijective, *if it is both injective and surjective.*

(1.104) EXERCISE.

Is a cryptographic hash-function as defined in Example 1.94 injective? ♠

(1.105) EXERCISE.

Suppose that

$$S = \{1,2,3\} \qquad \text{and} \qquad T = \{1,2,3,4\}$$

and that the function $f : S \to T$ is defined by the table

x	1	2	3
$f(x)$	1	2	4

Is f injective? Is it surjective? Is it possible to adjust the table so that f becomes injective? Is it possible to adjust the table so that f becomes surjective? ♠

(1.106) EXERCISE.

Consider the function $f : S \to T$ given by

$$f(x) = x^2,$$

where $S = T = \mathbb{R}$. Is f injective? Is f surjective? Suggest how to change S and T so that $f : S \to T$ becomes bijective. ♠

(1.107) EXERCISE.

Consider the function $f : \mathbb{Z} \to \mathbb{Z}$ given by

$$f(x) = x + 1$$

Show that f is bijective. ♠

(1.108) EXERCISE.

Write down precisely how the truth table for $p \implies q$ may be expressed in terms of a function $f : S \to T$. What are the sets S and T in this case? ♠

1.9.6 The inverse function

If $f : S \to T$ is bijective, then we may define a function $g : T \to S$, so that $(f \circ g)(y) = y$ for every $y \in T$ and $(g \circ f)(x)$ for every $x \in S$. This function is denoted f^{-1}.

How do we define $f^{-1}(y)$ for $y \in T$? Well, since f is surjective, we may find $x \in S$ so that $y = f(x)$. Now, we simply define

$$f^{-1}(y) = x. \tag{1.20}$$

We cannot have $x_1 \neq x_2$ in S with $f(x_1) = f(x_2) = y$, since f is injective. We only have one choice for x in (1.20). Therefore (1.20) really is a good and sound definition.

(1.109) EXAMPLE.

Let $f : S \to S$, where $S = \{1,2,3\}$ be given by the table

x	1	2	3
$f(x)$	3	1	2

.

Then $f^{-1}$is given by the table

x	1	2	3
$f^{-1}(x)$	2	3	1

.

♠

(1.110) EXERCISE.

What if the definition of f in Example 1.109 is changed to

x	1	2	3
$f(x)$	3	2	2

.

Does f^{-1} make sense here? ♠

(1.111) EXERCISE.

What is the inverse function of $f : \mathbb{Z} \to \mathbb{Z}$ given by $f(x) = x + 1$? What is the inverse function of $g : S \to S$, where $g(x) = \sqrt{x}$ and $S = \{x \in \mathbb{R} \mid x \geq 0\}$? ♠

1.9.7 The preimage

(1.112) DEFINITION.

Consider a function

$$f : A \to B,$$

where A and B are sets. If $C \subseteq B$, then the preimage of C under f is defined by

$$f^{-1}(C) = \{x \in A \mid f(x) \in C\}.$$

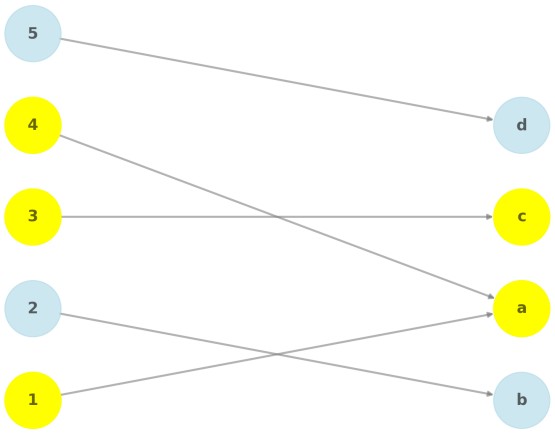
Definition 1.112 is short and sweet. Here is a first example of the preimage.

(1.113) EXAMPLE.

Consider the function $f : A \to B$, where $A = \{1,2,3,4,5\}$ and $B = \{a,b,c,d\}$ given by

x	1	2	3	4	5
$f(x)$	a	b	c	a	d

For $C = \{a,c\}$, $f^{-1}(C) = \{1,3,4\}$ as illustrated below.



♠

(1.114) EXERCISE.

What is $f^{-1}(C)$ when $A = \mathbb{R}, B = \mathbb{R}, f(x) = x^2 - 5x + 6$ and $C = (-\infty, 0]$? ♠

(1.115) QUIZ.

Quiz not included in static version. ♠

1.9.8 Neural networks

Having defined functions and composition of functions, we can deflate the term (deep) neural network, which is often clouded in magic and mystery.

A *neural network* is a special case of a function

$$f : A \to B, \tag{1.21}$$

where $A \subseteq \mathbb{R}^m$ and $B \subseteq \mathbb{R}^n$. Neural networks are often compositions of many intermediate functions called (hidden) layers.

(1.116) REMARK.

Recall from Definition 1.101 that a function such as (1.21) can be written

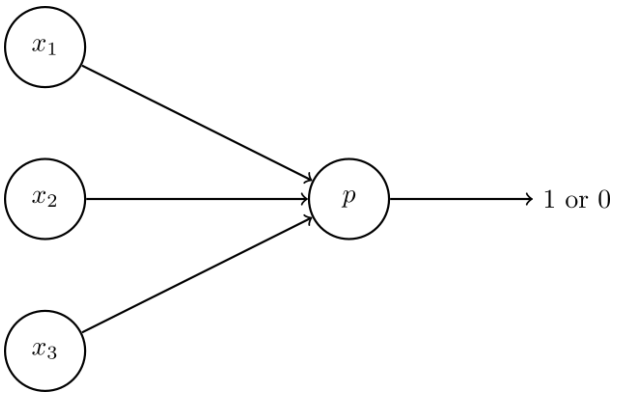
$$f(x_1,\dots,x_m) = (f_1(x_1,\dots,x_m),\dots,f_n(x_1,\dots,x_m)), \tag{1.22}$$

where $f_1,\dots,f_n$ are functions $A \to \mathbb{R}$. Check out Example 1.102.

In a neural network the functions $f_1, f_2, \dots, f_n$ are viewed as neurons[7]. Depending on their input they either fire or do not fire a signal. Classically this is modelled by the perceptron, which is a function $p : \mathbb{R}^n \to \mathbb{R}$ of the form

$$p(x_1,\dots,x_n) = \begin{cases} 1 & \text{if } w_1x_1 + \cdots + w_nx_n > b \\ 0 & \text{if } w_1x_1 + \cdots + w_nx_n \leq b \end{cases} \tag{1.23}$$

for fixed numbers $w_1,\dots,w_n$ (called weights) and a number b (called the threshold). If the weighted sum $w_1x_1 + \cdots + w_nx_n$ is above the threshold, the neuron fires (returns the value 1). If not it does not fire (returns the value 0).



(1.117) EXERCISE.

Consider the three perceptrons $p_1, p_2, p_3 : \mathbb{R}^2 \to \mathbb{R}$, where

$$p_1(x,y) = \begin{cases} 1 & \text{if } -x-y > -\frac{3}{2} \\ 0 & \text{if } -x-y \leq -\frac{3}{2} \end{cases}, \qquad p_2(x,y) = \begin{cases} 1 & \text{if } x+y > \frac{1}{2} \\ 0 & \text{if } x+y \leq \frac{1}{2} \end{cases},$$

and

$$p_3(x,y) = \begin{cases} 1 & \text{if } x+y > \frac{3}{2} \\ 0 & \text{if } x+y \leq \frac{3}{2} \end{cases}.$$

Let $f(x,y) = p_3(p_1(x,y), p_2(x,y))$. Then f is a composite function $f = g \circ h$ of two functions $h : \mathbb{R}^2 \to \mathbb{R}^2$ and $g : \mathbb{R}^2 \to \mathbb{R}$. Write down these functions.

Hint: Have a closer look at (1.22) in order to understand how functions from $\mathbb{R}^2$ to $\mathbb{R}^2$ are expressed. Notice that our notation is a bit inconsistent when it comes to types. For example, the function $p_1 : \mathbb{R}^2 \to \mathbb{R}$ should really be denoted $p_1((x,y))$ instead of $p_1(x,y)$, since it takes input from $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$. This is remedied in the (hopefully easy to understand) python code below.

[7] To be precise, the functions should be viewed as synapses

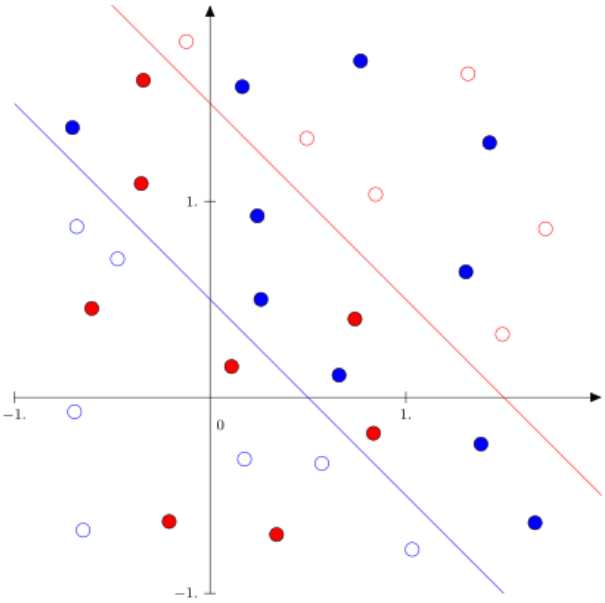
Interactive code not included in static version.

LLM

Explain the python code below to me.
def p1(v): (x, y) = v if -x-y > -3/2: return 1 else: return 0
def p2(v): (x, y) = v if x + y > 1/2: return 1 else: return 0
def p3(v): (x, y) = v if x+y > 3/2: return 1 else: return 0
def h(v): return (p1(v), p2(v))
def g(v): return p3(v)
def f(v): return g(h(v))

Compute $f(0,0), f(1,0), f(0,1)$ and $f(1,1)$.

Relate the perceptrons p_1 and p_2 to the illustration below. What do you think the red and blue line illustrate? What does it mean that a dot is solid compared to hollow? What is special about points between the red and blue lines? Try to relate $f(0,0), f(1,0), f(0,1)$ and $f(1,1)$ to the illustration.



(Illustration courtesy of William Heyman Krill).

♠

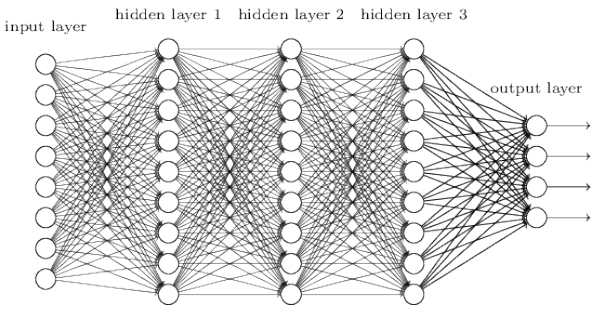
(1.118) EXERCISE.

Give weights w_1, w_2 and a threshold b for a perceptron $p : \mathbb{R}^2 \to \mathbb{R}$ that computes the logical and function $\wedge$ i.e, p must satisfy

$$p(0,0) = 0$$
$$p(1,0) = 0$$
$$p(0,1) = 0$$
$$p(1,1) = 1.$$

Do the same for the logical or function $\vee$. ♠

The output of one neuron can be used as input for other neurons in a potentially extremely complicated network:



The diagram above represents a neural network, which is a function $\mathbb{R}^8 \to \mathbb{R}^4$. This function is actually a composition (represented by the hidden layers 1, 2, 3 and the output layer):

$$\mathbb{R}^8 \to \mathbb{R}^9 \to \mathbb{R}^9 \to \mathbb{R}^9 \to \mathbb{R}^4.$$

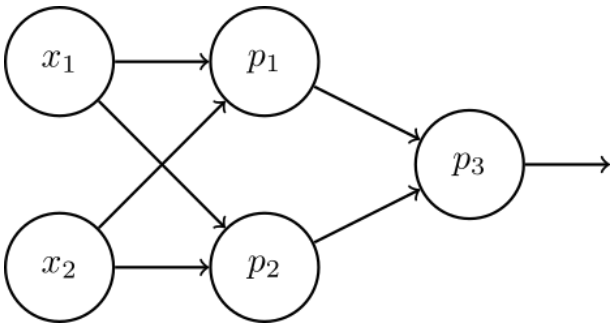
All of the nodes above, except the ones in the input layer, represent perceptrons.

(1.119) EXERCISE.

Is it possible to find a perceptron $p : \mathbb{R}^2 \to \mathbb{R}$, such that

$$\begin{aligned} p(0,0) &= 0 \\ p(1,0) &= 1 \\ p(0,1) &= 1 \\ p(1,1) &= 0? \end{aligned}$$

What if you are allowed to use a neural network composed as $\mathbb{R}^2 \to \mathbb{R}^2 \to \mathbb{R}$ (one hidden layer)



? ♠

Mathematically there is no reason to use special functions such as perceptrons in each node. One also uses a (smooth) version of the perceptron employing the sigmoid function. With the notation above, this function is given as

$$\sigma(x_1,\ldots,x_n) = \frac{1}{1+e^{-(w_1x_1+\cdots+w_nx_n)-b}}.$$

However, around 2011 it was observed that the perceptron activation function (ReLU) as defined in (1.23) led to better training of deep neural networks.

Chapter 2

Linear equations

Modern mathematical terminology may seem abstract, but a lot of it comes from equation solving. We will talk about linear equations in this chapter to motivate the concept of matrices in the next chapter.

Linear equations are equations, where the unknowns only appear to the first power. For example, $x^2+x+1=0$ is not a linear equation in the unknown x, since x to the second power (x^2) appears in the equation, whereas $2x-3=1$ is. We may also consider several linear equations with several unknowns, such as

$$\begin{aligned} x+y+z &= 3 \\ x-y+z &= 1 \\ x+y-z &= 1 \end{aligned} \tag{2.1}$$

consisting of three linear equations with the three unknowns x, y and z. To be completeply precise, a solution to (2.1) is a triple $(x,y,z) \in \mathbb{R}^3$, such that the predicate

$$(x+y+z=3) \wedge (x-y+z=1) \wedge (x+y-z=1)$$

is true.

(2.1) EXERCISE.

Try to come up with a solution to (2.1) i.e., find numbers x,y,z satisfying all three equations. Do not use a computer. Is there more than one solution?

Write down two linear equations with two unknowns, which do not have a solution. ♠

Do the exercise above, before you evaluate the Sage code below, which uses the dreaded solve function. The solve function should always be used as a last resort.

Interactive code not included in static version.

2.1 One linear equation with one unknown

Very simple rules apply when solving linear equations.

Consider as an example the linear equation $2x-3=1$ in the unknown x. Solving this equation amounts to

reducing to an expression $x =$ a number. This is called *isolating* x. The process is very mechanical:

$$\begin{gathered} 2x-3=1 \\ \Updownarrow \\ 2x-3+3=1+3 \\ \Updownarrow \\ 2x=4 \\ \Updownarrow \\ \left(\frac{1}{2}\right)2x=\left(\frac{1}{2}\right)4 \\ \Updownarrow \\ x=2 \end{gathered}$$

If you look closely, you will see that we have used the rules

$$\begin{aligned} a=b \quad &\Longleftrightarrow \quad a+c=b+c \\ a=b \quad &\Longleftrightarrow \quad ta=tb, \end{aligned}$$

where a,b,c are numbers and t is a number $\neq 0$.

(2.2) EXERCISE.

Point out the mistake(s) in the argument[1] below showing that $2=1$.

$$\begin{aligned} a=b &\iff \\ a^2=ab &\iff \\ a^2-b^2=ab-b^2 &\iff \\ (a+b)(a-b)=b(a-b) &\iff \\ a+b=b &\iff \\ 2b=b &\iff \\ 2=1. & \end{aligned}$$

♠

(2.3) QUIZ.

Quiz not included in static version. ♠

(2.4) EXERCISE.

Diophantus's youth lasted $1/6$ of his life. He grew a beard after $1/12$ more. After $1/7$ more he got married. Five years later he had a son. The son lived half as long as the father and Diophantus died four years after the son. At what age did Diophantus die?

Link/Hint: You can read about Diophantus and the solution to the puzzle in the Wikipedia entry about him. Please try solving the problem on your own first. ♠

[1] This teaser was presented at the workshop for new teaching assistants, August 2020.

2.2 Several linear equations with several unknowns

The linear equation $2x-3=1$ has only one unknown with the unique solution $x=2$. *If one linear equation has more than one unknown, then it has infinitely many solutions*. Consider as an example the linear equation $2x-3y=1$ with the unknowns x and y. Using the procedure as before, we get

$$\begin{aligned} 2x-3y &= 1 \\ &\Updownarrow \\ x &= \frac{1}{2}+\frac{3}{2}y \end{aligned}$$

Here we are free to choose y in infinitely many ways giving infinitely many solutions $(x,y)\in\mathbb{R}^2$.

2.2.1 Several equations

Several equations with several unknowns also make sense. Consider

$$\begin{aligned} x+y &= 3 \\ 2x-3y &= 1 \end{aligned} \tag{2.2}$$

Two numbers x and y form a solution $(x,y)\in\mathbb{R}^2$ if both equations are satisfied. From the example above, we know that

$$x=\frac{1}{2}+\frac{3}{2}y. \tag{2.3}$$

This can be inserted for x in the first equation and we get

$$3=x+y=\frac{1}{2}+\frac{3}{2}y+y=\frac{1}{2}+\frac{5}{2}y.$$

Here we end up with one linear equation in one variable y. The solution is $y=1$, which is inserted in the equation (2.3) giving $x=2$. Therefore the solution to the equations is $(x,y)=(2,1)$.

(2.5) REMARK.

To be completely precise about these steps, let us use predicate logic. A solution to the system of equations above is a pair of numbers $(x,y)\in\mathbb{R}^2$ satisfying the predicate

$$(x+y=3)\wedge(2x-3y=1).$$

Now use the rules in Proposition 1.13 and substitution to rewrite systematically:

$$\begin{aligned} &(x+y=3)\wedge(2x-3y=1) \iff \\ &(x=3-y)\wedge(2x-3y=1) \iff \\ &(x=3-y)\wedge(2(3-y)-3y=1) \iff \\ &(x=3-y)\wedge(6-5y=1) \iff \\ &(x=3-y)\wedge(y=1) \iff \\ &(x=2)\wedge(y=1). \end{aligned}$$

Tracing back we have actually proved that

$$(x+y=3)\wedge(2x-3y=1) \iff (x=2)\wedge(y=1).$$

(2.6) EXERCISE.

Why is $x = 2 \wedge y = 1$ the only solution to the equations in (2.2)? How can we be so sure that there are no other values for x and y satisfying (2.2)? ♠

(2.7) QUIZ.

Quiz not included in static version. ♠

2.3 Gauss elimination

When solving systems of several linear equations, it is natural to fix one of the equations, isolate an unknown and then insert in the other equations.

Let us study this procedure focusing on an example with two equations and three unknowns:

$$\begin{aligned} x+2y+z &= 8 \\ 2x+y+z &= 7 \end{aligned}$$

In the first equation we isolate $x = 8-2y-z$, which is then inserted into the second equation:

$$2x+y+z = 2(8-2y-z)+y+z = -3y-z+16 = 7 \implies -3y-z = -9.$$

It makes perfect sense to multiply the first equations by 2 and subtract from the second equations. This operation gives

$$-3y-z = -9.$$

It is not a coincidence that these two operations give the same result.

(2.8) THEOREM.

Suppose that

$$\begin{aligned} a_1x_1+a_2x_2+\cdots+a_nx_n &= c_1 \\ b_1x_1+b_2x_2+\cdots+b_nx_n &= c_2 \end{aligned}$$

are two linear equations in the unknowns $x_1,\ldots,x_n$ with $a_1 \neq 0$. The equation gotten by first isolating x_1 in the first equation and then inserting in the second equation is identical to the equation you get by adding the first equation multiplied by $-b_1/a_1$ to the second equation.

Proof. Isolating x_1 in the first equation inserted in the second equation gives the equation

$$b_1\left(\frac{c_1}{a_1}-\frac{a_2}{a_1}x_2-\cdots-\frac{a_n}{a_1}x_n\right)+b_2x_2+\cdots+b_nx_n = c_2 \tag{2.4}$$

Adding $-b_1/a_1$ multiplied to the first equation to the second equation gives

$$\left(b_2-\frac{b_1a_2}{a_1}\right)x_2+\cdots+\left(b_n-\frac{b_1a_n}{a_1}\right)x_n = c_2-\frac{b_1}{a_1}c_1 \tag{2.5}$$

Using basic arithmetic you can see that (2.4) can be rewritten to (2.5). □

Multiplying an equation by a number and then adding to another equation is easier to handle than the method of isolating and inserting. We have showed above that they produce the same result. Below is an extended example.

(2.9) EXAMPLE.

We wish to solve the system of equations

$$\begin{array}{rcrcrcl} 2x & + & y & + & z & = 7 & \\ x & + & 2y & + & z & = 8. & \\ x & + & y & + & 2z & = 9 & \end{array} \tag{2.6}$$

The first step is subtracting the third equation from the second:

$$\begin{array}{rcrcrcl} 2x & + & y & + & z & = 7 & \\ x & + & 2y & + & z & = 8 & \\ x & + & y & + & 2z & = 9 & \end{array} \quad \Longleftrightarrow \quad \begin{array}{rcrcrcl} 2x & + & y & + & z & = 7 & \\ & & y & - & z & = -1 & \\ x & + & y & + & 2z & = 9 & \end{array}$$

Then we multiply the third equation by 2 and subtract from the first:

$$\begin{array}{rcrcrcl} 2x & + & y & + & z & = 7 & \\ & & y & - & z & = -1 & \\ x & + & y & + & 2z & = 9 & \end{array} \quad \Longleftrightarrow \quad \begin{array}{rcrcrcl} & - & y & - & 3z & = -11 & \\ & & y & - & z & = -1 & \\ x & + & y & + & 2z & = 9 & \end{array}$$

Finally we add the second equation to the first:

$$\begin{array}{rcrcrcl} & - & y & - & 3z & = -11 & \\ & & y & - & z & = -1 & \\ x & + & y & + & 2z & = 9 & \end{array} \quad \Longleftrightarrow \quad \begin{array}{rcrcrcl} & & & - & 4z & = -12 & \\ & & y & - & z & = -1 & \\ x & + & y & + & 2z & = 9 & \end{array}$$

We have now reduced the original system of equations (2.6) to

$$\begin{array}{rcrcrcl} & & & - & 4z & = -12 & \\ & & y & - & z & = -1\,, & \\ x & + & y & + & 2z & = 9 & \end{array}$$

where the first equation shows that $z = 3$. Now $z = 3$ can be inserted into the second equation, giving $y - 3 = -1$, which is solved by $y = 2$. Finally $y = 2$ and $z = 3$ are inserted into the third equations giving the equation $x + 8 = 9$, which is solved by $x = 1$.

One very important observation here is that $x = 1, y = 2$ and $z = 3$ is the only solution to (2.6). This is a logical consequence of the bi-implication arrows $\Longleftrightarrow$ throughout the above calculations.

Interactive code not included in static version.

♠

The elimination or substitution method for solving systems of linear equations is old and well known. Sir Isaac Newton described in 1720 the methods eloquently as follows.

> *And you are to know, that by each Æquation one unknown Quantity may be taken away, and consequently, when there are as many Æquations and unknown Quantities, all at length may be reduc'd into one, in which there shall be only one Quantity unknown.*

The mathematical rockstar Carl Friedrich Gauss used the method to determine the orbit for the asteroid Pallas. The mathematical analysis of the observations lead him to the famous least squares method and a system of six linear equations with six unknowns.

The method is known today by the term *Gaussian elimination* even though Gauss was not the first to introduce it. In fact it appeared already in *The Nine Chapters on the Mathematical Art*, which is an ancient Chinese mathematics book compiled over several centuries from the 10th century BCE to the 2nd century CE. This book contains several practical problems and their solutions. An example is

> There are three categories of corn. Three bundles of the first class, two of the second and one of the third make 39 measures. Two of the first, three of the second, and one of the third make 34 measures. Finally one of the first, two of the second and three of the third make 26 measures. How many measures of graín are contained in one bundle of each class?

(2.10) QUIZ.

Quiz not included in static version. ♠

(2.11) QUIZ.

Quiz not included in static version. ♠

(2.12) EXERCISE.

Find the solutions to

$$\begin{array}{rcrcrcr} x & + & 3y & + & z & = & 2 \\ -2x & - & 5y & + & 3z & = & 4. \end{array}$$

by expressing x and y in terms of z i.e., isolate x on the left hand side, such that

$$x = \dots$$
$$y = \dots,$$

where $\dots$ indicate an expression only in the unknown z. ♠

(2.13) EXERCISE.

Your enemy transmits secret codes (x_1, x_2, x_3, x_4) consisting of four integers x_1, x_2, x_3, x_4 over the internet. He does not transmit the code itself but an encrypted version (y_1, y_2, y_3, y_4) given by

$$\begin{array}{rcrcrcrcr} y_1 & = & 2x_1 & + & x_2 & + & 3x_3 & + & 4x_4 \\ y_2 & = & x_1 & + & 2x_2 & + & 3x_3 & + & 4x_4 \\ y_3 & = & 3x_1 & + & 3x_2 & + & x_3 & + & x_4 \\ y_4 & = & 4x_1 & + & 4x_2 & + & 2x_3 & + & 3x_4 \end{array}.$$

You have knowledge of the encryption method above and by listening in on a recent communication, you learn that the encryption $(15, 16, 12, 20)$ was sent. What was the original secret code before the encryption?

Extra credit: Suppose that you only know that the encryption scheme is

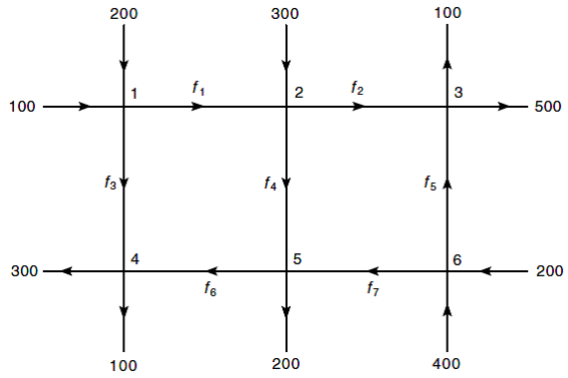
$$\begin{array}{rcrcrcrcr} y_1 & = & a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & a_{14}x_4 \\ y_2 & = & a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & + & a_{24}x_4 \\ y_3 & = & a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & + & a_{34}x_4 \\ y_4 & = & a_{41}x_1 & + & a_{42}x_2 & + & a_{43}x_3 & + & a_{44}x_4 \end{array},$$

and that you have no knowledge of the numbers $a_{11}, \ldots, a_{44}$. How many transmissions do you need to know at the minimum to find these encryption numbers?

♠

(2.14) EXERCISE.

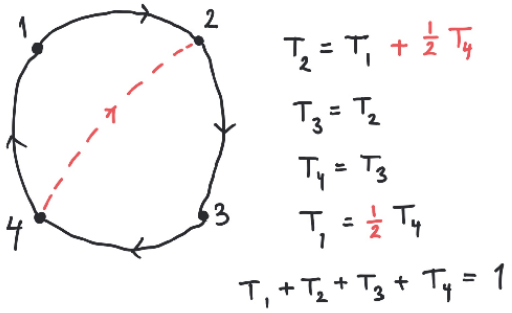
The diagram below shows a network of roads and 6 intersections. Every road is labeled by a number indicating the average number of cars per hour on the road. Some of these numbers $f_1, \ldots, f_7$ are unknowns. Write up a system of linear equations for finding $f_1, \ldots, f_7$.



Compute $f_1, f_2, f_3, f_4, f_5, f_6$ supposing that $f_1 = 200$ and $f_7 = 100$. ♠

(2.15) EXAMPLE.

This example relates to the famous Google page rank algorithm.



Suppose we have a very simple internet with only four webpages as depicted above with arrows indicating that a webpage links to another.

We wish to study traffic in this network in the sense that we let a random websurfer jump from a given webpage to another by selecting a link randomly.

If you look at the network without the punctured red arrow, it is almost clear the a random websurfer will spend 25 % of the time uniformly in each of the four nodes.

However, if we introduce the puntured red arrow, then the percentages in each node are given by the linear equations above. Here it turns out that website 1 only gets around 14 % of the time (the other websites get double this time each).

♠

(2.16) EXERCISE.

Which webpage in the above diagram loses most traffic when a link is added from 3 to 1? ♠

(2.17) EXAMPLE.

You may try out the python code below to simulate a random tour of the small internet in Example 2.15.

Interactive code not included in static version.

The list (or matrix)

Interactive code not included in static version.

encodes the graph of links between the four nodes $0,1,2,3$. From A you can see that 2 links to 0 and 1 and that 0 links to 1. The command

Interactive code not included in static version.

simulates a random surf with $1,000$ clicks starting in node 0.

The linear equations really seem to give the right result! ♠

2.4 Polynomials

Before going further into examples of linear equations we need to introduce (non-linear) functions called *polynomials*. A polynomial of degree n is a function $f : \mathbb{R} \to \mathbb{R}$ of the form

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0, \tag{2.7}$$

where $a_0, \ldots, a_n$ are real numbers and $a_n \neq 0$. We call $a_0, \ldots, a_n$ the *coefficients* of f. The degree of the polynomial f is denoted $\deg(f)$. As an example,

$$x^3 - 2x + 17$$

is a polynomial of degree 3 with

$$\begin{aligned} a_3 &= 1 \\ a_2 &= 0 \\ a_1 &= -2 \\ a_0 &= 17. \end{aligned}$$

In addition to the polynomials defined in (2.7) with $a_n \neq 0$, we also view the function $f(x) = 0$ as a polynomial, called the *zero polynomial*. The zero polynomial does not[2] have a degree.

[2] All its coefficients are zero!

The set of all polynomials is denoted $\mathbb{R}[x]$, so that for example it makes sense to write

$$x^2 - 5x + 6 \in \mathbb{R}[x].$$

It is probably the most natural functions from $\mathbb{R}$ to $\mathbb{R}$ you can come up with. If you look at (2.7), you will see that the output is formed by using addition and multiplication (by x and selected real numbers).

You can compute with polynomials treating the *variable* x as a number obeying the rules in Proposition 1.13. For example,

$$(3x^2 + 2x + 1)(2x + 1) = 6x^3 + 7x^2 + 4x + 1.$$

In that sense polynomials obey the same arithmetic rules as numbers. A fundamental difference is the x is a placeholder or a smbol where you can insert a number from $\mathbb{R}$. In general a polynomial of degree m times a polynomial of degree n is a polynomial of degree $m+n$.

In the sage window below we encounter for the first time the `sympy` library. The input format and commands for handling polynomials should be clear from the context.

Interactive code not included in static version.

You have already seen polynomials of degree one. They have the form

$$f(x) = ax + b,$$

where a and b are real numbers and $a \neq 0$. Similarly polynomials of degree two are called quadratic polynomials. They look like

$$f(x) = ax^2 + bx + c,$$

where a, b and c are real numbers and $a \neq 0$.

To get a feeling for the behavior of polynomials you should experiment in the sage window below. Try varying the degree and the coefficients of the polynomial in the plot. Also adjust the plot interval for the right view.

Interactive code not included in static version.

(2.18) EXERCISE.

Suppose that

$$f(x) = ax^2 + bx + c.$$

To compute $f(x)$ it seems that you need 3 multiplications ($a \cdot x \cdot x$ and $b \cdot x$) and 2 additions. Can you compute $f(x)$ with only 2 multiplications and 2 additions?

Try to generalize to the computation of $f(x)$, where f is a polynomial

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

of degree n (you should only need n multiplications and n additions here). ♠

2.4.1 Polynomial division

Division is sometimes referred to as long division when focusing on the method for division. Let us look at the situation for integers first.

The remainder of 14 divided by 4 is 2, since

$$14 = 3 \cdot 4 + 2.$$

Here the remainder 2 is strictly less than the divisor 4.

For polynomials we have a similar situation, where the degree is taken into account. For example, the remainder of x^3+x+1 divided by x^2+x+1 is $x+2$, since

$$x^3+x+1=(x-1)(x^2+x+1)+(x+2). \tag{2.8}$$

Here the degree of the remainder 1 is strictly less than the degree of the divisor 2.

The Python library `sympy` contains a wealth of functions for symbolic mathematics. In the window below, it is shown how the polynomial division (2.8) is computed using the Polynomial Manipulaton section of the `sympy` documentation.

Interactive code not included in static version.

The (division) algorithm for carrying out (long) division of polynomials is explained by an example in the video below.

(2.19) VIDEO.

Link to video

(2.20) EXERCISE.

Watch the five minute video above and carry out (do not use a computer) the polynomial division alluded to in (2.8).

Also interact with a chatbot of your choice below.

LLM

Please explain the division algorithm for polynomials to me. I want you to do this by an example.

♠

The general result about division of polynomials is given below.

(2.21) THEOREM.

Let $d(x)\in\mathbb{R}[x]$ be a non-zero polynomial. Then for every polynomial $f(x)\in\mathbb{R}[x]$, there exists polynomials $q(x),r(x)\in\mathbb{R}[x]$, such that

$$f(x)=q(x)d(x)+r(x), \tag{2.9}$$

where $r(x)=0$ or $\deg(r(x))<\deg(d(x))$.

Proof. We will prove this using induction on $n=\deg(f)$. Suppose that

$$f(x)=a_nx^n+\cdots \qquad \text{and} \qquad d(x)=b_mx^m+\cdots$$

In general if $\deg(d(x))=m>n$, then

$$f(x)=0\cdot d(x)+f(x)$$

satisfies the assumptions for the identity in (2.9) with $q(x) = 0$ and $r(x) = f(x)$.

If $m \leq n$, then $f(x) - a_n b_m^{-1} x^{n-m} d(x)$ is a polynomial of degree $< n$. So by induction we may find polynomials $q_0(x)$ and $r_0(x)$, such that

$$f(x) - a_n b_m^{-1} x^{n-m} d(x) = q_0(x)d(x) + r_0(x).$$

Therefore

$$f(x) = (q_0(x) + a_n b_m^{-1} x^{n-m})d(x) + r_0(x)$$

giving the desired result with $q(x) = q_0(x) + a_n b_m^{-1} x^{n-m}$ and $r(x) = r_0(x)$. □

2.4.2 Roots of polynomials

A real number $\alpha \in \mathbb{R}$ is called a *root* of the polynomial $f(x) \in \mathbb{R}[x]$ if $f(\alpha) = 0$. This is a very fundamental definition. It is mirrored beautifully in the following result.

(2.22) PROPOSITION.

A real number α is a root of the polynomial $f(x) \in \mathbb{R}[x]$ if and only if

$$f(x) = q(x)(x - \alpha),$$

for some polynomial $q(x) \in \mathbb{R}[x]$.

Proof. By Theorem 2.21, we may write

$$f(x) = q(x)(x - \alpha) + r(x), \tag{2.10}$$

where $r(x) = 0$ or $r(x)$ is a non-zero polynomial of degree zero i.e., a non-zero constant. Now the result follows, since $f(\alpha) = q(\alpha)(\alpha - \alpha) + r(\alpha) = r(\alpha)$ using (2.10). □

(2.23) EXERCISE.

Is there an easy way of deciding if a polynomial $d(x) = ax + b$ of degree one divides a polynomial $f(x)$ without performing the (long) division of $f(x)$ by $d(x)$. Here divides means that $f(x) = q(x)d(x)$ for some polynomial $q(x)$. ♠

A quadratic polynomial

$$ax^2 + bx + c$$

has at most two roots given by the formula (one root for $+$ and one for $-$ in $\pm$ below)

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \tag{2.11}$$

if its *discriminant* $b^2 - 4ac$ is ≥ 0.

Deriving the formula (2.11) comes from a classical algebraic trick called *completing the square*. Looking at the quadratic equation $ax^2+bx+c=0$, what bothers us is the term bx. If $b=0$ we could solve the equation rewriting to

$$x^2=-\frac{c}{a}$$

and then taking square roots. The first step in this direction is rewriting the equation

$$ax^2+bx+c=0$$

to

$$x^2+\frac{b}{a}x=-\frac{c}{a}. \tag{2.12}$$

We would like to add a number d^2 to both sides of (2.12) so that the left hand side comes to look like

$$(x+d)^2=x^2+2xd+d^2. \tag{2.13}$$

This is what is called completing the square.
Comparing the left hand side of (2.12) with the right hand side of (2.13), we find that

$$d=\frac{b}{2a}$$

works. Therefore (2.12) implies

$$\left(x+\frac{b}{2a}\right)^2=-\frac{c}{a}+\left(\frac{b}{2a}\right)^2.$$

This identity can be rewritten into the formula (2.11) for solving the quadratic equation.

For polynomials of degree three (cubic polynomials) there is a formula, but these days nobody remembers it. Also for polynomials of degree four (quartic polynomials) there is a formula. But for polynomials of degree five (quintic polynomials) and up, one can prove that a formula cannot exist!

An exceedingly important result is quoted and proved below: the degree of a polynomial is an upper bound for its number of roots.

(2.24) THEOREM.

A non-zero polynomial $f(x)\in\mathbb{R}[x]$ of degree $n>0$ can have at most n roots.

Proof. We will prove this by induction starting with $n=1$. Here $f(x)=ax+b$ for $a,b\in\mathbb{R}[x]$ and

$$f(\alpha)=0 \iff \alpha=-a^{-1}b.$$

Therefore $f(x)$ has precisely one root. Suppose now that we have proved that polynomials of degree n has at most n roots. Assume that $f(x)$ is a polynomial of degree $n+1$. If $f(x)$ has no roots, we are done with the proof. Suppose that $f(\alpha)=0$ i.e., α is a root in f. Then

$$f(x)=q(x)(x-\alpha)$$

by Proposition 2.22. Here $q(x)$ has to be a polynomial of degree n and therefore by induction, $q(x)$ has at most n roots. However, if $f(\beta)=q(\beta)(\beta-\alpha)=0$, then either $\beta=\alpha$ or $q(\beta)=0$. We have proved that $f(x)$ cannot have more than $n+1$ roots. □

(2.25) REMARK.

Theorem 2.24 has a few interesting consequences. First it implies that two identical polynomials i.e., $f(x) = g(x)$ for every $x \in \mathbb{R}$ must have the same coefficients.

Secondly if two polynomials $f(x)$ and $g(x)$ of degree n satisfy $f(x_i) = g(x_i)$ for distinct points $x_1, \ldots, x_{n+1}$, then $f(x) = g(x)$.

(2.26) EXERCISE.

In Remark 2.25 it is stated that if two polynomials $f(x)$ and $g(x)$ of degree n satisfy $f(x_i) = g(x_i)$ for distinct points $x_1, \ldots, x_{n+1}$, then $f(x) = g(x)$. How does this follow from Theorem 2.24? ♠

It might happen that a polynomial of degree n has precisely n roots, but it could have less or even no roots: the polynomials

$$x^2 + 1, x^4 + 1, x^6 + 1, \ldots$$

have no roots, whereas for example

$$x^2 - 2x + 1$$

is a quadratic polynomial with only one root. However polynomials of degree $1, 3, 5, \ldots$ always have at least one root.

(2.27) THEOREM.

A polynomial of odd degree always has a root.

The proof of this result is beyond our scope now and will have to wait for tools from analysis (Chapter 5).

(2.28) EXERCISE.

Compute (without using a computer!) the roots of the quartic

$$x^4 - 5x^2 + 6.$$

♠

(2.29) EXERCISE.

Give an example of a polynomial of degree 17 with precisely one root. ♠

(2.30) EXERCISE.

Suppose that α, β are two roots of the quadratic polynomial

$$f(x) = x^2 + bx + c.$$

How can b and c be computed in terms of α and β? Show concretely how this can be applied to the polynomial $g(x) = x^2 - 5x + 6$: if you know that $g(2) = 0$ how can you easily find the other root?

Hint: Show that $f(x) = (x - \alpha)(x - \beta)$ and use this. ♠

2.5 Applications of linear equations to polynomials

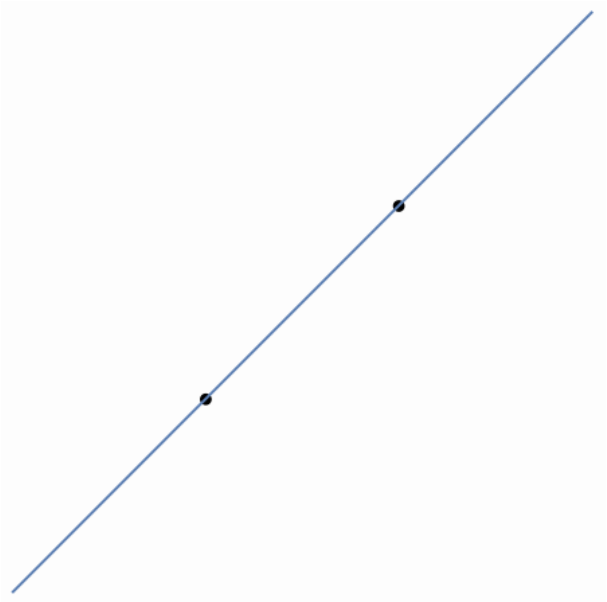
A line in the plane is given by its equation $y = ax + b$, where a is the slope and b is the intersection with the y-axis. Two lines in the plane are either parallel or intersect in a single point.

(2.31) EXERCISE.

The two lines $y = x + 1$ and $y = -x + 2$ have a single point of intersection. Compute this point.

Give an example of two parallel lines and their equations. ♠

Through two (distinct) points (x_1, y_1) and (x_2, y_2) with $x_1 \neq x_2$ passes a unique line



You can find the equation for this line by solving two equations with two unknowns a and b:

$$x_1 a + b = y_1$$
$$x_2 a + b = y_2$$

We might as well apply Gauss elimination to solve this system. First we subtract the second equation from the first. This gives $(x_1 - x_2)a = y_1 - y_2$. Therefore

$$a = \frac{y_1 - y_2}{x_1 - x_2}.$$

Inserting this a in the first equation we get

$$b = \frac{x_1 y_2 - x_2 y_1}{x_1 - x_2}.$$

We can also in a quite explicit way just write

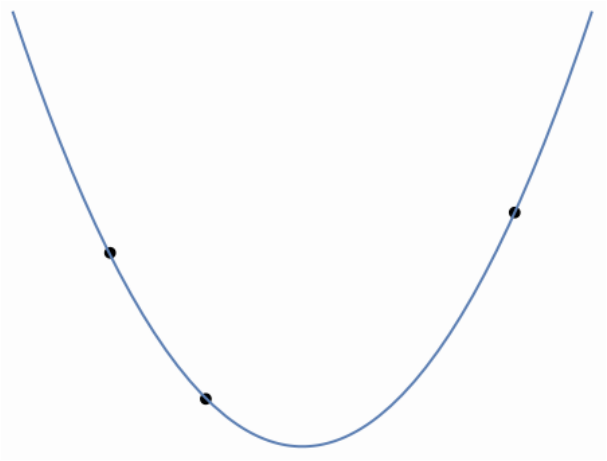
$$y = f(x) = y_1 \frac{x - x_2}{x_1 - x_2} + y_2 \frac{x - x_1}{x_2 - x_1}. \tag{2.14}$$

The function $f(x)$ in (2.14) is a polynomial of degree one with $f(x_1) = y_1$ and $f(x_2) = y_2$.

In almost the same way we may find a unique quadratic polynomial

$$y = ax^2 + bx + c$$

through three points $(x_1, y_1), (x_2, y_2)$ and (x_3, y_3) with distinct x-values:



Here we end up with three linear equations in the unknowns a, b and c:

$$\begin{aligned} x_1^2 a + x_1 b + c &= y_1 \\ x_2^2 a + x_2 b + c &= y_2 \\ x_3^2 a + x_3 b + c &= y_3 \end{aligned} \tag{2.15}$$

It is not immediately obvious that this system of equations has a solution. But watch the following trick evolve.

We may explicitly construct the quadratic polynomial passing through the three points as

$$\begin{aligned} y = f(x) = & y_1 \frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)} + y_2 \frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)} \\ & + y_3 \frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)} \end{aligned} \tag{2.16}$$

Take a moment and verify that $f(x_1) = y_1, f(x_2) = y_2$ and $f(x_3) = y_3$. This also proves that the system of equations in (2.15) can be solved.

(2.32) REMARK.

Notice in (2.16) that

$$y = y_1 L_1(x) + y_2 L_2(x) + y_3 L_3(x),$$

where (for example) L_1 is a polynomial of degree two satisfying

$$L_1(x_1) = 1, \quad L_1(x_2) = 0, \quad \text{and} \quad L_1(x_3) = 0.$$

What about L_2 and L_3 with respect to x_1, x_2 and x_3?

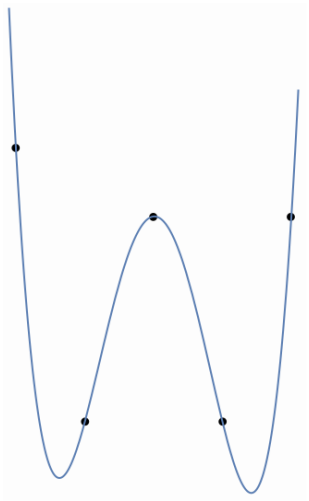
(2.33) EXERCISE.

Compute the polynomial you get when you apply (2.16) to $x_1 = 1, x_2 = 2, x_3 = 3$ and $y_1 = 1, y_2 = 2, y_3 = 3$. How do you explain this result in terms of the points $(x_1, y_1), (x_2, y_2)$ and (x_3, y_3) plotted in plane? ♠

The natural generalization is that there exists a unique polynomial of degree $\leq n$ passing through $n+1$ points $(x_1, y_1), \ldots, (x_{n+1}, y_{n+1})$ with distinct x-values.

The rather miraculous trick above in (2.16) is called Lagrange interpolation and can be generalized to polynomials of arbitrary degree.

Below is an example of five points defining a unique polynomial of degree four.



2.5.1 The magic of Lagrange polynomials

(2.34) EXAMPLE.

Let us explain with a simple numerical example what happens in (2.16). Suppose we wish to find a polynomial $f(x) = a_0 + a_1 x + a_2 x^2$ through the points

$$(1,2), \qquad (2,3) \qquad \text{and} \qquad (3,5).$$

More precisely we wish to find numbers a_0, a_1 and a_2, such that

$$\begin{aligned} f(1) &= a_0 + a_1 + a_2 = 2 \\ f(2) &= a_0 + 2a_1 + 4a_2 = 3 \\ f(3) &= a_0 + 3a_1 + 9a_2 = 5. \end{aligned}$$

This is a system of three linear equations which in this case has a unique solution in a_0, a_1 and a_2.
We may, however, attack this problem in another way. Suppose that $L_1(x), L_2(x)$ and $L_3(x)$ are polynomials of degree at most two, such that

$$\begin{array}{lll} L_1(1) = 1 & L_1(2) = 0 & L_1(3) = 0 \\ L_2(1) = 0 & L_2(2) = 1 & L_2(3) = 0 \\ L_3(1) = 0 & L_3(2) = 0 & L_3(3) = 1 \end{array}$$

Then

$$f(x) = 2L_1(x) + 3L_2(x) + 5L_3(x)$$

really is the polynomial we wish to find. The insight is that these $L_1(x), L_2(x)$ and $L_3(x)$ can be explicitly written down as

$$\begin{aligned} L_1(x) &= \frac{(x-2)(x-3)}{(1-2)(1-3)} \\ L_2(x) &= \frac{(x-1)(x-3)}{(2-1)(2-3)} \\ L_3(x) &= \frac{(x-1)(x-2)}{(3-1)(3-2)}. \end{aligned}$$

♠

Example 2.34 can be generalized: suppose we have n numbers

$$x_1, x_2, \ldots, x_n. \tag{2.17}$$

Then these numbers give n polynomials each of degree $n-1$:

$$L_i(x) = \frac{1}{C_i}(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n),$$

where $C_i = (x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)$ for $i = 1, \ldots, n$.

The polynomial $L_i(x)$ is called the i-th Lagrange basis polynomial associated to the n numbers $x_1, \ldots, x_n$. It satisfies $L_i(x_1) = \cdots = L_i(x_{i-1}) = 0$, $L_i(x_i) = 1$ and $L_i(x_{i+1}) = \cdots = L_i(x_n) = 0$ i.e., $L_i(x)$ is equal to zero evaluated at all of the numbers $x_1, \ldots, x_n$ except at x_i where it evaluates to 1.

The Lagrange basis polynomials allow us to construct a polynomial f of degree $\leq n$ through $n+1$ points $(x_1,y_1),\ldots,(x_{n+1},y_{n+1})$ i.e., a polynomial f such that

$$\begin{aligned} f(x_1) &= y_1 \\ &\vdots \\ f(x_{n+1}) &= y_{n+1} \end{aligned}$$

simply as

$$f(x) = y_1 L_1(x) + \cdots + y_{n+1} L_{n+1}(x).$$

However, $f(x)$ does not have to have degree n. For example, it could come out as a line through three points $(x_1,y_1),(x_2,y_2)$ and (x_3,y_3) (see Exercise 2.33).

(2.35) EXERCISE.

Compute $a_0,a_1,a_2,a_3 \in \mathbb{R}$ so that

$$\begin{aligned} f(-2) &= -1 \\ f(-1) &= 1 \\ f(1) &= 1 \\ f(2) &= 1, \end{aligned}$$

where

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3.$$

You can do this either by Lagrange interpolation or by solving linear equations. Which one do you prefer? ♠

(2.36) EXAMPLE.

Can you predict the next number in the sequence starting with

$$15, \quad 34, \quad 65, \quad 111, \quad 175, \quad 260, \quad 369? \tag{2.18}$$

This question was posed[3] by the tutors in a class session for new computer science students. Let us put the sequence (2.18) inside a table like

n	1	2	3	4	5	6	7
$f(n)$	15	34	65	111	175	260	369

,

where $f : \mathbb{N} \to \mathbb{N}$ is the secret function responsible for the sequence. We would like to compute $f(8)$. Assuming that the $f(n)$ is a polynomial function, we may simply compute the unique polynomial of degree ≤ 6 through the 7 points

$$(1,15), \quad (2,34), \quad (3,65), \quad (4,111), \quad (5,175), \quad (6,260), \quad (7,369).$$

We know how to do this either by solving linear equations or computing with Lagrange polynomials. It turns out that Sage has built in functions helping us here.

Interactive code not included in static version.

Press the button to see what next number is in the sequence (computed using the secret polynomial). See also the description of Neville's algorithm in Wikipedia for an easier approach to computing $f(8)$.

♠

[3] Thanks to Tobias Bendsen Poulsen for notifying me about this.

2.6 Shamir secret sharing

Lagrange interpolation is used in cryptography in Shamir's secret sharing. Secret sharing is important in many practical situations. Here is an example quoted from Wikipedia:

> A company needs to secure their vault's passcode. They could encrypt it, but what if the beholder of the secret key is unavailable or turns rogue?
>
> One needs to distribute the secret. This is where SSS comes in. It can be used to encrypt the vault's passcode and generate a certain number of shares, where a certain number of shares can be allocated to each executive within the company. Now, only if they pool their shares can they unlock the vault. The threshold can be appropriately set for the number of executives, so the vault is always able to be accessed by the authorized individuals. Should a share or two fall into the wrong hands, they couldn't open the passcode unless the other executives cooperated.

The mathematics that takes care of this is surprisingly simple. Suppose the secret is the number a_0. Then we construct the polynomial

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m \tag{2.19}$$

for some other numbers $a_1, \ldots, a_m$. We know that this polynomial is uniquely given by its values in $m+1$ distinct numbers (see Remark 2.25). So if there are n trusted people we could distribute the shares

$$(1, f(1)), (2, f(2)), \ldots, (n, f(n))$$

to them. Here we suppose that $n > m$. In this setting, if there are less than $m+1$ of the people present they cannot open the vault. If $m+1$ or more people are present they can reconstruct the polynomial in (2.19), find the secret code a_0 and open the vault.

(2.37) EXERCISE.

You are in a study group consisting of four people. The professor has decided that you submit your project using a secret code that is distributed to the group members with Shamir secret sharing. At least three group members need to agree on submission.

On the day of the deadline three group members with shares

$$(1, 7035), \quad (2, 19748) \quad \text{and} \quad (3, 39373)$$

are present. What is the secret code they may use to submit their project? ♠

2.7 Fitting data

Given a data set

$$\mathscr{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

one would often like to find a model (i.e. some function) that describes the data well. With Lagrange interpolation we can find a polynomial f fitting the sample data $\mathscr{D}$ perfectly, i.e. satisfying $f(x_i) = y_i$ for $i = 1, 2, \ldots n$. Is f an optimal model? For the given data set it seems so, but we have been a bit imprecise in formulating the goal of a model.

Actually, we are not very interested in modeling the data at hand with extreme precision. What we want is a model that fits new data well. Let us look at a concrete example.

Consider the data set

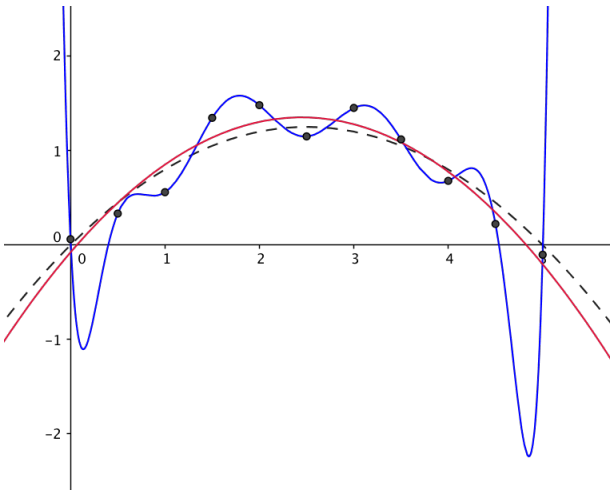
$$\mathscr{D} = \{(0, 0.06), (0.5, 0.33), (1, 0.56), (1.5, 1.35), (2, 1.48), (2.5, 1.15),$$
$$(3, 1.45), (3.5, 1.12), (4, 0.68), (4.5, 0.22), (5, -0.10)\}$$

The data points (x_i, y_i), $i = 1, 2, \ldots, 11$, were generated as $(x_i, p(x_i) + \varepsilon_i)$ where $p(x) = -0.2x^2 + x$ is a quadratic polynomial and $\varepsilon_i \in [-0.4, 0.4]$ is a random number to simulate noise. The polynomial p is the best possible model for unknown data as there will always by noise that can not be modeled. In real life p is what needs to be modeled based on the available data.

In the figure below is a fit with a degree 2 and degree 10 polynomial respectively. As we see, the degree 2 polynomial is pretty close to the target p compared to the degree 10 polynomial that nevertheless fits the data $\mathscr{D}$ perfectly. Generally a simple model is preferred over a complex, as the latter will have a tendency to fit noise. This phenomenon is called overfitting and is an extremely important topic.

An interactive version of this illustration with a little more bells and whistles can be found here.

(2.38) FIGURE.



Fitting a degree 2 *polynomial (in red) and a degree* 10 *polynomial (in blue) to the sample data* $\mathscr{D}$*. The target function* p *is the dashed curve. We see that the simple quadratic fit is much closer to the target function and hence performes better on new data.*

In later chapters we will see how the degree two polynomial fit was obtained. This is a nice example of a convex optimization problem.

Chapter 3

Matrices

Handling linear equations and keeping track of the unknowns can be a pain. At a certain point one needs to simplify the notation. This is done introducing matrices.

For example, the system of equations

$$\begin{array}{rcrcrcr} & & 2y & + & 4z & = & -2 \\ 3x & + & 2y & + & 7z & = & 4 \end{array} \tag{3.1}$$

can be represented by the rectangular array (matrix)

$$\begin{pmatrix} 0 & 2 & 4 & -2 \\ 3 & 2 & 7 & 4 \end{pmatrix} \tag{3.2}$$

of numbers. Many of the operations we do to solve linear equations might as well be done on this array forgetting about the unknowns.

3.1 Matrices

3.1.1 Definitions

A rectangular array of numbers is called a *matrix*. A matrix with m *rows* and n *columns* is called an $m \times n$ (m by n) matrix. The notation for an $m \times n$ matrix A is

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mj} & \cdots & a_{mn} \end{pmatrix}, \tag{3.3}$$

where $A_{ij} = a_{ij}$ denotes the number or *entry* in the i-th row and j-th column. If the matrix in (3.2) is denoted A, then it has 2 rows and 4 columns with $A_{14} = -2$.

Two matrices are equal if they have the same number of rows and columns and their entries are identical.

A very useful (and famous) open source library in python (with 1000+ contributors) for handling matrices is `NumPy`. Here is how the matrix in (3.2) is entered in `NumPy`.

Interactive code not included in static version.

1. A matrix whose entries are all 0 is called a zero matrix. It is denoted simply by 0, when it is clear from the context what its numbers of rows and columns are.

2. A matrix is called *quadratic* if it has an equal number of rows and columns.

   The first two matrices below are quadratic, whereas the third is not.

$$\begin{pmatrix}1\end{pmatrix}, \qquad \begin{pmatrix}1 & 2 & 3\\ 4 & 5 & 6\\ 7 & 8 & 9\end{pmatrix}, \qquad \begin{pmatrix}0 & 1 & 0\\ 1 & 0 & 1\end{pmatrix}.$$

3. The *diagonal* in a matrix is defined as the entries in the matrix with the same row- and column indices.

   Below we have a 3×4 matrix with the diagonal elements marked

$$\begin{pmatrix}1 & 3 & 0 & 1\\ 3 & 2 & 1 & 5\\ 1 & 0 & 3 & 6\end{pmatrix}.$$

   A matrix is called a *diagonal matrix*, if all its entries outside the diagonal are $=0$. Below is an example of a square diagonal matrix

$$\begin{pmatrix}1 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 3\end{pmatrix}.$$

4. A matrix is called a *row vector* if it has only one row.

   For example,

$$\begin{pmatrix}1 & 2 & 3\end{pmatrix}$$

   is a row vector with three columns.

5. A matrix is called a *column vector* if it has only one column.

   For example,

$$\begin{pmatrix}1\\ 2\\ 3\end{pmatrix}$$

   is a column vector with three rows.

6. The rows in a matrix are called the *row vectors* of the matrix. The i-th row in a matrix A is denoted A_i.

   The matrix A in (3.2) contains the row vectors

$$A_1 = \begin{pmatrix}0 & 2 & 4 & -2\end{pmatrix} \qquad \text{and} \qquad A_2 = \begin{pmatrix}3 & 2 & 7 & 4\end{pmatrix}.$$

7. The columns in a matrix are called the *column vectors* of the matrix. The j-th column in a matrix A is denoted A^j[1].

   The matrix A in (3.2) contains the column vectors

$$A^1 = \begin{pmatrix}0\\ 3\end{pmatrix}, \quad A^2 = \begin{pmatrix}2\\ 2\end{pmatrix}, \quad A^3 = \begin{pmatrix}4\\ 7\end{pmatrix} \quad \text{and} \quad A^4 = \begin{pmatrix}-2\\ 4\end{pmatrix}.$$

8. A row- or column vector is referred to as a *vector*.

9. Even though we have used the notation $\mathbb{R}^n$ for the n-th cartesian product of $\mathbb{R}$, we will use $\mathbb{R}^n$ henceforth to denote the set of column vectors with n rows (entries). This definition is almost identical with the previous one, except that the tuple is formatted as a column vector.

   Illustrated by an example,

$$\begin{pmatrix}1\\ 2\\ 3\end{pmatrix} \in \mathbb{R}^3 \qquad \text{instead of} \qquad (1,2,3) \in \mathbb{R}^3.$$

[1] Not to be confused with powers of the matrix A introduced later.

3.2 Linear maps

In the first chapter we encountered a miniature version of a neural network. Neural networks are generally incredibly complicated functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. The function $f : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x^7 y + \cos(xy)e^{x^2+y^2-1} \\ 2xy^2 - \sin(x+y)(x^3+y^3) \end{pmatrix},$$

even though it looks complicated, is simple in comparison.

You probably agree that the function $g : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$g\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2x+3y \\ 3x-2y \end{pmatrix}$$

is even simpler. This function (or map) is an example of a linear map.

In general, a *linear map* $f : \mathbb{R}^n \to \mathbb{R}^m$ has the form

$$f\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{pmatrix},$$

where $a_{11}, \ldots, a_{mn}$ are mn real numbers.

Using matrices we will use the notation

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{pmatrix}.$$

In this way, we can write the map f as

$$f(v) = Av,$$

where A is the $m \times n$ matrix

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

and v is the vector

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

in $\mathbb{R}^n$.

Basically a linear map is a system of linear equations without the right hand side (including $=$). In fact, we may write the system of linear equations in (3.1) as

$$\begin{pmatrix} 0 & 2 & 4 \\ 3 & 2 & 7 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -2 \\ 4 \end{pmatrix}.$$

(3.1) EXERCISE.

Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be the linear map given by the 2×2 matrix

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

Does there exist $u \in \mathbb{R}^2$, such that

$$f(u) = \begin{pmatrix} 3 \\ 7 \end{pmatrix}?$$

Quite generally, can we find $u \in \mathbb{R}^2$, such that

$$f(u) = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}?$$

for arbitrary $b_1, b_2 \in \mathbb{R}$? ♠

(3.2) EXERCISE.

Suppose you know that $f : \mathbb{R}^n \to \mathbb{R}^m$ is a linear map and that you have a black box giving you output $f(v) \in \mathbb{R}^m$ if you supply the input $v \in \mathbb{R}^n$. How would you find the matrix defining f? ♠

3.3 Matrix multiplication

Suppose we are given two linear maps $f : \mathbb{R}^2 \to \mathbb{R}^2$ and $g : \mathbb{R}^2 \to \mathbb{R}^2$. Then it turns out that the composition $f \circ g : \mathbb{R}^2 \to \mathbb{R}^2$ is also a linear map. A word of advice: the computations below look large and intimidating. They are not. It is important that you carry them out on your own. Do not look and copy or tell yourself that it looks okay. Do the computations yourself and ask me or fellow students if you get stuck.

Let us look at an example. Suppose that

$$g\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ -1 & -2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \qquad \text{and} \qquad f\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}.$$

Then

$$(f \circ g)\begin{pmatrix} x \\ y \end{pmatrix} = f\left(g\begin{pmatrix} x \\ y \end{pmatrix}\right) = f\left(\begin{pmatrix} 2 & 3 \\ -1 & -2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} 1 & 2 \\ 1 & -2 \end{pmatrix}\left(\begin{pmatrix} 2 & 3 \\ -1 & -2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}\right)$$

$$= \begin{pmatrix} 1 & 2 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} 2x+3y \\ -x-2y \end{pmatrix} = \begin{pmatrix} -y \\ 4x+7y \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 4 & 7 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

In terms of the matrices of the linear maps, we write this as

$$\begin{pmatrix} 1 & 2 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} 2 & 3 \\ -1 & -2 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 4 & 7 \end{pmatrix} \tag{3.4}$$

There is nothing special about the numbers in this example. We might as well do the computation in general: suppose that

$$g\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \qquad \text{and} \qquad f\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}.$$

Then

$$(f\circ g)\begin{pmatrix}x\\y\end{pmatrix}=f\left(g\begin{pmatrix}x\\y\end{pmatrix}\right)=f\left(\begin{pmatrix}b_{11}&b_{12}\\b_{21}&b_{22}\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}\right)=\begin{pmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{pmatrix}\left(\begin{pmatrix}b_{11}&b_{12}\\b_{21}&b_{22}\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}\right)$$

$$=\begin{pmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{pmatrix}\begin{pmatrix}b_{11}x+b_{12}y\\b_{21}x+b_{22}y\end{pmatrix}=\begin{pmatrix}a_{11}(b_{11}x+b_{12}y)+a_{12}(b_{21}x+b_{22}y)\\a_{21}(b_{11}x+b_{12}y)+a_{22}(b_{21}x+b_{22}y)\end{pmatrix}$$

$$=\begin{pmatrix}(a_{11}b_{11}+a_{12}b_{21})x+(a_{11}b_{12}+a_{12}b_{22})y\\(a_{21}b_{11}+a_{22}b_{21})x+(a_{21}b_{12}+a_{22}b_{22})y\end{pmatrix}$$

$$=\begin{pmatrix}a_{11}b_{11}+a_{12}b_{21}&a_{11}b_{12}+a_{12}b_{22}\\a_{21}b_{11}+a_{22}b_{21}&a_{21}b_{12}+a_{22}b_{22}\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}.$$

Again, in terms of the matrices of the linear maps, we write this as

$$\begin{pmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{pmatrix}\begin{pmatrix}b_{11}&b_{12}\\b_{21}&b_{22}\end{pmatrix}=\begin{pmatrix}a_{11}b_{11}+a_{12}b_{21}&a_{11}b_{12}+a_{12}b_{22}\\a_{21}b_{11}+a_{22}b_{21}&a_{21}b_{12}+a_{22}b_{22}\end{pmatrix}\tag{3.5}$$

The equation above is the formula for matrix multiplication for two 2×2 matrices, precisely as it was introduced by Cayley around 1857.

Upon closer inspection (and colored in (3.5) for $i=2$ and $j=1$), you will see that the number in the i-th row and j-th column in the product matrix is the *row-column multiplication* between the i-th row and the j-th column in the two matrices:

The *row-column multiplication* between a row vector

$$x=(x_1x_2\ldots x_n)\qquad\text{and a column vector}\qquad y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}$$

with the same number of entries is defined as

$$xy=x_1y_1+x_2y_2+\cdots+x_ny_n.$$

(3.3) DEFINITION.

Let A be an $m\times n$ matrix and B an $n\times r$ matrix. Then the matrix product AB is defined as the $m\times r$ matrix C given by the row-column multiplication

$$C_{ij}=A_iB^j=A_{i1}B_{1j}+A_{i2}B_{2j}+\cdots+A_{in}B_{nj}$$

for $1\le i\le m$ and $1\le j\le r$.

If A is an $m\times n$ matrix and B is an $r\times s$, then the matrix product AB only makes sense if $n=r$: the number of columns in A must equal the number of rows in B.

(3.4) QUIZ.

Quiz not included in static version. ♠

(3.5) VIDEO.

I have been told that my pronunciation of column *in the video below is wrong. In the area of the US, where I got my PhD, people for some reason had this (Irish?) rare pronunciation.*

Link to video

Using matrix product notation, the system of linear equations in (3.1) can now be written as

$$\begin{pmatrix} 0 & 2 & 4 \\ 3 & 2 & 7 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -2 \\ 4 \end{pmatrix}$$

Here we multiply a 2×3 with a 3×1 matrix. The row-column multiplication gives the 2×1 matrix

$$\begin{pmatrix} 2y+4z \\ 3x+2y+7z \end{pmatrix}.$$

This matrix must equal the 2×1 matrix on the right hand side for (3.1) to be true. This is in agreement with our convention for writing linear maps in section 3.2.

(3.6) QUIZ.

Quiz not included in static version. ♠

3.3.1 Matrix multiplication in `numpy`

Matrix multiplication in `numpy` is represented by the function `dot`:

Interactive code not included in static version.

3.3.2 The identity matrix

The identity matrix I_n of order n is the $n \times n$ diagonal matrix with 1 in the diagonal. Below is the identity matrix of order 5.

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The identity matrix I_n has the crucial property that

$$I_n A = A I_n = A \tag{3.6}$$

for all $n \times n$ matrices A.

Interactive code not included in static version.

(3.7) EXERCISE.

Prove that the two identities in (3.6) are true for $n \times n$ matrices. ♠

3.3.3 Examples of matrix multiplication

Matrix multiplication is omnipresent in mathematics. Below we give an example, which is a baby version of Google's famous page rank algorithm.

(3.8) EXAMPLE.

Suppose that 20% of the people living in the suburbs move to the big city and that 30% of the people living in the big city move to the suburbs per year.

Aiming for a model using probabilities, let us be a bit more precise.

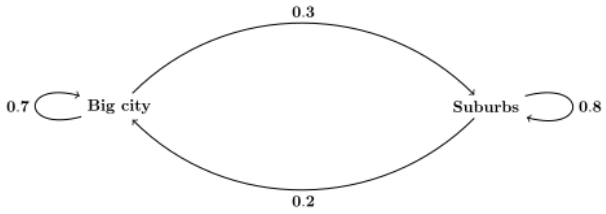
(i) If you live in the suburbs, the probability that you move to the big city is 0.2,

(ii) If you live in the suburbs, the probability that you do not move is 0.8.

(iii) If you live in the big city the probability that you move to the suburbs is 0.3.

(iv) If you live in the big city the probability that you do not move is 0.7.

All of the above probabilities are per year and can be illustrated in the diagram below



We are interested in predicting, using this model, how many people live in the big city and the suburbs given that we know how many people live in the big city, x_0 and in the suburbs y_0 to begin with i.e., setting the time $t = 0$ (years).

How many people x_1 and y_1 live in the two places after the first year ($t = 1$)?

The population of the big city will decrease by 30%, but there are newcomers amounting to 20% of the population in the suburbs. Therefore

$$x_1 = 0.7x_0 + 0.2y_0.$$

In the same way,

$$y_1 = 0.3x_0 + 0.8y_0.$$

Using matrix multiplication, these two equations can be written

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}.$$

For $t = 2$ years, we can repeat the procedure and the result becomes

$$\begin{aligned} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} &= \begin{pmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{pmatrix} \left(\begin{pmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \right) \\ &= \left(\begin{pmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{pmatrix} \begin{pmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{pmatrix} \right) \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = P^2 \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}, \end{aligned} \tag{3.7}$$

where

$$P = \begin{pmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{pmatrix}. \tag{3.8}$$

In general we have the formula

$$\begin{pmatrix} x_n \\ y_n \end{pmatrix} = P^n \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}, \tag{3.9}$$

giving the distribution of the populations for $t = n$ years.

Let us experiment a little:

$$\begin{aligned} P^2 &= \begin{pmatrix} 0.55 & 0.3 \\ 0.45 & 0.7 \end{pmatrix} \\ P^3 = PP^2 &= \begin{pmatrix} 0.475 & 0.35 \\ 0.525 & 0.65 \end{pmatrix} \\ P^4 = PP^3 &= \begin{pmatrix} 0.4375 & 0.375 \\ 0.5625 & 0.625 \end{pmatrix} \\ &\vdots \\ P^{15} &= \begin{pmatrix} 0.400018 & 0.399951 \\ 0.599982 & 0.600012 \end{pmatrix} \\ P^{16} &= \begin{pmatrix} 0.400009 & 0.399994 \\ 0.599991 & 0.600006 \end{pmatrix} \end{aligned}$$

It seems that the distribution stabilizes around 40% living in the big city and 60% living in the suburbs of the original total population.

Interactive code not included in static version.

The matrix P is an example of a stochastic 2×2 matrix. In general, a square matrix is called a *stochastic matrix* if its entries are ≥ 0 and the sum of the entries in its column vectors are 1. ♠

(3.9) EXAMPLE.

A simple example of the page rank algorithm is given in Example 2.15. There you encountered the equations

$$\begin{aligned} T_2 &= T_1 + \tfrac{1}{2}T_4 \\ T_3 &= T_2 \\ T_4 &= T_3 \\ T_1 &= \tfrac{1}{2}T_4 \\ T_1 + T_2 + T_3 + T_4 &= 1. \end{aligned}$$

In terms of matrix multiplication the first four equations can be rewritten to

$$\begin{pmatrix} 0 & 0 & 0 & \frac{1}{2} \\ 1 & 0 & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{pmatrix} = \begin{pmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{pmatrix}.$$

Putting

$$P = \begin{pmatrix} 0 & 0 & 0 & \frac{1}{2} \\ 1 & 0 & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

we get a stochastic matrix and may again iterate and compute $P, P^2, P^3, \ldots$.

> Interactive code not included in static version.

Is there a connection between the entries of P^N, where N is very big and the solutions to the linear equations? ♠

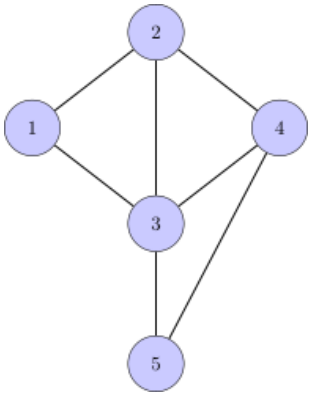
(3.10) EXERCISE.

In the end of Example 3.8 (above) a stochastic matrix is defined. Show that the matrix product of two $n \times n$ stochastic matrices is a stochastic matrix. ♠

Below is an example, where matrix multiplication occurs in networks.

(3.11) EXAMPLE.

Suppose we have five cities connected with roads as shown below



This network has a so called 5×5 *incidence matrix*, where city i is associated with the i-th row and i-th column. A 1 in the matrix in the (i, j) entry means that there is a road from city i to city j, whereas a 0 means that city i and city j are not connected by a road:

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

Here

$$A^2 = \begin{pmatrix} 2 & 1 & 1 & 2 & 1 \\ 1 & 3 & 2 & 1 & 2 \\ 1 & 2 & 4 & 2 & 1 \\ 2 & 1 & 2 & 3 & 1 \\ 1 & 2 & 1 & 1 & 2 \end{pmatrix} \quad \text{and} \quad A^3 = \begin{pmatrix} 2 & 5 & 6 & 3 & 3 \\ 5 & 4 & 7 & 7 & 3 \\ 6 & 7 & 6 & 7 & 6 \\ 3 & 7 & 7 & 4 & 5 \\ 3 & 3 & 6 & 5 & 2 \end{pmatrix}.$$

What is the interpretation of A^2, A^3 and A^n in general? It turns out that the entry (i, j) in the matrix A^n exactly is the number of paths of length n from city i to city j.

For example, there are 3 paths from city 1 to city 5 of length 3 corresponding to the paths $1245, 1345, 1235$. The 2 paths from city 1 to city 1 of length 3 are $1231, 1321$ and the 5 paths of length 3 from city 1 to city 2 are $1342, 1242, 1312, 1212, 1232$.

A deeper explanation: Suppose that we have a network with m cities and incidence matrix A.

The general proof of the observations above in our special example, builds on the fact that a path of length n from city i to city j has to end with a road from a neighboring city k to j. For every one of these neighboring cities, we may count the number of paths of length $n-1$ from city i. If A^{n-1}_{gh} is the number of paths of length $n-1$ from city g to city h, then matrix multiplication tells us that

$$A^n_{ij} = A^{n-1}_{i1}A_{1j} + \cdots + A^{n-1}_{im}A_{mj}$$

This number is exactly the number of paths of length n from city i to city j, since $A_{kj} = 1$ only when k is a neighboring city to city j (and 0 otherwise). ♠

3.4 Matrix arithmetic

Matrix multiplication is very different from ordinary multiplication of numbers: it is not commutative. Consider the matrices

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

Then

$$AB = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \qquad \text{and} \qquad BA = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

i.e.., $AB \neq BA$.

Interactive code not included in static version.

Addition of matrices is like ordinary addition, except that you add all the entries of the involved matrices.

3.4.1 Matrix addition

Addition of two matrices with the same number of rows and columns is defined below.

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mn} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & \cdots & a_{1n}+b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & \cdots & a_{mn}+b_{mn} \end{pmatrix}.$$

The zero matrix is the $(m \times n)$ matrix containing zero in all its entries. When its number of rows and columns are clear from the context it is simply denoted by 0. For 2×3 matrices for example, we write

$$0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

(3.12) EXERCISE.

Given an example of a non-zero 2×2 matrix, such that

$$A^2 = 0.$$

♠

3.4.2 Multiplication of a number and a matrix

A matrix may be multiplied by a number λ by multiplying each entry by the number:

$$\lambda \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = \begin{pmatrix} \lambda a_{11} & \cdots & \lambda a_{1n} \\ \vdots & \ddots & \vdots \\ \lambda a_{m1} & \cdots & \lambda a_{mn} \end{pmatrix}.$$

(3.13) EXERCISE.

Does there exists a number λ, such that

$$\lambda \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 4 & 6 \\ 8 & 10 & 15 \end{pmatrix}?$$

♠

(3.14) EXERCISE.

Let A be a 2×2 matrix, such that

$$AB = BA,$$

for every other 2×2 matrix B. Show that A is a diagonal matrix of the form

$$A = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix},$$

where $a \in \mathbb{R}$ i.e., $A = aI_2$. ♠

3.4.3 The distributive law

Ordinary numbers a, b, c satisfy $a(b+c) = ab+ac$. This rule also holds for matrices and is called the distributive law (multiplication is distributed over plus)

(3.15) PROPOSITION.

Let B and C be $m \times n$ matrices, A an $r \times m$ matrix and D an $n \times s$ matrix. Then

$$A(B+C) = AB+AC \qquad \textit{and} \qquad (B+C)D = BD+CD.$$

Proof. Let us start by looking at $A(B+C) = AB+AC$. Here it suffices to do the proof, when A is a row vector and B, C column vectors, since

$$(A(B+C))_{ij} = A_i(B+C)^j = A_i(B^j + C^j).$$

For $(B+C)D = BD+CD$, we may reduce to the case, where B, C are row vectors and D a column vector, since

$$((B+C)D)_{ij} = (B+C)_i D^j = (B_i + C_i)D^j.$$

Both of these cases follow using the distributive law for ordinary numbers. □

(3.16) EXERCISE.

Suppose that A and B are two 2×2 matrices. Is it true that

$$(A+B)^2 = A^2 + B^2 + 2AB?$$

What about

$$(A+B)(A-B) = A^2 - B^2?$$

♠

3.4.4 The miraculous associative law

It does not make sense to multiply three matrices A, B and C. We have only defined matrix multiplication for two matrices. There are two natural ways of evaluating ABC:

$$(AB)C \qquad \text{and} \qquad A(BC).$$

We can begin by multiplying A by B and then multiply C from the right. However, we may just as well start by multiplying B by C and then multiply A from the left.

It is in no way clear, that these two computations give the same result!

That this turns out to be true, is just one of many miracles in the universe (there is a rather cool mathematical explanation, though, addressed in an exercise below).

(3.17) THEOREM.

Let A be an $m \times n$ matrix, B an $n \times r$ matrix and C an $r \times s$ matrix. Then

$$(AB)C = A(BC).$$

Proof. We must prove that

$$((AB)C)_{ij} = (A(BC))_{ij}$$

for $1 \leq i \leq m$ og $1 \leq j \leq s$. The left hand side can be written

$$\begin{aligned}(AB)_i C^j &= (A_i B^1, \ldots, A_i B^r) C^j \\ &= (A_i B^1) C_{1j} + (A_i B^2) C_{2j} + \cdots + (A_i B^r) C_{rj}.\end{aligned} \tag{3.10}$$

The right hand side is

$$A_i (BC)^j = A_i \begin{pmatrix} B_1 C^j \\ \vdots \\ B_n C^j \end{pmatrix} = A_{i1}(B_1 C^j) + \cdots + A_{in}(B_n C^j). \tag{3.11}$$

Writing the row-column multiplications in (3.10), we get

$$\begin{aligned} &A_{i1}B_{11}C_{1j} + \cdots + A_{in}B_{n1}C_{1j} + \\ &A_{i1}B_{12}C_{2j} + \cdots + A_{in}B_{n2}C_{2j} + \\ &\vdots \\ &A_{i1}B_{1r}C_{rj} + \cdots + A_{in}B_{nr}C_{rj}. \end{aligned} \tag{3.12}$$

Writing the row-column multiplications in (3.11), we get

$$\begin{aligned} &A_{i1}B_{11}C_{1j}+\cdots+A_{i1}B_{1r}C_{rj}+\\ &A_{i2}B_{21}C_{1j}+\cdots+A_{i2}B_{2r}C_{rj}+\\ &\vdots\\ &A_{in}B_{n1}C_{1j}+\cdots+A_{in}B_{nr}C_{rj}. \end{aligned} \tag{3.13}$$

The rows in the sum in (3.12) correspond to the columns in the sum (3.13). Therefore these sums are equal and $((AB)C)_{ij} = (A(BC))_{ij}$. □

(3.18) REMARK.

The associative law $(AB)C = A(BC)$ is true, but in computing ABC there can be a (big) difference in the number of multiplications in the two computations $A(BC)$ and $(AB)C$ i.e., efficiency is not associative for matrix multiplication. In the notation of Theorem 3.17, computing $(AB)C$ requires

$$mnr + mrs = mr(n+s)$$

multiplications, whereas computing $A(BC)$ requires

$$nrs + mns = ns(m+r)$$

multiplications. If for example $m = 10000, n = 10, r = 10000$ and $s = 10$, then computing $(AB)C$ requires $2 \cdot 10^9$ multiplications, whereas computing $A(BC)$ requires $2 \cdot 10^6$ multiplications!

(3.19) EXERCISE.

Verify the associative law for the three matrices

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \qquad \text{and} \qquad C = \begin{pmatrix} 6 & 3 \\ 5 & 2 \\ 4 & 1 \end{pmatrix}$$

by showing by explicit computation that

$$(AB)C = A(BC).$$

♠

(3.20) EXERCISE.

There is in fact a high tech explanation that the associative law for matrices holds. An explanation that makes the calculations in the above proof superfluous and shows the raw power of abstract mathematics: suppose that $f : \mathbb{R}^n \to \mathbb{R}^m, g : \mathbb{R}^r \to \mathbb{R}^n$ and $h : \mathbb{R}^s \to \mathbb{R}^r$ are linear maps. Then $f \circ (g \circ h)$ and $(f \circ g) \circ h$ are both linear maps from $\mathbb{R}^s \to \mathbb{R}^m$, such that

$$(f \circ (g \circ h))(x) = ((f \circ g) \circ h)(x) = f(g(h(x)))$$

for every $x \in \mathbb{R}^s$. How does this relate to the associative law for matrix multiplication? ♠

3.5 The inverse matrix

You are allowed to divide by a number provided it is $\neq 0$. Does it makes sense to divide by matrices?

It does, but there are some matrices that correspond to the number 0 that we are not allowed to divide by.

(3.21) EXERCISE.

Let A, B and C be $n \times n$ matrices. Show that

$$BA = I_n$$

and

$$AC = I_n$$

implies that $B = C$. ♠

(3.22) DEFINITION.

An $n \times n$ matrix A is called invertible, if there exists an $n \times n$ matrix B, such that

$$AB = BA = I_n.$$

In this case, B is called the inverse matrix of A and denoted A^{-1}.

As a reality check, you should convince yourself that the 2×2 matrix

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

is not invertible. In fact, here the associative law from Theorem 3.17 is incredibly useful: if A is an invertible matrix with inverse matrix B and $AC = 0$, then $C = 0$:

$$AC = 0 \implies B(AC) = B0 = 0 \implies (BA)C = I_n C = C = 0.$$

(3.23) EXERCISE.

Show that a quadratic matrix with a column or row consisting entirely of zeros cannot be invertible. ♠

(3.24) EXERCISE.

Suppose that

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with $D = ad - bc \neq 0$. Prove that A is invertible with

$$A^{-1} = \frac{1}{D} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

♠

(3.25) EXERCISE.

When is a quadratic diagonal matrix invertible? Look first at the 2×2 case:

$$\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}.$$

♠

The inverse matrix can be computed in `numpy`:

Interactive code not included in static version.

The inverse matrix enters the picture when solving n linear equations with n unknowns:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned}$$

can be rewritten using matrix notation as

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

or more compactly as $Ax = b$.

If A is invertible, then the associative law gives the following:

$$\begin{aligned} Ax = b &\iff \\ A^{-1}(Ax) = A^{-1}b &\iff \\ (A^{-1}A)x = A^{-1}b &\iff \\ Ix = A^{-1}b &\iff \\ x = A^{-1}b. & \end{aligned}$$

The inverse matrix gives the solution to the linear equations $Ax = b$ just by one matrix multiplication!

(3.26) EXAMPLE.

The system of linear equations

$$\begin{array}{rcrcl} 5x & + & 3y & = & 13 \\ 3x & + & 2y & = & 8 \end{array} \tag{3.14}$$

can be rewritten using matrix multiplication to

$$Av = b,$$

where

$$A = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}, \qquad v = \begin{pmatrix} x \\ y \end{pmatrix} \qquad \text{and} \qquad b = \begin{pmatrix} 13 \\ 8 \end{pmatrix}.$$

Here A is invertible and

$$A^{-1} = \begin{pmatrix} 2 & -3 \\ -3 & 5 \end{pmatrix}.$$

One simple matrix multiplication

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 & -3 \\ -3 & 5 \end{pmatrix} \begin{pmatrix} 13 \\ 8 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

shows the solution we expect from (3.14). ♠

The product of two invertible matrices (when this makes sense) is an invertible matrix. This is the content of the following result.

(3.27) PROPOSITION.

The product AB of two invertible matrices A and B is invertible and $(AB)^{-1} = B^{-1}A^{-1}$.

Proof. We must check that

$$(B^{-1}A^{-1})(AB) = I \qquad \text{and} \qquad AB(B^{-1}A^{-1}) = I.$$

Let us check the first condition using the associative law:

$$\begin{aligned} (B^{-1}A^{-1})(AB) &= ((B^{-1}A^{-1})A)B \\ &= (B^{-1}(A^{-1}A))B \\ &= (B^{-1}I)B = B^{-1}(IB) = B^{-1}B = I, \end{aligned}$$

where I denotes the identity matrix. The condition $AB(B^{-1}A^{-1}) = I$ is verified in the same way. □

(3.28) EXERCISE.

We have defined a matrix A to be invertible if there exists a matrix B, such that $AB = I$ and $BA = I$. Suppose that only $BA = I$. Can we then conclude that $AB = I$?

Find the mistake in the argument below.

Suppose that $BA = I$. Then for every $y \in \mathbb{R}^n$ we have $Ax = y \implies x = (BA)x = By$. Therefore $A(By) = (AB)y = y$ for every $y \in \mathbb{R}^n$ and we have proved that $AB = I$.

♠

(3.29) EXERCISE.

Let

$$N = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Compute the powers N^k for $k \geq 2$ i.e., $N^2, N^3, \ldots$. Now let

$$A = I + N,$$

where $I = I_4$. Show that A is invertible, and

$$A^{-1} = I - N + N^2 - N^3.$$

Compute A^{-1}.

Do you see a way of generalizing this computation to $n \times n$ matrices N with a property shared by the 4×4 matrix above?

♠

3.5.1 Well, how do I find the inverse of a matrix?

Finding the inverse of a matrix or deciding that the matrix is not invertible is a matter of solving systems of linear equations.

Given an $n \times n$ matrix A, we need to see if there exists an $n \times n$ matrix B, such that

$$AB = I, \tag{3.15}$$

where $I = I_n$ is the identity matrix of order n. We can do this by computing the columns of B. From the definition in (3.15), the j-th column B^j of B must satisfy

$$AB^j = I^j. \tag{3.16}$$

This follows from the definition of matrix multiplication!

The identity in (3.16) is a system of n linear equations in n unknowns. The unknowns are the entries in the j-th column B^j of the inverse matrix A^{-1} (if it exists).

(3.30) EXAMPLE.

Suppose that A is a 2×2 matrix. Then the inverse matrix B (if it exists) can be computed from the systems of linear equations below.

$$AB^1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \text{and} \qquad AB^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Writing

$$\begin{pmatrix} x \\ y \end{pmatrix} = B^1 \qquad \text{and} \qquad \begin{pmatrix} u \\ v \end{pmatrix} = B^2$$

for the first and second columns, the systems of linear equations can be written as

$$\begin{array}{rcrcl} A_{11}x & + & A_{12}y & = & 1 \\ A_{21}x & + & A_{22}y & = & 0 \end{array} \qquad \text{and} \qquad \begin{array}{rcrcl} A_{11}u & + & A_{12}v & = & 0 \\ A_{21}u & + & A_{22}v & = & 1' \end{array}$$

where

$$B = \begin{pmatrix} x & u \\ y & v \end{pmatrix}.$$

A concrete example along with a useful way of keeping track of the computation is presented in the video below.

(3.31) VIDEO.

Link to video ♠

(3.32) EXERCISE.

Compute the inverse of the matrix

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix}$$

by employing the method of solving linear equations above. Explain the steps in your computation. You may find it useful to collect inspiration from the video in Example 3.30. ♠

3.6 The transposed matrix

The transpose of an $m \times n$ matrix A is the $n \times m$ matrix $A^\top$ given by

$$A^\top_{ij} = A_{ji},$$

where $1 \leq i \leq m$ and $1 \leq j \leq n$. As an example,

$$\begin{pmatrix} 0 & 2 & 4 & -2 \\ 3 & 2 & 7 & 4 \end{pmatrix}^\top = \begin{pmatrix} 0 & 3 \\ 2 & 2 \\ 4 & 7 \\ -2 & 4 \end{pmatrix}.$$

Notice also that $(A^\top)^\top = A$ for an arbitrary matrix A.

(3.33) PROPOSITION.

Let A be an $m \times r$ matrix and B an $r \times n$ matrix. Then

$$(AB)^\top = B^\top A^\top.$$

Proof. By definition $(AB)^\top_{ij} = (AB)_{ji}$. This entry is given by row-column multiplication of the j-th row in A and the i-th column in B, which is the row-column multiplication of the i-th row in $B^\top$ and the j-th column in $A^\top$. □

(3.34) EXERCISE.

Let A be a quadratic matrix. Prove that A is invertible if and only if $A^\top$ is invertible. ♠

(3.35) EXERCISE.

In the sage window below, you are supposed to experiment a bit by entering an arbitrary matrix B and studying the quadratic matrix $BB^\top$. Is there anything special about this product? Press the *Further explanation* button below the sage window to display the rest of the exercise after(!) you have completed your experimentation.

Interactive code not included in static version.

Further explanation: A quadratic matrix A is called symmetric if $A = A^\top$. Prove that

$$BB^\top$$

is a symmetric matrix, where B is an arbitrary matrix. ♠

3.7 Symmetric matrices

A (quadratic) matrix A is called symmetric if $A = A^\top$. Visually, this means that A is symmetric around the diagonal like the 3×3 matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 4 \\ 3 & 4 & 6 \end{pmatrix},$$

but not like the 3×3 matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

(3.36) EXERCISE.

Show that

$$B^\top AB$$

is a symmetric matrix, when A is a symmetric matrix and B is an arbitrary matrix. Both matrices are assumed quadratic of the same dimensions. ♠

If A is a symmetric $n \times n$ matrix, we define the function $f_A : \mathbb{R}^n \to \mathbb{R}$ given by

$$f_A(v) = v^\top Av.$$

This definition is rather compact. Let us consider the following example for $n = 2$.

$$A = \begin{pmatrix} a & c \\ c & b \end{pmatrix} \quad \text{and} \quad v = \begin{pmatrix} x \\ y \end{pmatrix}.$$

Then

$$\begin{aligned} [emph]v^\top Av &= \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} a & c \\ c & b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} ax+cy \\ cx+by \end{pmatrix} \\ &= x(ax+cy) + y(cx+by) = ax^2 + by^2 + 2cxy. \end{aligned}$$

You are also encouraged to watch the short video below for an example with concrete numbers.

(3.37) VIDEO.

Link to video

Inside the set of the symmetric matrices we find two very important subsets of matrices: the positive definite and the positive semi-definite matrices. They correspond to positive and non-negative real numbers.

3.7.1 Positive definite matrices

A symmetric matrix A is called *positive definite* if

$$f_A(v) > 0$$

for every $v \in \mathbb{R}^n \setminus \{0\}$. Probably the first example of a positive definite 2×2-matrix one thinks of is A being the identity matrix. Here

$$\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = x^2 + y^2.$$

Of course, here $x^2 + y^2 = 0$ if and only if $x = y = 0$ or $\begin{pmatrix} x \\ y \end{pmatrix} = 0$.

(3.38) EXERCISE.

Give examples of (non-zero) 1×1 and 2×2 matrices that are positive definite and ones that fail to be positive definite.

When is a 2×2 diagonal matrix positive definite? ♠

(3.39) EXERCISE.

Let A be a symmetric $n \times n$ matrix. Show that A is not positive definite if $A_{11} < 0$. ♠

3.7.2 Positive semi-definite matrices

A symmetric matrix A is called *positive semi-definite* if

$$f_A(v) \geq 0$$

for every $v \in \mathbb{R}^n$. A positive definite matrix is positive semi-definite. Probably the first example of a non positive definite, but positive semi-definite 2×2-matrix one thinks of is A being the zero matrix. Here

$$\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0,$$

for every $x, y \in \mathbb{R}$.

(3.40) EXERCISE.

Give an example of a non-zero matrix that is positive semi-definite, but not positive definite.

When is a 2×2 diagonal matrix positive semi-definite? ♠

3.7.3 Symmetric reductions

As you probably have noticed, it is rather straightforward to see when a diagonal matrix is positive (semi)definite. For a general symmetric matrix, one needs to reduce to the case of a diagonal matrix. This is done using the following result.

(3.41) PROPOSITION.

Let A be a symmetric $n \times n$ matrix and B an invertible $n \times n$ matrix. Then A is positive (semi) definite if and only if

$$B^\top AB$$

is positive (semi) definite.

Proof. Every vector $v \in \mathbb{R}^n$ is equal to Bu for a unique $u \in \mathbb{R}^n$, since B is invertible. Why? The upshot is that the equation

$$v = Bu$$

can be solved by multiplying both sides by B^{-1} giving

$$v = Bu \iff B^{-1}v = B^{-1}(Bu) = (B^{-1}B)u = u.$$

So we get

$$v^\top Av = (Bu)^\top A(Bu) = u^\top (B^\top AB)u.$$

This computation shows that A is positive (semi) definite if $B^\top AB$ is positive semi-definite. The same reasoning with $u = B^{-1}v$ shows that $B^\top AB$ is positive (semi) definite if A is positive (semi) definite.

Notice that it is important that $Bv = 0$ only happens when $v = 0$. □

(3.42) EXERCISE.

Let

$$D = \begin{pmatrix} d & 0 \\ 0 & e \end{pmatrix}$$

be a diagonal matrix. What conditions must the diagonal entries d and e satisfy in order for D to be positive definite?

Let

$$A = \begin{pmatrix} a & c \\ c & b \end{pmatrix}$$

denote a symmetric 2×2 matrix, where $a \neq 0$. Let

$$B = \begin{pmatrix} 1 & -\frac{c}{a} \\ 0 & 1 \end{pmatrix}.$$

Show that B is invertible and compute

$$B^\top AB.$$

Use this to show that A is positive definite if and only if $a > 0$ and $ab - c^2 > 0$.

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function defined by

$$f(x,y) = 2x^2 + 3y^2 + 4xy.$$

Show that $f(x,y) \geq 0$ for every $x, y \in \mathbb{R}$. ♠

Chapter 4

What is optimization?

In this chapter we will denote the set of column vectors with d rows by $\mathbb{R}^n$. The arithmetic of $n \times 1$ matrices apply i.e., we may add vectors in $\mathbb{R}^n$ and multiply them by a number in $\mathbb{R}$.

In the next chapter we will introduce them as *euclidean* vector spaces. The term euclidean refers to a norm: a function measuring the size of a vector. In this chapter we only need the structure as column vectors.

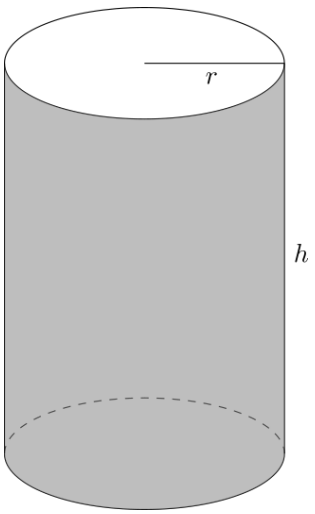
4.1 What is an optimization problem?

An optimization problem consists of maximizing or minimizing a function subject to constraints.

Below are two classical examples related to minimizing (non-linear) functions subject to (non-linear) constraints. These are actually examples of convex optimization problems. More about that later.

(4.1) EXAMPLE.

A cylindrical can is supposed to have a volume of V m^3. The material used for the top and bottom costs T DKK per m^2 and the material used for the side costs S DKK per m^2. Give the dimensions r and h of the can minimizing the price of the materials.



The cost of the top and bottom pieces are $2\pi r^2 T$. The cost of the side material is $2\pi rhS$. The constraint is that

the volume must be V. This is expressed in the equation $\pi r^2 h = V$. All in all the optimization problem is

$$\begin{array}{ll} \text{Minimize} & 2\pi r^2 T + 2\pi rhS \\ \text{with constraints} & \\ & \pi r^2 h = V \\ & r \geq 0 \\ & h \geq 0, \end{array}$$

where V, T and S are constants.

(4.2) EXERCISE.

Can you see a way of solving this optimization problem by eliminating h in the constraint $\pi r^2 h = V$?

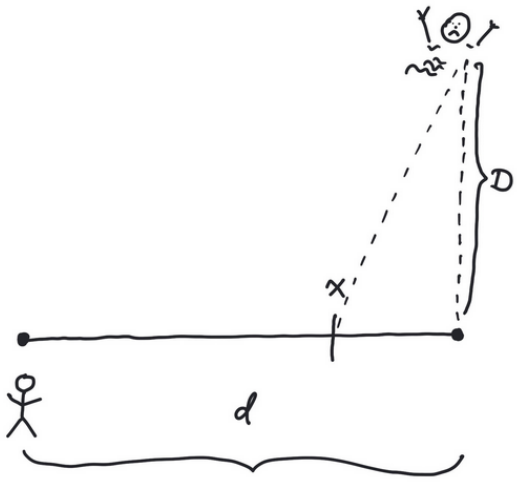
Hint:

$$\pi r^2 h = V \iff h = \frac{V}{\pi r^2}$$

and h can be inserted in $2\pi r^2 T + 2\pi rhS$. Why is this helpful? ♠ ♠

(4.3) EXAMPLE.

A person is in distress D meters from the beach. The life guard spots the situation, but is d meters from where he would naturally jump in the water as indicated below. The life guard runs 8 m/s on the beach and swims 1 m/s in the water. How far (x) should he run along the beach before jumping into the water in order to minimize the time needed to reach the person in distress?



The time spent moving with a speed of v over a distance of s is

$$t = \frac{s}{v}.$$

If the life guard jumps in the water at the point x he will have to swim a distance of

$$\sqrt{D^2 + (d-x)^2}$$

using the Pythagorean theorem. Therefore the optimization problem becomes

$$\begin{array}{ll} \text{Minimize} & \dfrac{x}{8}+\sqrt{D^2+(d-x)^2} \\ \text{with constraints} & \\ & x \geq 0 \end{array}$$

Strictly speaking we do not need the constraint $x \geq 0$, as the life guard is free to run in the other direction. So the optimization problem is simply to minimize

$$\frac{x}{8}+\sqrt{D^2+(d-x)^2}$$

with no strings attached i.e., $x \in \mathbb{R}$ is just assumed to be any number.

Interactive code not included in static version.

♠

(4.4) EXERCISE.

You need to build a rectangular fence in front of your house for a herb garden. Your house will make up one side of the rectangle, so you only need to build three sides. Suppose you have 10 m of wire. What is the maximum area of the herb garden you can wall in? ♠

4.2 General definition

An optimization problem consists of a subset $D \subseteq \mathbb{R}^n$ and a function $f : D \to \mathbb{R}$. We will consider optimization problems in the context of minimization. Optimize in this situation means minimize.

(4.5) DEFINITION.

In our most general setting an optimization problem looks like

$$\begin{array}{ll} \textit{Minimize} & f(x) \\ \textit{with constraint} & \\ & x \in C, \end{array}$$

where C and D are subsets of $\mathbb{R}^n$ with $C \subseteq D$ and $f : D \to \mathbb{R}$ is a function. A solution to the optimization problem is a vector $x_0 \in C$, such that

$$f(x_0) \leq f(x)$$

for every $x \in C$. Here x_0 is called an optimum *and $f(x_0)$ is called the* optimal value.
We will often write the optimization problem defined above in short form as

$$\begin{array}{l} \textit{min}\, f(x) \\ x \in C. \end{array}$$

(4.6) REMARK.

The complexity of the problem depends very much on the nature of C and f. Also, we cannot even be certain that an optimization problem has a solution. Consider the problem

$$\begin{aligned} &\min x \\ &x \in C, \end{aligned}$$

where $C = \{x \in \mathbb{R} \mid x \leq 0\}$.
Here x can be made arbitrarily small subject to the constraint $x \in C$ and the problem has no optimal solution.

(4.7) REMARK.

We have deliberately not included maximization problems in Definition 4.5. This is because a maximization problem, such as

$$\begin{aligned} &\text{Maximize} && f(x) \\ &\text{with constraint} \\ & && x \in C \end{aligned} \tag{4.1}$$

can be formulated as the minimization problem

$$\begin{aligned} &\text{Minimize} && -f(x) \\ &\text{with constraint} \\ & && x \in C. \end{aligned} \tag{4.2}$$

Again, we will use the short notation

$$\begin{aligned} &\max f(x) \\ &x \in C \end{aligned}$$

for the maximization problem in (4.1). A solution to (4.1) is a vector $x_0 \in C$, such that

$$f(x_0) \geq f(x)$$

for every $x \in C$. Again, x_0 is called an optimum and $f(x_0)$ the optimal value.

(4.8) EXERCISE.

Suppose that the maximization problem

$$\begin{aligned} &\max f(x) \\ &x \in C \end{aligned} \tag{4.3}$$

is formulated as the minimization problem

$$\begin{aligned} &\min -f(x) \\ &x \in C. \end{aligned} \tag{4.4}$$

Show that $f(x_0)$ is the optimal value and x_0 the optimum for (4.3) if $-f(x_0)$ is the optimal value and x_0 the optimum for (4.4). ♠

(4.9) EXERCISE.

Suppose that $a > 0$. Solve the optimization problem

$$\begin{array}{ll} \text{Minimize} & ax^2+bx+c \\ \text{with constraint} & \\ & x \in \mathbb{R}. \end{array}$$

♠

4.3 Convex optimization

Particularly well behaved optimization problems are the convex ones. These are optimization problems, where $C \subseteq \mathbb{R}^n$ is a convex subset and $f : C \to \mathbb{R}$ a convex function in Definition 4.5.

To define these concepts we first introduce the notion of a line in $\mathbb{R}^n$.

(4.10) DEFINITION.

A line $L \subseteq \mathbb{R}^n$ is a subset of the form

$$L = \{u+tv \mid t \in \mathbb{R}\},$$

where $u, v \in \mathbb{R}^n$ with $v \neq 0$.

(4.11) EXAMPLE.

A line L in the plane $\mathbb{R}^2$ is (usually) given by its equation

$$y = ax+b. \tag{4.5}$$

This means that it consists of points $(x,y) \in \mathbb{R}^2$ satisfying $y = ax+b$. Here a can be interpreted as the slope of the line and b the intersection with the y-axis.

What about all the points (x,y) with $x = 0$? Certainly they also deserve to be called a line. However, they do not satisfy an equation like (4.5). Informally, this line has infinite slope.

Therefore we introduce the parametric representation of a line: a line is the set of points of the form

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} + t \begin{pmatrix} u_0 \\ v_0 \end{pmatrix}, \tag{4.6}$$

where $t \in \mathbb{R}$,

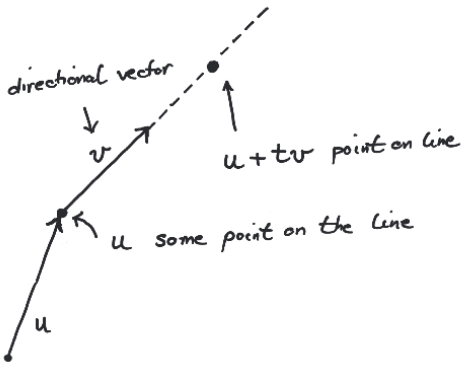
$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

is any point on the line and

$$\begin{pmatrix} u_0 \\ v_0 \end{pmatrix}$$

is a non-zero (directional) vector.

(4.12) FIGURE.



Example of a line in $\mathbb{R}^2$ with (directional) vector $v \in \mathbb{R}^2$ through the point $u \in \mathbb{R}^2$.

Given two distinct points

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} x_1 \\ y_1 \end{pmatrix},$$

there is one and only one line passing through them. This line is given by

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} + t \begin{pmatrix} x_1 - x_0 \\ y_1 - y_0 \end{pmatrix}. \tag{4.7}$$

How do we convert the line $y = ax + b$ in (4.5) to the parametric form (4.6)? Well, we know that the two distinct points

$$\begin{pmatrix} 0 \\ b \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} 1 \\ a+b \end{pmatrix}$$

are on the line. Therefore it is given by

$$\begin{pmatrix} 0 \\ b \end{pmatrix} + t \begin{pmatrix} 1 \\ a \end{pmatrix}$$

by (4.7). ♠

(4.13) EXERCISE.

Compute the parametric representation of the line L through the points $(1,1)$ and $(2,3)$. Also compute a and b in the representation $y = ax + b$ for L. ♠

(4.14) EXERCISE.

What is the parametric representation of the line consisting of the points (x,y) with $x = 0$? ♠

(4.15) EXERCISE.

Show in Definition 4.10 that if L is given by u and v, then you might as well replace v by sv, where s is a real number and $s \neq 0$. It gives the same line. ♠

(4.16) EXERCISE.

Show that there is a unique line passing through two distinct points $x, y \in \mathbb{R}^n$ and that it is given by $u = x$ and $v = y - x$ in Definition 4.10. ♠

(4.17) EXERCISE.

Do the points

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix}$$

lie on the same line in $\mathbb{R}^3$? ♠

(4.18) EXERCISE.

Show that the line through two distinct points $x, y \in \mathbb{R}^n$ is equal to the subset

$$\{(1-t)x + ty \mid t \in \mathbb{R}\} \subseteq \mathbb{R}^n.$$

♠

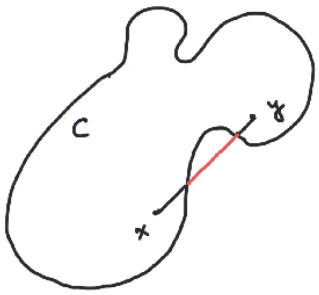
(4.19) DEFINITION.

A convex subset $C \subseteq \mathbb{R}^n$ is a subset that contains the line segment between any two of its points $x, y \in C$ i.e.,

$$(1-t)x + ty \in C$$

for every number t with $0 \leq t \leq 1$.

(4.20) FIGURE.



Example of non-convex subset of $\mathbb{R}^2$.

(4.21) QUIZ.

Quiz not included in static version. ♠

(4.22) EXERCISE.

A closed interval in $\mathbb{R}$ is a subset of the form

$$[a,b] = \{x \mid a \leq x \leq b\}$$

for $a \leq b$. Prove that $[a,b]$ is a convex subset of $\mathbb{R}$.

Hint:

Keep cool and just apply the definitions! First of all, $x \in [a,b]$ if and only if

$$a \leq x \wedge x \leq b. \tag{4.8}$$

Now pick any $t \in [0,1]$. We must show that if $x \in [a,b]$ and $y \in [a,b]$, then

$$(1-t)x + ty \in [a,b].$$

You may also write this out as

$$a \leq x \wedge x \leq b \qquad \wedge \qquad a \leq y \wedge y \leq b$$

implies that

$$a \leq (1-t)x + ty \qquad \wedge \qquad (1-t)x + ty \leq b.$$

Hint:

$$a \leq x \implies (1-t)a \leq (1-t)x \qquad \wedge \qquad a \leq y \implies ta \leq ty$$

implies that

$$(1-t)a + ta \leq (1-t)x + ty.$$

What is $(1-t)a + ta$?

♠

(4.23) EXERCISE.

Let A and B be convex subsets of $\mathbb{R}^n$. Prove that $A \cap B$ is a convex subset of $\mathbb{R}^n$. Generalize this to show that if $A_1, \ldots, A_n$ are any number of convex subsets of $\mathbb{R}^n$, then their intersection

$$A_1 \cap \cdots \cap A_n$$

is a convex subset of $\mathbb{R}^n$. Is the union of two convex subsets necessarily convex? ♠

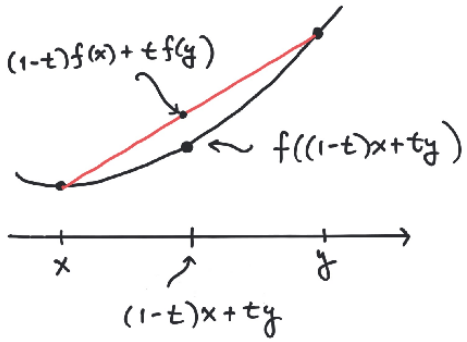
(4.24) DEFINITION.

A convex function is a function $f : C \to \mathbb{R}$ defined on a convex subset $C \subseteq \mathbb{R}^n$, such that

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$$

for every number t with $0 \leq t \leq 1$.

(4.25) FIGURE.



Graph of convex function. The line segment between $(x, f(x))$ and $(y, f(y))$ lies above the graph.

(4.26) EXERCISE.

i) Let the function $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = ax + b$, where $a, b \in \mathbb{R}$. Show that f is a convex function.

Hint: Try the case $a = 0$ first.

ii) Can you at this point prove that $f(x) = x^2$ is a convex function?

Hint: Simplify

$$(1-t)x^2 + ty^2 - ((1-t)x + ty)^2$$

to an expression that has to be non-negative.

Hint:

Interactive code not included in static version.

iii) Using that $f(x) = x^2$ is a convex function, prove that $g(x) = x^4$ is a convex function.

Hint: Use that $g(x) = f(x)^2$ and $a \leq b \implies a^2 \leq b^2$ if $a, b \geq 0$ (here we really need $a, b \geq 0$, since for example $-2 \leq -1$, but $4 \leq 1$ is not true) to conclude that

$$((1-t)x + ty)^4 = (((1-t)x + ty)^2)^2 \leq ((1-t)x^2 + ty^2))^2 \leq (1-t)x^4 + ty^4$$

for $x, y \in \mathbb{R}$.

iv) It is a fact that $f(x) = x^3$ is not a convex function, but can you explain this using the definition of a convex function?

Hint: Try $x = -1, y = 0$ and $t = \frac{1}{2}$.

♠

(4.27) LEMMA.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then the subset

$$C = \{x \in \mathbb{R}^n \mid f(x) \leq a\}$$

is a convex subset of $\mathbb{R}^n$, where $a \in \mathbb{R}$.

Proof. Suppose that $u, v \in C$ and $t \in [0,1]$. Looking at Definition 4.19 we must prove that

$$(1-t)u + tv \in C.$$

By the definition of f being convex (Definition 4.24), it follows that

$$f((1-t)u + tv) \leq (1-t)f(u) + tf(v).$$

But, since $f(u) \leq a$ and $f(v) \leq a$ we have

$$\begin{aligned} (1-t)f(u) &\leq (1-t)a \\ tf(v) &\leq ta \end{aligned}$$

and therefore

$$(1-t)f(u) + tf(v) \leq (1-t)a + ta = a.$$

Therefore,

$$f((1-t)u + tv) \leq a$$

and $(1-t)u + tv \in C$. □

Why are convex optimization problems interesting? We will return to this in the chapter on convex functions. As a sneak preview let me comment already now.

(4.28) REMARK.

In hunting for optimal solutions to an optimization problem one is often stuck with a point $x_0 \in \mathbb{R}^n$, which is optimal locally. This means that $f(x_0) \leq f(x)$ for every x that is sufficiently close to x_0 (we will explain what this means in the next chapter). The remarkable thing that happens in a convex optimization problem is that if x_0 is optimal locally, then it is a global optimum! It satisfies $f(x_0) \leq f(x)$ not only for x close to x_0, but for every $x \in C$.

The optimization problem in Exercise 4.9 is a very typical convex optimization problem.

(4.29) EXAMPLE.

Below you see a plot of the function (press Compute)

$$f(x) = x^3 + 2x^2 + x + 1$$

restricted to the interval $[-1.5, 0]$. You can see that it has a local minimum around -0.3 and also that this minimum is not a global minimum (certainly $f(-1.4)$ is smaller). So $f(x)$ is not a convex function on this

interval according to Remark 4.28(but if you look at it more locally on the interval $[-0.6, 0]$ it is a convex function).

Interactive code not included in static version.

♠

(4.30) EXERCISE.

Solve the optimization problem

$$\begin{array}{ll} \text{Minimize} & x^3+2x^2+x+1 \\ \text{with constraint} & \\ & x \in C \end{array}$$

for $C = [-0.6, 0]$ and $C = [-2, 0]$. ♠

4.4 Linear optimization

We will start this section with a concrete example.

A company produces two products A and B. The product A is selling for 350 USD and B is selling for 300 USD. There are certain limited ressources in the production of A and B. Two raw materials S_1 and S_2 are needed along with employee work time. The production of A requires 18 minutes, one unit of S_1 and six units of S_2. The production of B requires 12 minutes, one unit of S_1 and eight units of S_2. There are 3132 minutes of employee work time, 200 units of S_1 and 1440 units of S_2 available. These constraints in the production can be outlined in the diagram below

	minutes	S_1	S_2
A	18	1	6
B	12	1	8
constraint	3132	200	1440

How many units x of A and y of B should the company produce to maximize its profit?

You can rewrite this as the optimization problem

$$\begin{array}{ll} \text{Maximize} & 350x+300y \\ \text{with constraints} & \\ & 18x+12y \leq 3132 \\ & x+y \leq 200 \\ & 6x+8y \leq 1440 \\ & x \geq 0 \\ & y \geq 0 \end{array}$$

This optimization problem is a special case of linear optimization, which arguably is one of the most succesful applications of mathematics (after the introduction of the simplex algorithm following World War II). We will give a taste of the mathematical setup here.

The simplest convex optimization problems are the linear ones. Recall that a linear function $f : \mathbb{R}^n \to \mathbb{R}$ has the form

$$f\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = c_1 x_1 + \cdots + c_n x_n$$

for $c_1, \ldots, c_n \in \mathbb{R}$. Usually we write this with matrix notation as

$$f(x) = c^\top x,$$

where

$$c = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} \qquad \text{and} \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

(4.31) EXERCISE.

Show that a linear function is convex. ♠

A linear optimization problem is not about minimizing a linear function over an arbitrary convex subset. We choose the convex subset as an intersection of subsets of the form

$$\{x \in \mathbb{R}^n \mid a^\top x \leq b\},$$

where $a \in \mathbb{R}^n$ is a non-zero vector and $b \in \mathbb{R}$ a number i.e., a linear optimization problem has the form

$$\begin{aligned} &\min c^\top x \\ &x \in C, \end{aligned}$$

where

$$\begin{aligned} C &= \{x \in \mathbb{R}^n \mid a_1^\top x \leq b_1, \ldots, a_m^\top x \leq b_m\} \\ &= \{x \in \mathbb{R}^n \mid a_1^\top x \leq b_1\} \cap \cdots \cap \{x \in \mathbb{R}^n \mid a_m^\top x \leq b_m\} \end{aligned} \tag{4.9}$$

and $c, a_1, \ldots, a_m \in \mathbb{R}^n$ and $b_1, \ldots, b_m \in \mathbb{R}$.

(4.32) EXERCISE.

Use a selection of previous exercises to show that the subset C defined in (4.9) is a convex subset of $\mathbb{R}^n$. ♠

Using matrix notation we write C as

$$C = \{x \in \mathbb{R}^n \mid Ax \leq b\},$$

where A is the $m \times d$ matrix with row vectors $a_1^\top, \ldots, a_m^\top$ and

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}.$$

(4.33) EXAMPLE.

Here is a concrete example for $d = 2$. The optimization problem

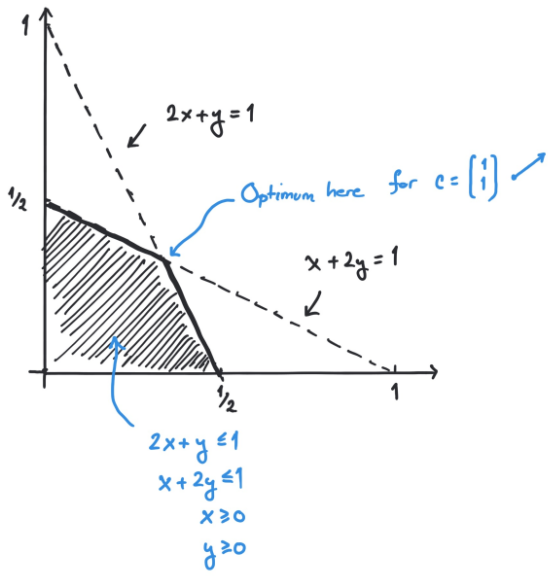
$$\begin{aligned} &\text{Maximize} && x+y \\ &\text{with constraints} && \\ & && 2x+y \leq 1 \\ & && x+2y \leq 1 \\ & && x \geq 0 \\ & && y \geq 0 \end{aligned}$$

translates into matrix notation with the matrices

$$c = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \\ -1 & 0 \\ 0 & -1 \end{pmatrix} \qquad \text{and} \qquad b = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

In this case it is helpful to draw the optimization problem in the plane $\mathbb{R}^2$.

This is done below.



Constraints pictured as shaded area above. Optimum occurs in a vertex (corner).

♠

We will give a general (but rather slow) algorithm below for solving linear optimization problems. In fact it all boils down to solving systems of linear inequalities. Sometimes linear optimization is referred to as linear programming. The basic theory of linear programming was pioneered, among others, by one of the inventors of the modern computer, John von Neumann.

John von Neumann (1903-1957). Picture from LANL.

Sage has much more advanced algorithms built in for solving (integer) linear optimization problems. I have translated the linear optimization problem in Example 4.33 into Sage below.

(4.34) EXAMPLE.

Interactive code not included in static version.

♠

4.5 Fourier-Motzkin elimination

Fourier-Motzkin elimination is a classical method (dating back to 1826) for solving linear inequalities. It is also a key ingredient in an algorithm for solving linear optimization problems.

I am convinced that the best way to explain this method is by way of an extended example. For more formalities you may consult Chapter 1 of my book Undergraduate Convexity.

Consider the linear optimization problem

$$\begin{array}{lll} \text{Maximize} & x+y & \\ \text{with constraints} & & \\ & \begin{array}{rcrcl} 2x & + & y & \leq & 6 \\ x & + & 2y & \leq & 6 \\ x & + & 2y & \geq & 2 \\ x & & & \geq & 0 \\ & & y & \geq & 0. \end{array} & \end{array} \tag{4.10}$$

We might as well write this as

$$\begin{array}{lll} \text{Maximize} & z & \\ \text{with constraints} & & \\ & \begin{array}{rcrcl} & & z & = & x+y \\ 2x & + & y & \leq & 6 \\ x & + & 2y & \leq & 6 \\ x & + & 2y & \geq & 2 \\ x & & & \geq & 0 \\ & & y & \geq & 0 \end{array} & \end{array}$$

by adding the extra variable z. This enables us to reformulate the problem as follows: Find the maximal value of z, such that there exists $(x,y) \in \mathbb{R}^2$ with

$$(x,y,z) \in P,$$

where $P \subseteq \mathbb{R}^3$ is the set of solutions to the system

$$\begin{array}{rcrcl} & & z & = & x+y \\ 2x & + & y & \leq & 6 \\ x & + & 2y & \leq & 6 \\ x & + & 2y & \geq & 2 \\ x & & & \geq & 0 \\ & & y & \geq & 0 \end{array} \tag{4.11}$$

of inequalities[1].

We have the Gauss elimination method for solving systems of linear equations. How do we now solve (4.11), where we also have inequalities?

Well, at first we can actually do a Gauss elimination step by eliminating x in the equation $z = x+y$ i.e., by putting $x = z-y$. This is then inserted into the inequalities in (4.11):

$$\begin{array}{rcrcl} 2(z-y) & + & y & \leq & 6 \\ (z-y) & + & 2y & \leq & 6 \\ (z-y) & + & 2y & \geq & 2 \\ (z-y) & & & \geq & 0 \\ & & y & \geq & 0 \end{array}$$

and we get the system

$$\begin{array}{rcrcl} 2z & - & y & \leq & 6 \\ z & + & y & \leq & 6 \\ z & + & y & \geq & 2 \\ z & - & y & \geq & 0 \\ & & y & \geq & 0 \end{array}$$

of inequalities in the variables z and y. Now we only have inequalities left and we have to invent a trick for eliminating y. Let us isolate y on one side of the inequality signs $\leq$ and $\geq$:

$$\begin{array}{rcrcl} 2z & - & 6 & \leq & y \\ 6 & - & z & \geq & y \\ 2 & - & z & \leq & y \\ & & z & \geq & y \\ & & 0 & \leq & y \end{array}$$

Written a little differently this is the same as

$$\begin{array}{rcrcl} 2z & - & 6 & \leq & y \\ & & & & y \leq 6 - z \\ 2 & - & z & \leq & y \\ & & & & y \leq z \\ & & 0 & \leq & y \end{array} \tag{4.12}$$

Now the scene is set for elimination of y. Listen carefully. First the inequalities in (4.12) can be boiled down to the following two inequalities

[1] An equality $a = b$ is logically equivalent to the two inequalities $a \leq b$ and $a \geq b$ in the sense that $(a \leq b) \wedge (a \geq b) \iff a = b$.

$$\begin{array}{rcl} \max(2z-6,2-z,0) & \leq & y \\ & & y \leq \min(6-z,z) \end{array} \tag{4.13}$$

by using (repeatedly) that $\max(a,b) \leq c \iff a \leq c \wedge b \leq c$ and $c \leq \min(a,b) \iff c \leq a \wedge c \leq b$ for three numbers $a,b,c \in \mathbb{R}$.

Then, finally comes the (Fourier-Motzkin) elimination step: The existence of a solution to (4.13) is equivalent to the single inequality

$$\max(2z-6,2-z,0) \leq \min(6-z,z). \tag{4.14}$$

This single inequality can be exploded or expanded (see Exercise 4.36) into the following $6 = 3 \cdot 2$ inequalities

$$\begin{aligned} 2z-6 &\leq 6-z \\ 2z-6 &\leq z \\ 2-z &\leq 6-z \\ 2-z &\leq z \\ 0 &\leq 6-z \\ 0 &\leq z. \end{aligned}$$

Similarly to (4.12) we now isolate z from the above inequalities:

$$\begin{array}{rcl} & z \leq & 4 \\ & z \leq & 6 \\ 1 \leq & z & \\ & z \leq & 6 \\ 0 \leq & z & \end{array}$$

and find that

$$\begin{array}{rcl} \max(1,0) = 1 & \leq & z \\ & z \leq & \min(4,6) = 4. \end{array}$$

Therefore the maximum in the optimization problem (4.10) is $z = x+y = 4$. How do we now find numbers $x,y \in \mathbb{R}$ satisfying the constraints in the optimization problem (4.10) with $z = x+y = 4$?

This is simply done inserting first $z = 4$ in (4.13). Here you get the two inequalities $2 \leq y$ and $y \leq 2$. Therefore $y = 2$. Since we had $x = z-y$ from the very beginning we therefore get $x = 2$ and we have the unique solution to the optimization problem.

(4.35) EXERCISE.

What is the solution if we replace Maximize with Minimize in the optimization problem (4.10)? ♠

(4.36) EXERCISE.

Prove the following:

Let $x_1,\ldots,x_m,y_1,\ldots,y_n \in \mathbb{R}$ be $m+n$ numbers. Then

$$\max(x_1,\ldots,x_m) \leq \min(y_1,\ldots,y_n)$$

if and only if the mn inequalities

$$\begin{array}{cccc} x_1 \leq y_1 & x_1 \leq y_2 & \dots & x_1 \leq y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m \leq y_1 & x_m \leq y_2 & \dots & x_m \leq y_n \end{array}$$

are satisfied. ♠

(4.37) EXERCISE.

The following is Exercise 1.8 from my book Undergraduate Convexity.

A vitamin pill P is produced using two ingredients M_1 and M_2. The pill needs to satisfy four constraints for the vital vitamins V_1 and V_2. It must contain at least 6 milligrams and at most 15 milligrams of V_1 and at least 5 milligrams and at most 12 milligrams of V_2. The ingredient M_1 contains 3 milligrams of V_1 and 2 milligrams of V_2 per gram. The ingredient M_2 contains 2 milligrams of V_1 and 3 milligrams of V_2 per gram:

	V_1	V_2
M_1	3	2
M_2	2	3

Let x denote the amount of M_1 and y the amount of M_2 (measured in grams) in the production of a vitamin pill. Write down a system of linear inequalities in x and y describing the constraints above.

We want a vitamin pill of minimal weight satisfying the constraints. How many grams of M_1 and M_2 should we mix?

Use Fourier-Motzkin elimination to solve this problem.

Check your solution by modifying the input to the Sage code in Example 4.34 using Remark 4.7.

One may also force minimization by inserting the following option

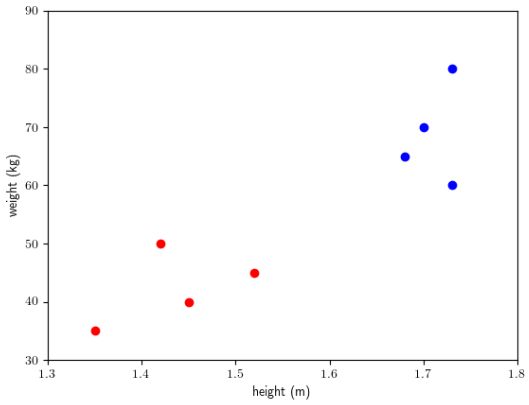
```
LP = MixedIntegerLinearProgram(maximization=False, solver = "GLPK").
```

in Example 4.34. ♠

4.6 Application in machine learning and data science

To start with, consider a toy example of a machine learning problem: we wish to tell the gender of a person based on a data point consisting of the height and weight of the person.

To do this we train our model by measuring the height and weight of a lot of people. Each of these measured data points are labeled female or male according to the gender of the person.



Given a new data point, we wish to tell if the person is female or male. Here we consider a very simple model for doing this. First we need to introduce some new mathematical terms. We will introduce the terms generally for data points in $\mathbb{R}^n$ and not just in $\mathbb{R}^2$ as above.

(4.38) DEFINITION.

A hyperplane in $\mathbb{R}^n$ is defined as a subset

$$H = \{v \in \mathbb{R}^n \mid a^\top v + b = 0\}$$

where $a \in \mathbb{R}^n$ is a non-zero vector (called a normal vector of H) and b is a number. A hyperplane divides $\mathbb{R}^n$ into two subsets: the points above the hyperplane satisfying $a^\top v + b > 0$ and the ones below the hyperplane satisfying $a^\top v + b < 0$.

(4.39) EXAMPLE.

In $\mathbb{R}^2$ a hyperplane is a line. The line $y = 2x + 1$ is the hyperplane

$$H = \{v \in \mathbb{R}^2 \mid a^\top v + b = 0\},$$

where

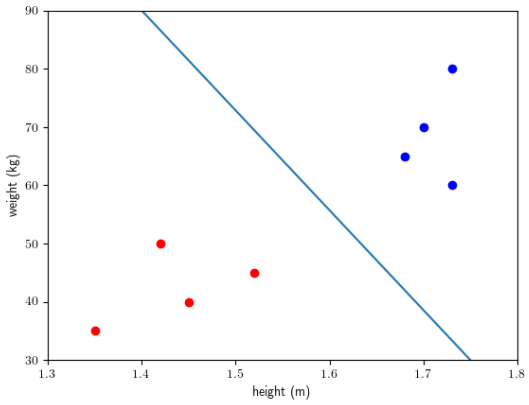
$$a = \begin{pmatrix} 2 \\ -1 \end{pmatrix} \qquad \text{and} \qquad b = 1.$$

The points $(0,1)$ and $(1,3)$ are on the hyperplane, while the point $(0,0)$ here is below the hyperplane. Notice however that above and below depend on the choice of a normal vector a. We might as well have picked

$$a = \begin{pmatrix} -2 \\ 1 \end{pmatrix} \qquad \text{and} \qquad b = -1$$

and then the point $(0,0)$ would have been above the hyperplane. ♠

Suppose we are given a data set as a finite set of points in $\mathbb{R}^n$ and that each of these points are labeled with either a blue or a red color. We wish to find a hyperplane, such that the blue points are above and the red points are below the hyperplane.



We may then use the hyperplane to predict the label of a point. This could be gender, if you win or lose money buying a stock, anything with a binary classifier.

4.6.1 Formulation as a linear optimization problem

Suppose that the points labeled blue are $x_1,\ldots,x_m \in \mathbb{R}^n$ and the points labeled red are $y_1,\ldots,y_n \in \mathbb{R}^n$. Then we wish to find $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, such that

$$a^\top x_i > b$$

for $i = 1,\ldots,m$ and

$$a^\top y_j < b$$

for $j = 1,\ldots,n$. One can show that these strict inequalities may be solved for a and b if and only if the inequalities

$$\begin{aligned} a^\top x_i &\geq b+1 \\ a^\top y_j &\leq b-1 \end{aligned}$$

are solvable for a and b, where $i = 1,\ldots,m$ and $j = 1,\ldots,n$.

It is, however, not realistic to expect data to behave this nicely. Instead one invents the rather ingenious linear optimization problem

$$\begin{aligned} &\text{Minimize} && \tfrac{1}{m}(u_1+\cdots+u_m)+\tfrac{1}{n}(v_1+\cdots+v_n) \\ &\text{with constraints} && \\ & && a^T x_i + u_i \geq b+1 \\ & && a^T y_j - v_j \leq b-1 \\ & && u_i \geq 0 \\ & && v_j \geq 0 \end{aligned} \tag{4.15}$$

for $i = 1,\ldots,m$ and $j = 1,\ldots,n$. This linear optimization problem has optimal value zero if and only if data can be separated strictly. Otherwise, it finds a hyperplane minimizing the mean errors for the points involved.

The linear optimization problem (4.15) may look untied to the real world, but it has been used very successfully in the diagnosis and prognosis of breast cancer. See Mangasarian et al.

In the sage window below we have implemented the solution of the linear optimization problem (4.15), where the output is a graphical illustration of the optimal line, that separates the points in `xpts` and `ypts` with the smallest mean error as defined in the function to be minimized in (4.15).

Interactive code not included in static version.

Chapter 5

Euclidean vector spaces

Big data are made up of many numbers in data sets. Such data sets can be represented as vectors in a high dimensional euclidean vector space. A vector is nothing but a list of numbers, but we need to talk mathematically about the size of a vector and perform operations on vectors. The term euclidean refers to vectors with a dot product as known from the plane $\mathbb{R}^2$.

The purpose of this chapter is to set the stage for this, especially by introducing the dot product (or inner product) for general vectors. Having a dot product is immensely useful and we give several applications like linear regression and the perceptron learning algorithm

In the last part of the chapter we will list rudimentary basics of analysis starting with bounded, open, closed and compact subsets of euclidean spaces leading to continuous functions and the socalled extreme value theorem, Theorem 5.66. This result states that a huge class of optimization problems always have a solution.

5.1 Vectors in the plane

The dot product (or inner product) between two vectors $u, v \in \mathbb{R}^2$ is given by

$$u \cdot v = x_1 x_2 + y_1 y_2, \tag{5.1}$$

where

$$u = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \qquad \text{and} \qquad v = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}. \tag{5.2}$$

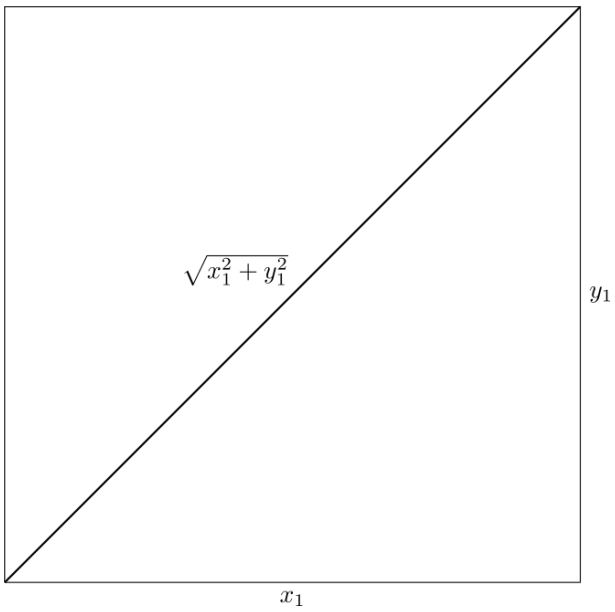
We may also interpret u and v as 2×1 matrices (or column vectors). Then the dot product in (5.1) may be realized as the matrix product:

$$u \cdot v = u^\intercal v.$$

The length or *norm* of the vector $u \in \mathbb{R}^2$ is given by

$$|u| = \sqrt{u \cdot u} = \sqrt{u^\intercal u} = \sqrt{x_1^2 + y_1^2}.$$

This follows from the Pythagorean theorem:



The distance $d(u,v)$ between the two vectors u and v is given by

$$d(u,v) = |u-v| = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$$

Also, the cosine of the angle θ between u and v is given by

$$\cos(\theta) = \frac{u\cdot v}{|u|\,|v|} \qquad \text{or} \qquad u\cdot v = |u|\,|v|\cos(\theta).$$

We will not go into this formula. It is a byproduct of considering the projection of a vector on another vector (see Exercise 5.6).

All of these rather natural notions in the plane $\mathbb{R}^2$ generalize naturally to $\mathbb{R}^n$ for $n > 2$.

5.2 Higher dimensions

We denote the set of column vectors with n rows by $\mathbb{R}^n$ and call it the euclidean vector space of dimension n. An element $v \in \mathbb{R}^n$ is called a vector and it has the form (column vector with d entries)

$$v = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

A vector in $\mathbb{R}^n$ is a model for a data set in real life. A collection of d numbers, which could signify d measurements. You will see an example of this below, where a vector represents a data set counting words in a string.

Being column vectors, vectors in $\mathbb{R}^n$ can be added and multiplied by numbers:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1+y_1 \\ x_2+y_2 \\ \vdots \\ x_n+y_n \end{pmatrix} \qquad \lambda \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \lambda x_1 \\ \lambda x_2 \\ \vdots \\ \lambda x_n \end{pmatrix}.$$

The dot product generalizes as follows to higher dimensions.

5.2.1 Dot product, norm and cosine

(5.1) DEFINITION.

Suppose that

$$u = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \text{and} \qquad v = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

are vectors in $\mathbb{R}^n$.

1. *The dot product between u and v is defined by*

$$u \cdot v = u^\top v = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n. \tag{5.3}$$

2. *Two vectors* $u, v \in \mathbb{R}^n$ *are called* orthogonal *if* $u \cdot v = 0$. *We write this as* $u \perp v$.

3. *The norm of* $u \in \mathbb{R}^n$ *is defined by*

$$|u| = \sqrt{u \cdot u} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}. \tag{5.4}$$

4. *The distance between the two vectors u and v is defined by*

$$d(u,v) = |u - v| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

5. *The cosine of the angle between u and v is defined by*

$$\frac{u \cdot v}{|u||v|}$$

provided that they both are non-zero.

All of the definitions above are present in modern machine learning frameworks. Below we see their incarnations in the python library `numpy`.

(5.2) EXAMPLE.

Interactive code not included in static version.

♠

(5.3) EXERCISE.

Show that

$$u \perp u \iff u = 0,$$

where $u \in \mathbb{R}^n$. ♠

(5.4) EXERCISE.

Use the definition in (5.3) to show that

$$u \cdot (v+w) = u \cdot v + u \cdot w$$
$$(\lambda u) \cdot v = u \cdot (\lambda v) = \lambda (u \cdot v)$$

for $u, v, w \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$. ♠

(5.5) EXERCISE.

Let $u \in \mathbb{R}^n$ be a nonzero vector and $\lambda \in \mathbb{R}$. Use the definition in (5.4) to show that $|\lambda u| = |\lambda|\,|u|$ and that

$$\frac{1}{|u|} u$$

is a unit vector.

Hint: You could perhaps use Exercise 5.4 to do this. Notice also that $|\lambda|$ is the absolute value for λ if $\lambda \in \mathbb{R}$. ♠

(5.6) EXERCISE.

Given two vectors $u, v \in \mathbb{R}^n$ with $v \neq 0$, find $\lambda \in \mathbb{R}$, such that $u - \lambda v$ and v are orthogonal, i.e.

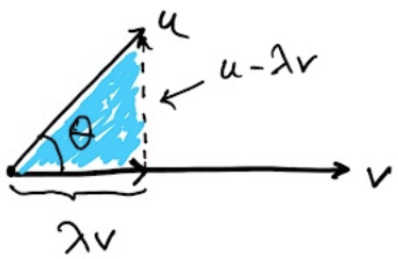
$$(u - \lambda v) \cdot v = 0.$$

Hint:

$$(u - \lambda v) \cdot v = 0 \iff (u \cdot v) - \lambda (v \cdot v) = 0.$$

This is an equation, where λ is unknown!

For $d = 2$, it is sketched below that if $u - \lambda v$ and v are orthogonal, then $u, \lambda v$ and $u - \lambda v$ are the sides in a right triangle.



In this case, if θ is the angle between u and v, show that

$$|u| \cos(\theta) = |v|\,\lambda.$$

Use this to show that

$$u \cdot v = |u|\,|v| \cos(\theta).$$

Finally show that

$$\cos(A-B) = \cos(A)\cos(B) + \sin(A)\sin(B),$$

where A and B are two angles.

Hint: In the last question, you could use that the vectors

$$\begin{pmatrix}\cos(A)\\ \sin(A)\end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix}\cos(B)\\ \sin(B)\end{pmatrix}$$

are unit vectors. ♠

(5.7) EXERCISE.

Given two vectors $u, v \in \mathbb{R}^n$, solve the minimization problem

$$\begin{array}{ll} \text{Minimize} & |u-\lambda v| \\ \text{with constraint} & \\ & \lambda \in \mathbb{R}. \end{array}$$

Hint: First convince yourself that λ minimizes $|u-\lambda v|$ if and only if it minimizes

$$(u-\lambda v)\cdot(u-\lambda v) = |v|^2\lambda^2 - 2(u\cdot v)\lambda + |u|^2,$$

which happens to be a quadratic polynomial in λ. ♠

5.3 The unreasonable effectiveness of the dot product

(5.8) QUIZ.

Quiz not included in static version. ♠

5.3.1 The dist formula from high school

The infamous dist formula from high school says that the distance from the point (x_1, y_1) to the line given by $y = ax + b$ is

$$\frac{|ax_1 + b - y_1|}{\sqrt{a^2+1}}. \tag{5.5}$$

Where does this magical formula come from? Consider a general line L in parametrized form (see Definition 4.10)

$$L = \{u + tv \mid t \in \mathbb{R}\} \subseteq \mathbb{R}^n.$$

If $w \in \mathbb{R}^n$, then the distance from w to L is given by the solution t_0 to the optimization problem

$$\begin{array}{l} \min (w-(u+tv))\cdot(w-(u+tv)) \\ t \in \mathbb{R}. \end{array} \tag{5.6}$$

This looks scary, but simply boils down to finding the top of a parabola. The solution is

$$t_0 = \frac{v \cdot w - u \cdot v}{v \cdot v}$$

and the point on L closest to w is $u + t_0 v$.

Now we put (see Example 4.11)

$$u = \begin{pmatrix} 0 \\ b \end{pmatrix}, \quad v = \begin{pmatrix} 1 \\ a \end{pmatrix} \quad \text{and} \quad w = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$$

in order to derive (5.5). The solution to (5.6) becomes

$$t_0 = \frac{x_1 + a y_1 - ab}{1 + a^2}.$$

We must compute the distance D from w to $u + t_0 v$ in this case. The distance squared is

$$\begin{aligned} D^2 &= |w - (u + t_0 v)|^2 = (w - (u + t_0 v)) \cdot (w - (u + t_0 v)) \\ &= (x_1 - t_0)^2 + (y_1 - b - t_0 a)^2. \end{aligned}$$

This is a mouthful and I have to admit that I used symbolic software (see below) to verify that

$$D^2 = \frac{a^2 x_1^2 + 2abx_1 - 2ax_1 y_1 + b^2 - 2by_1 + y_1^2}{1 + a^2} = \frac{(ax_1 + b - y_1)^2}{1 + a^2}.$$

Interactive code not included in static version.

5.3.2 The perceptron algorithm

Already at this point we have the necessary definitions for explaining the perceptron algorithm. This is one of the early algorithms of machine learning. It aims at finding a high dimensional line (hyperplane) that separates data organized in two clusters. In terms of the dot product, the idea of the algorithm is described below in dimension two.

A line in the plane is given by an equation

$$ax + by + c = 0$$

for $a, b, c \in \mathbb{R}$, where $(a, b) \neq (0, 0)$. Given finitely many points

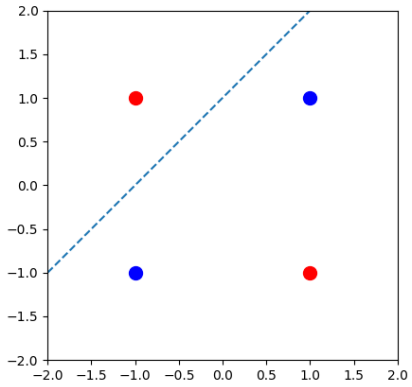
$$v_1 = (x_1, y_1), v_2 = (x_2, y_2), \ldots, v_n = (x_n, y_n)$$

each with a label $\ell_1, \ldots, \ell_n$ of ± 1 (or blue and red for that matter), we wish to find a line (given by a, b, c), such that

$$\begin{aligned} ax_i + by_i + c > 0 \quad &\text{if} \quad \ell_i = 1 \\ ax_i + by_i + c < 0 \quad &\text{if} \quad \ell_i = -1 \end{aligned}$$

for $i = 1, \ldots, n$. Such a line is called a separating line for the labeled points.

In some cases this is impossible (an example is illustrated below).



(5.9) EXERCISE.

Show that it is imposible to find a line separating the red and blue points above. The red points are $(-1,1)$ and $(1,-1)$. The blue points are $(-1,-1)$ and $(1,1)$. ♠

A clever approach to finding such a line, if it exists, is to reformulate the problem by looking at the vectors given by

$$\hat{v}_1 = (\ell_1 x_1, \ell_1 y_1, \ell_1), \qquad \hat{v}_2 = (\ell_2 x_2, \ell_2 y_2, \ell_2), \qquad \dots \qquad , \qquad \hat{v}_n = (\ell_n x_n, \ell_n y_n, \ell_n) \tag{5.7}$$

in $\mathbb{R}^3$. Then the existence of the line is equivalent to the existence of a vector $\alpha \in \mathbb{R}^3$ with $\alpha \cdot \hat{v}_i > 0$ for $i = 1, \dots, n$. If $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ is such a vector, then we have for $i = 1, \dots, n$,

$$\begin{aligned} \alpha_1 x_i + \alpha_2 y_i + \alpha_3 > 0 \quad &\text{if} \quad \ell_i = 1 \\ -\alpha_1 x_i - \alpha_2 y_i - \alpha_3 > 0 \quad &\text{if} \quad \ell_i = -1. \end{aligned} \tag{5.8}$$

Therefore we may take $a = \alpha_1, b = \alpha_2$ and $c = \alpha_3$ as the line.

A ridiculously simple algorithm

In view of the approach introduced in (5.8), the the following general question is interesting.

> Given finitely many vectors $v_1, \dots, v_m \in \mathbb{R}^n \setminus \{0\}$, can we find $\alpha \in \mathbb{R}^n$, such that
>
> $$\alpha \cdot v_i > 0$$
>
> for every $i = 1, \dots, m$?

(5.10) EXERCISE.

Come up with a simple example, where this problem is unsolvable i.e., come up with vectors $v_1, \dots, v_n \in \mathbb{R}^n$, where such an α does not exist.

Hint: Try out some simple examples for $d = 1$ and $d = 2$. ♠

In case α exists, the following ridiculously simple algorithm works in computing α. It is called the *perceptron (learning) algorithm*.

(i) Begin by putting $\alpha = 0$.

(ii) If there exists $v_i \in \{v_1, \ldots, v_m\}$ with $\alpha \cdot v_i \leq 0$, then replace α by $\alpha + v_i$ and repeat this step. Otherwise α is the desired output vector.

(5.11) EXAMPLE.

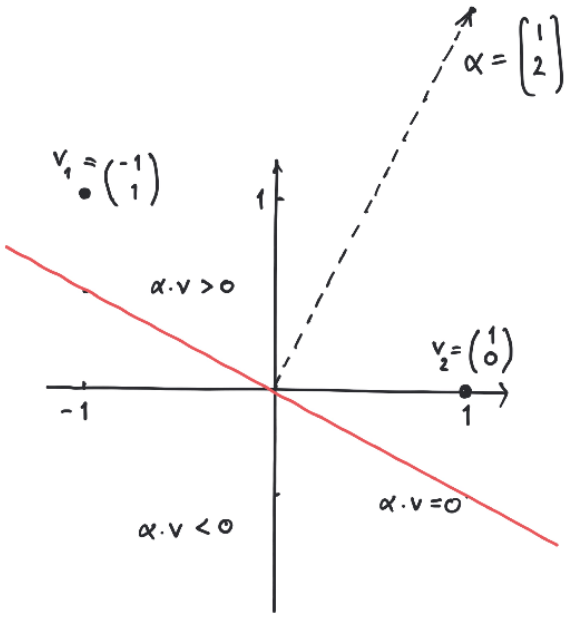
Let us try out the algorithm on the simple example of just two points in $\mathbb{R}^2$ given by

$$v_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \qquad \text{and} \qquad v_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

In this case the algorithm proceeds as pictured below.

$$\alpha = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \xrightarrow{+v_1} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \xrightarrow{+v_2} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \xrightarrow{+v_2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \xrightarrow{+v_1} \begin{pmatrix} 0 \\ 2 \end{pmatrix} \xrightarrow{+v_2} \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

It patiently crawls its way ending with the vector $\alpha = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, which satisfies $\alpha \cdot v_1 > 0$ and $\alpha \cdot v_2 > 0$.



♠

Let us see how (5.7) works in a concrete example.

(5.12) EXAMPLE.

Consider the points

$$v_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad v_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \text{and} \qquad v_3 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

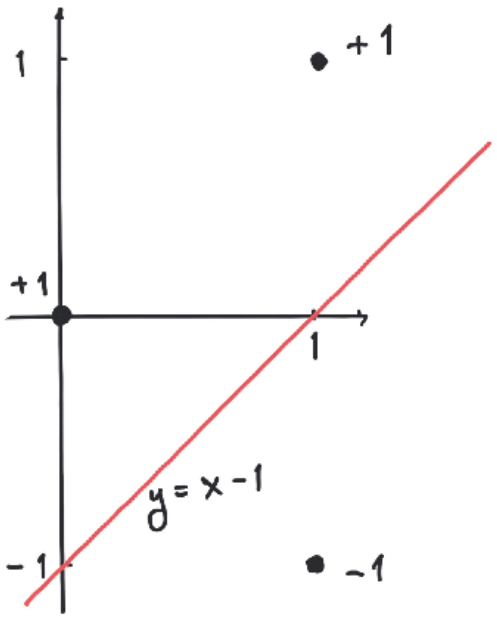
in $\mathbb{R}^2$, where v_1 and v_2 are labeled by $+1$ and v_3 is labeled by -1. Then we let

$$\hat{v}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \qquad \hat{v}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad \text{and} \qquad \hat{v}_3 = \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}.$$

Now we run the simple algorithm above Example 5.11:

$$\hat{\alpha} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \xrightarrow{+\hat{v}_1} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \xrightarrow{+\hat{v}_3} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \xrightarrow{+\hat{v}_1} \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}.$$

From the last vector we see that $x - y - 1 = 0$ determines a line separating the labeled points.



♠

Below is an implementation of the perceptron (learning) algorithm in python (with numpy) with input from Example 5.12 (it also works in higher dimensions).

Interactive code not included in static version.

(5.13) EXERCISE.

Consider the points

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

in $\mathbb{R}^2$, where the first point is labeled with -1 and the rest by 1. Use the perceptron algorithm to compute a separating hyperplane.

What happens when you run the perceptron algorithm on the above points, but where the label of

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

is changed from 1 to -1?

♠

5.3.3 Why does the perceptron algorithm work?

We will assume that there exists $\alpha \in \mathbb{R}^n$, such that

$$\alpha \cdot v_i > 0$$

for every $i = 1, \ldots, m$. Therefore $\mu = \min(\alpha \cdot v_1, \ldots, \alpha \cdot v_m) > 0$ and if we put

$$\alpha^* = \frac{1}{\mu}\alpha, \tag{5.9}$$

then $\alpha^* \cdot v_i \geq 1$ for every $i = 1, \ldots, m$.

The basic insight is the following

(5.14) PROPOSITION.

Let $r = \max\{|v_1|, \ldots, |v_m|\}$. After k iterations of the perceptron algorithm, α satisfies

$$\alpha \cdot \alpha^* \geq k \qquad \text{and} \qquad kr^2 \geq |\alpha|^2,$$

where α^* is defined in (5.9).

Proof. The algorithm starts with $\alpha = 0$. In the second step we update α to $\alpha + v_i$ if $\alpha \cdot v_i \leq 0$. For such a v_i we have the following inequalities

$$(\alpha + v_i) \cdot \alpha^* = \alpha \cdot \alpha^* + v_i \cdot \alpha^* \geq \alpha \cdot \alpha^* + 1$$

and

$$(\alpha + v_i) \cdot (\alpha + v_i) = |\alpha|^2 + 2v_i \cdot \alpha + |v_i|^2 \leq |\alpha|^2 + |v_i|^2 \leq |\alpha|^2 + r^2.$$

If the second step of the algorithm is executed after k steps, then we get for the new $\alpha + v_i$ that

$$(\alpha + v_i) \cdot \alpha^* \geq \alpha \cdot \alpha^* + 1 \geq k + 1 \qquad \text{and} \qquad |\alpha + v_i|^2 \leq |\alpha|^2 + r^2 \leq kr^2 + r^2 = (k+1)r^2.$$

□

Proposition 5.14 implies that

$$k \leq |\alpha|\,|\alpha^*| \leq \sqrt{k}r\,|\alpha^*|.$$

Therefore we get $k \leq r^2\,|\alpha^*|^2$ and there is an upper bound on the number of iterations used in the second step. So after a finite number of steps, we must have $\alpha \cdot v_i > 0$ for every $i = 1, \ldots, m$.

5.4 Pythagoras and the least squares method

The result below is a generalization of the theorem of Pythagoras about right triangles to higher dimensions.

(5.15) PROPOSITION.

If $u, v \in \mathbb{R}^n$ and $u \perp v$, then

$$|u + v|^2 = |u|^2 + |v|^2.$$

Proof. This follows from

$$(u+v)\cdot(u+v) = u\cdot u + u\cdot v + v\cdot u + v\cdot v = u\cdot u + v\cdot v = |u|^2 + |v|^2,$$

since $u\cdot v = v\cdot u = 0$. □

The dot product and the norm have a vast number of applications. One of them is the method of least squares: suppose that you are presented with a system

$$Ax = b \tag{5.10}$$

of linear equations, where A is an $m \times n$ matrix.

You may not be able to solve (5.10). There could be for example 17 equations and only 2 unknowns making it impossible for all the equations to hold. As an example, the system

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} \tag{5.11}$$

of three linear equations and two unknowns does not have any solutions.

The method of (linear) least squares seeks the best approximate solution x_0 to (5.10) as a solution to the minimization problem

$$\begin{array}{ll} \text{Minimize} & |b-Ax|^2 \\ \text{with constraint} & \\ & x \in \mathbb{R}^n. \end{array} \tag{5.12}$$

There is a surprising way of finding optimal solutions to (5.12):

(5.16) THEOREM.

If $x_0 \in \mathbb{R}^n$ is a solution to the system

$$(A^\top A)x = A^\top b \tag{5.13}$$

of n linear equations with n unknowns, then x_0 is an optimal solution to (5.12). If x_0 on the other hand is an optimal solution to (5.12), then x_0 is a solution to (5.13).

Proof. Suppose we know that $b - Ax_0$ is orthogonal to Av for every $v \in \mathbb{R}^n$. Then

$$|b-Ax|^2 = |b-Ax_0+A(x_0-x)|^2 = |b-Ax_0|^2 + |A(x-x_0)|^2$$

for every $x \in \mathbb{R}^n$ by Proposition 5.15. So, in the case that $b - Ax_0 \perp Av$ for every $v \in \mathbb{R}^n$ we have

$$|b-Ax|^2 \geq |b-Ax_0|^2$$

for every $x \in \mathbb{R}^n$ proving that x_0 is an optimal solution to (5.12).

Now we wish to show that $b - Ax_0$ is orthogonal to Av for every $v \in \mathbb{R}^n$ if and only if $A^\top Ax_0 = A^\top b$. This is a computation involving the matrix arithmetic introduced in Chapter 3:

$$\begin{aligned} &(b-Ax_0)\cdot Av = \\ &(b-Ax_0)^\top Av = \\ &b^\top Av - x_0^\top A^\top Av = \\ &(b^\top A - x_0^\top A^\top A)v = 0 \end{aligned}$$

for every $v \in \mathbb{R}^n$ if and only if $b^\top A - x_0^\top A^\top A = 0$. But

$$(b^\top A - x_0^\top A^\top A)^\top = A^\top b - A^\top A x_0$$

so that $(A^\top A)x_0 = A^\top b$.

On the other hand, if $|b - Ax_0|^2 \leq |b - Ax|^2$ for every $x \in \mathbb{R}^n$, then $(b - Ax_0) \cdot Av = 0$ for every $v \in \mathbb{R}^n$: if we could find v with $(b - Ax_0) \cdot Av < 0$, then

$$|b - A(x_0 - \varepsilon v)|^2 < |b - Ax_0|^2$$

for a small number $\varepsilon > 0$. This follows, since

$$|b - A(x_0 - \varepsilon v)|^2 = ((b - Ax_0) + \varepsilon Av)^2,$$

which is

$$|b - Ax_0|^2 + 2\varepsilon(b - Ax_0) \cdot Av + \varepsilon^2 (Av)^2$$

By picking $\varepsilon > 0$ sufficiently small,

$$\varepsilon(2(b - Ax_0) \cdot Av + \varepsilon(Av)^2) < 0.$$

□

In a future course on linear algebra you will see that the system of linear equations in Theorem 5.16 is always solvable i.e., an optimal solution to (5.12) can always be found in this way.

(5.17) EXERCISE.

Show that (5.11) has no solutions. Compute the best approximate solution to (5.11) using Theorem 5.16. ♠

(5.18) EXAMPLE.

The classical application of the least squares method is to find the best line $y = \alpha x + \beta$ through a given set of points

$$(x_1, y_1), \quad (x_2, y_2), \quad \ldots \quad , (x_n, y_n)$$

in the plane $\mathbb{R}^2$.

Usually we cannot find a line matching the points precisely. This corresponds to the fact that the system of equations

$$\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

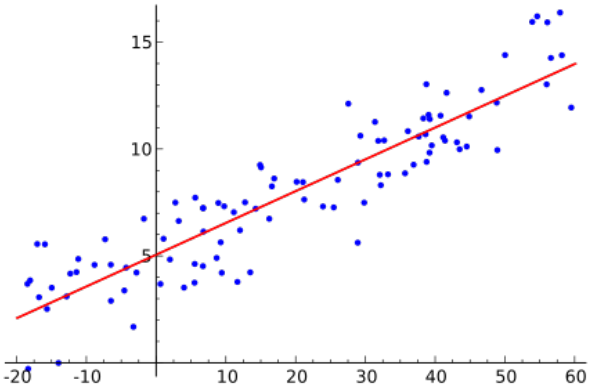
has no solutions.

Working with the least squares solution, we try to compute the best line $y = \alpha x + \beta$ in the sense that

$$(y_1 - \alpha x_1 - \beta)^2 + (y_2 - \alpha x_2 - \beta)^2 + \cdots + (y_n - \alpha x_n - \beta)^2$$

is minimized.

(5.19) FIGURE.



Best fit of line to random points from Wikipedia.

We might as well have asked for the best quadratic polynomial

$$y = \alpha x^2 + \beta x + \gamma$$

passing through the points

$$(x_1, y_1), \quad (x_2, y_2), \quad \dots \quad , (x_n, y_n)$$

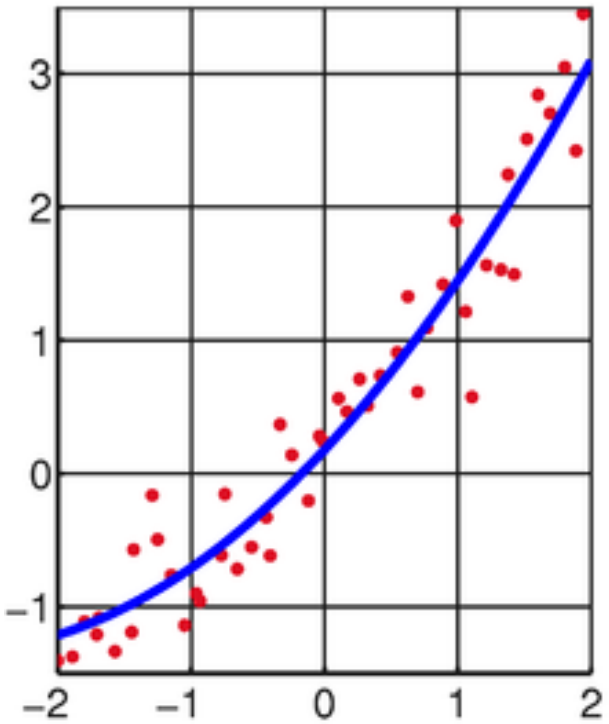
in $\mathbb{R}^2$.

The same method gives us the system

$$\begin{pmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \\ x_n^2 & x_n & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

of linear equations.

(5.20) FIGURE.



Best fit of quadratic polynomial to random points from Wikipedia.

The method generalizes naturally to finding the best polynomial of degree m

$$y = a_m x^m + a_{m-1} x^{m-1} + \cdots + a_1 x + a_0$$

through a given set of points. ♠

(5.21) EXERCISE.

Find the best line $y = \alpha x + \beta$ through the points $(1,2),(2,1)$ and $(4,3)$ and the best quadratic polynomial $y = ax^2 + bx + c$ through the points $(-2,2),(-1,1),(0,0),(1,1)$ and $(2,2)$.

It is important here, that you write down the relevant system of linear equations according to Theorem 5.16. It is however ok to solve the equations on a computer (or check your best fit on WolframAlpha).

Also, you can get a graphical illustration of your result in the sage window below.

Interactive code not included in static version.

♠

(5.22) EXERCISE.

A circle with center (a,b) and radius r is given by the equation

$$(x-a)^2 + (y-b)^2 = r^2. \tag{5.14}$$

1. Explain how (5.14) can be rewritten to the equation

   $$2ax + 2by + c = x^2 + y^2, \tag{5.15}$$

   where $c = r^2 - a^2 - b^2$.

2. Explain how fitting a circle to the points $(x_1,y_1),\ldots,(x_n,y_n)$ in the least squares context using (5.15) leads to the system

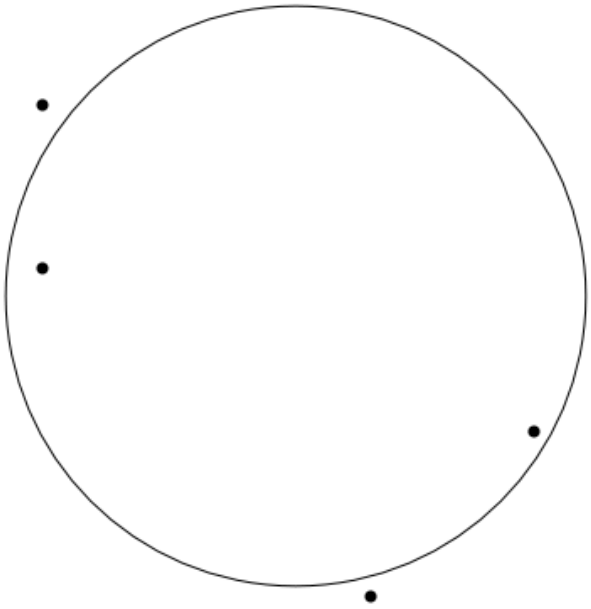
   $$\begin{pmatrix} 2x_1 & 2y_1 & 1 \\ 2x_2 & 2y_2 & 1 \\ \vdots & \vdots & \vdots \\ 2x_n & 2y_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} x_1^2 + y_1^2 \\ x_2^2 + y_2^2 \\ \vdots \\ x_n^2 + y_n^2 \end{pmatrix},$$

   of linear equations.

3. Compute the best circle through the points

   $$(0,2), \quad (0,3), \quad (2,0) \quad \text{and} \quad (3,1)$$

   by giving the center coordinates and radius with two decimals. Use the Sage window below to plot your result too see if it matches the drawing.



Interactive code not included in static version.

♠

5.5 The Cauchy-Schwarz inequality

Take another look at 5 in Definition 5.1. It is actually a small miracle that no matter which (non-zero) vectors u and v you use as input to the `cosine` function defined in Example 5.2, you always get a number between -1 and 1. The mathematics behind this is rather elegant. It is a consequence of the famous Cauchy-Schwarz inequality stated and proved below.

(5.23) THEOREM.

For two vectors $u, v \in \mathbb{R}^n$,

$$|u \cdot v| \leq |u|\,|v|\,.$$

Proof. We consider the function $q : \mathbb{R} \to \mathbb{R}$ given by

$$q(x) = (xu+v)\cdot(xu+v) = |u|^2 x^2 + 2(u\cdot v)x + |v|^2$$

Then $q(x)$ is a quadratic polynomial with $q(x) \geq 0$. Therefore its discriminant must be ≤ 0 i.e.,

$$4(u\cdot v)^2 - 4\,|u|^2\,|v|^2 \leq 0,$$

which gives the result. □

(5.24) EXERCISE.

Why are the two inequalities

$$-1 \leq \frac{u\cdot v}{|u||v|} \leq 1$$

a consequence of Theorem 5.23? ♠

(5.25) EXERCISE.

For arbitrary two numbers $x, y \in \mathbb{R}$,

$$2(x^2+y^2) \geq (x+y)^2,$$

since

$$2(x^2+y^2)-(x+y)^2 = x^2+y^2-2xy = (x-y)^2 \geq 0.$$

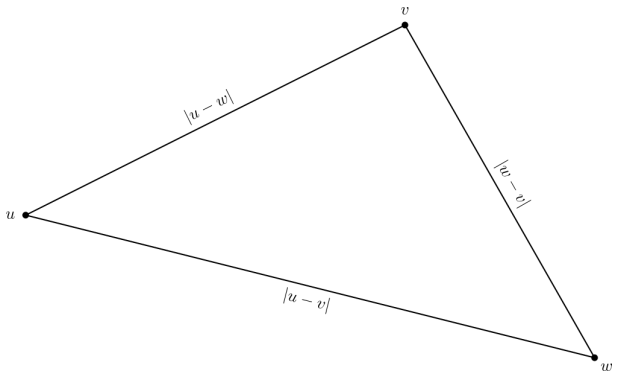
Why is

$$n(x_1^2+\cdots+x_n^2) \geq (x_1+\cdots+x_n)^2$$

for arbitary n numbers $x_1, \ldots, x_n \in \mathbb{R}$? ♠

5.5.1 The triangle inequality

Another nice consequence of the Cauchy-Schwarz inquality is the triangle inquality.



(5.26) COROLLARY.

For three vectors $u, v, w \in \mathbb{R}^n$,

$$d(u,w) \leq d(u,v)+d(v,w).$$

Proof. From the Cauchy-Schwarz inequality (Theorem 5.23) it follows that

$$|v_1+v_2|^2 = (v_1+v_2)\cdot(v_1+v_2) = |v_1|^2+2v_1\cdot v_2+|v_1|^2 \leq |v_1|^2+2|v_1||v_2|+|v_2|^2$$

for two vectors $v_1, v_2 \in \mathbb{R}^n$. Since the right hand side of this inequality is $(|v_1|+|v_2|)^2$, we have

$$|v_1+v_2| \leq |v_1|+|v_2|.$$

By the definition of $d(u,w)$, we then get the desired inequality as

$$d(u,w) = |u-w| = |(u-v)+(v-w)| \leq |u-v|+|v-w| = d(u,v)+d(v,w).$$

□

(5.27) EXERCISE.

Apply the triangly inequality in the form

$$|u+v| \leq |u|+|v|$$

for $u, v \in \mathbb{R}^n$ to show that

$$\begin{aligned} ||u|-|v|| &\leq |u-v| \\ ||u|-|v|| &\leq |u+v|\,. \end{aligned}$$

♠

5.5.2 Cosine similarity in machine learning

When vectors two vectors $u, v \in \mathbb{R}^n$ are interpreted as data sets, the number in 5 of Definition 5.1 is known as the *cosine similarity*. It measures the correlation between the vectors u and v.

A very primitive way of modelling sentences in a language is the socalled *one-hot encoding* of its words. We will illustrate this by an example. Suppose that our language consists of the words

'a', 'and', 'applicable', 'are', 'fun', 'is', 'mathematics', 'matrices', 'matrix', 'useful'

Each word gets embedded into $\mathbb{R}^{10}$ with a vector associated to its row below

$$\begin{array}{lcccccccccc} \text{a} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{and} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{applicable} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{are} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{fun} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \text{is} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \text{mathematics} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \text{matrices} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \text{matrix} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \text{useful} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \tag{5.16}$$

Now onsider the two sentences "*mathematics is fun and a matrix is useful*" and "*mathematics is fun and matrices are applicable*".

From the words in the two strings we form the following vectors in $\mathbb{R}^{10}$ using the one-hot embedding in (5.16).

$$\begin{array}{lcccccccccc} \text{mathematics is fun and a matrix is useful} & 1 & 1 & 0 & 0 & 1 & 2 & 1 & 0 & 1 & 0 \\ \text{mathematics is fun and matrices are applicable} & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \end{array}$$

Here a sentence is mapped to the vector, which is the sum of all the vectors corresponding to the words in the sentence, where each vector is multiplied by its multiplicity i.e., how many times the word occurs. The closer the cosine gets to 1 (corresponding to an angle of 0 degrees), the more similar we consider the sentences. Use the python snippet below to experiment and compute the cosine similarity in the example.

Interactive code not included in static version.

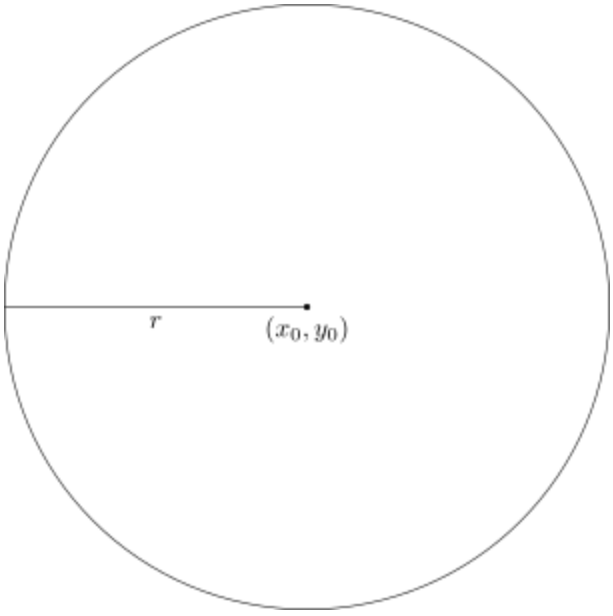
Cosine similarity is crucial in machine learning, especially in NLP.

The one-hot embedding is very crude and does not really capture the semantics of a sentence. The bread and butter of modern (large) language models is more advanced (dense) embeddings constructed using deep learning. The embeddings even take whole sentences as input! The recent breakthroughs can be traced back to 2013, where Google introduced the word embedding word2vec. When embedding a sentence one usually considers tokens and not words. This means that every sentence (as input to ChatGPT) must be broken down into a sequence of tokens. Modern large language models typically operate with around 50,000 tokens. Each token is embedded into a euclidean space of dimension usually > 1000.

5.6 Special subsets of euclidean spaces

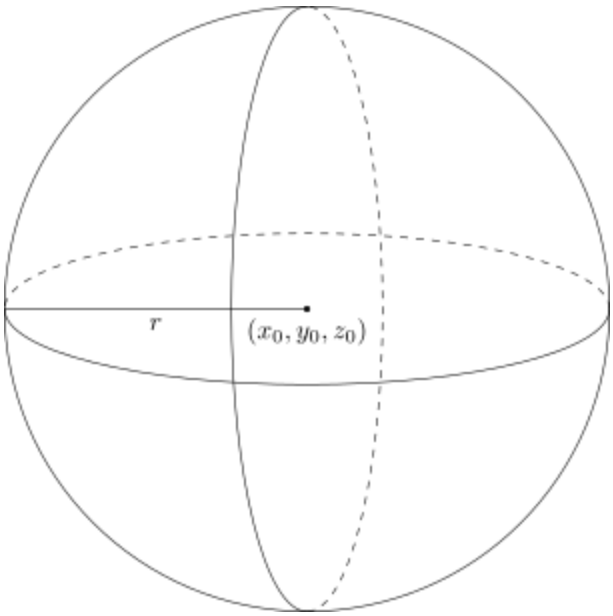
Recall that a circle (or an open disk) centered at $(x_0, y_0) \in \mathbb{R}^2$ with radius $r \in \mathbb{R}$ is defined as the subset

$$\begin{aligned}
&\left\{(x,y) \in \mathbb{R}^2 \middle| (x-x_0)^2 + (y-y_0)^2 < r^2\right\} = \\
&\left\{(x,y) \in \mathbb{R}^2 \middle| \sqrt{(x-x_0)^2 + (y-y_0)^2} < r\right\} = \\
&\left\{(x,y) \in \mathbb{R}^2 \middle| d((x_0,y_0),(x,y)) < r\right\}.
\end{aligned}$$



Similarly an open ball in $\mathbb{R}^3$ centered at $(x_0, y_0, z_0) \in \mathbb{R}^3$ with radius $r \in \mathbb{R}$ is defined as the subset

$$\begin{aligned}
&\left\{(x,y,z) \in \mathbb{R}^3 \middle| (x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2 < r^2\right\} = \\
&\left\{(x,y,z) \in \mathbb{R}^3 \middle| \sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2} < r\right\} = \\
&\left\{(x,y,z) \in \mathbb{R}^3 \middle| d((x_0,y_0,z_0),(x,y,z)) < r\right\}.
\end{aligned}$$



The natural generalization of this definition to higher dimensions is given below.

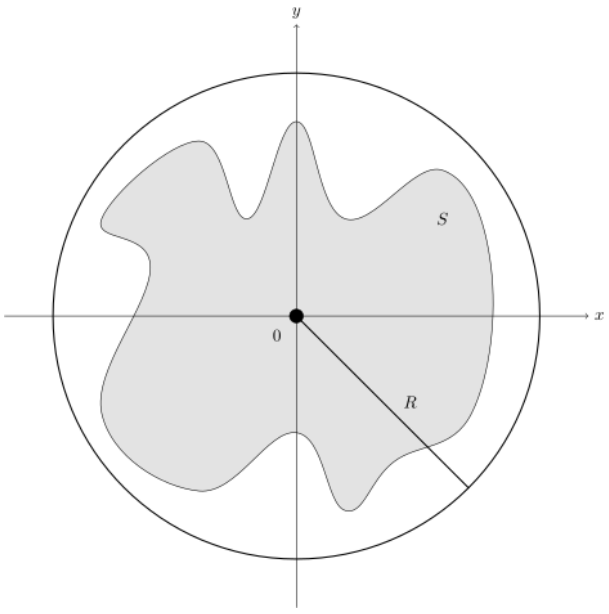
(5.28) DEFINITION.

The open ball centered at $u \in \mathbb{R}^n$ with radius $r \in \mathbb{R}$ is defined as

$$B(u,r) = \{v \in \mathbb{R}^n \mid d(u,v) < r\}.$$

5.6.1 Bounded subsets

It makes sense to defined bounded subsets as subsets that can be contained in a large enough open ball centered at 0:



(5.29) DEFINITION.

A subset $S \subseteq \mathbb{R}^n$ is called bounded if there exists $R \in \mathbb{R}$, such that

$$S \subseteq B(0,R).$$

(5.30) REMARK.

Written out Definition 5.29 says that

$$|u| = \sqrt{x_1^2 + \cdots + x_n^2} \leq R$$

for every $u = (x_1, \ldots, x_n) \in S$. Boundedness of S is also equivalent to the following two conditions

(i) There exists R, such that

$$x_1^2 + \cdots + x_n^2 \leq R$$

for every $(x_1, \ldots, x_n) \in S$.

(ii) There exists R, such that

$$|x_i| \leq R$$

for $i = 1, \ldots, n$ and every $(x_1, \ldots, x_n) \in S$.

Every finite subset is bounded (why?). For $d = 1$, Definition 5.29 simply says

$$\exists R \in \mathbb{R}\, \forall x \in S : |x| \leq R$$

This implies that an interval $S = [a,b]$ is bounded by putting $R = \max(|a|, |b|)$ in Definition 5.29.

LLM

```
I find the definition below quite hard to understand. It is about bounded subsets.
Please explain it to me patiently, give some examples and test me afterwards.
'''
A subset $S \subseteq \mathbb{R}^n$ is called bounded if there exists
$R\in \mathbb{R}$, such that
$$
S \subseteq B(0, R),
$$
where $B(0, R) =  \{v\in \mathbb{R}^n \mid d(0, v) < R\}$ and $d$ is the
euclidean distance function.
'''
```

(5.31) EXERCISE.

Show precisely that the subset $\mathbb{N}$ of $\mathbb{R}$ is not bounded, whereas the subset $\{1, \frac{1}{2}, \frac{1}{3}, \ldots\}$ is. ♠

(5.32) EXERCISE.

Sketch why

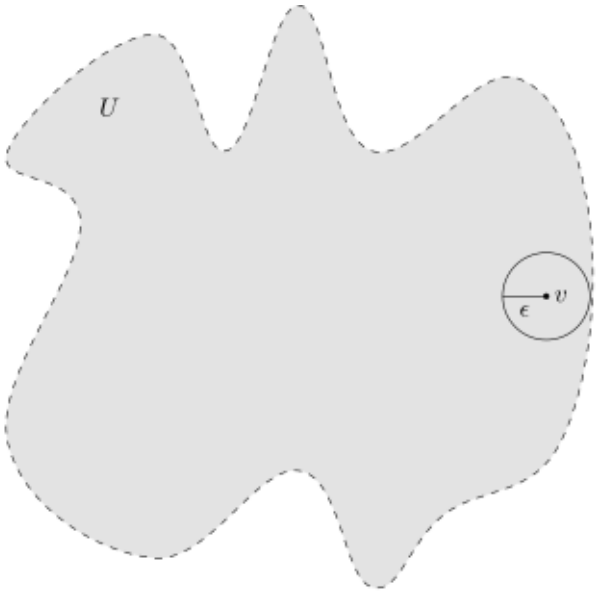
$$S = \{(x,y) \mid x \geq 0, y \geq 0, x+y \leq 1\} \subseteq \mathbb{R}^2$$

is bounded. Now use Fourier-Motzkin elimination to show the same without sketching. ♠

5.6.2 Open, closed and compact subsets and boundaries and interiors of subsets

Open subsets

An open subset of $\mathbb{R}^n$ is a subset consisting of points, that are interior in the following sense:



(5.33) DEFINITION.

A subset $U \subseteq \mathbb{R}^n$ is called open *if for every $v \in U$, there exists $\varepsilon > 0$, such that*

$$B(v,\varepsilon) \subseteq U.$$

(5.34) EXERCISE.

Decide whether each of the subsets given below are open.

(a) $\{1\}$

(b) $\mathbb{R}^n$

(c) $[0,1]$

(d) $(0,1)$

♠

(5.35) EXERCISE.

Prove that an open ball given by $B(v,\varepsilon) \subseteq \mathbb{R}^n$ is an open subset.

Hint: Suppose that $u \in B(v,\varepsilon)$. Define a suitable $\varepsilon' > 0$ for $u \in B(v,\varepsilon)$ and use Corollary 5.26 to conclude that $B(u,\varepsilon') \subseteq B(v,\varepsilon)$. ♠

(5.36) EXERCISE.

Show that a finite subset of $\mathbb{R}^n$ is never open. ♠

We will need the result below.

(5.37) PROPOSITION.

If $U_1,\ldots,U_m \subseteq \mathbb{R}^n$ are open subsets, then

$$U_1 \cup \cdots \cup U_m \qquad \text{and} \qquad U_1 \cap \cdots \cap U_m$$

are open subsets.

Closed subsets

(5.38) DEFINITION.

A subset $F \subseteq \mathbb{R}^n$ is called closed *if $\mathbb{R}^n \setminus F$ is open.*

In analogy with Proposition 5.37 we have the result below.

(5.39) PROPOSITION.

If $F_1,\ldots,F_m \subseteq \mathbb{R}^n$ are closed subsets, then

$$F_1 \cup \cdots \cup F_m \quad \text{and} \quad F_1 \cap \cdots \cap F_m$$

are closed subsets.

(5.40) EXERCISE.

Decide whether each of the subsets given below are closed.

(a) $\{1\}$

(b) $\mathbb{R}^n$

(c) $[0,1]$

(d) $[0,1)$

♠

Open intervals

(5.41) PROPOSITION.

The following subsets

$$(a,\infty) = \{x \in \mathbb{R} \mid a < x\}$$
$$(-\infty,a) = \{x \in \mathbb{R} \mid x < a\}$$
$$(a,b) = \{x \in \mathbb{R} \mid a < x < b\}$$

are open subsets of $\mathbb{R}$ *for every* $a,b \in \mathbb{R}$.

Proof. Let us prove that (a,∞) is an open subset of $\mathbb{R}$. If $x \in (a,\infty)$, then we let $\varepsilon = x - a$. Suppose that $|y - x| < \varepsilon$. If $y > x$, then $y > a$ and $y \in (a,\infty)$. If $y < x$, then $x - y < \varepsilon = x - a$ and therefore $y > a$ and $y \in (a,\infty)$. We have proved that (a,∞) is an open subset.

A similiar proof shows that $(-\infty,a)$ is an open subset. If $a < b$, then

$$(a,b) = (-\infty,b) \cap (a,\infty),$$

which is an open subset by the above and Proposition 5.37. □

Closed intervals

We have a similar result for closed subsets.

(5.42) PROPOSITION.

The following subsets

$$[a,b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$$
$$[a,\infty) = \{x \in \mathbb{R} \mid a \leq x\}$$
$$(-\infty,a] = \{x \in \mathbb{R} \mid x \leq a\}$$

are closed subsets of $\mathbb{R}$ *for every* $a,b \in \mathbb{R}$.

Proof. The proof follows from Definition 5.38 and Proposition 5.41. For example,

$$\mathbb{R} \setminus [a,\infty) = (-\infty,a).$$

□

Compact subsets

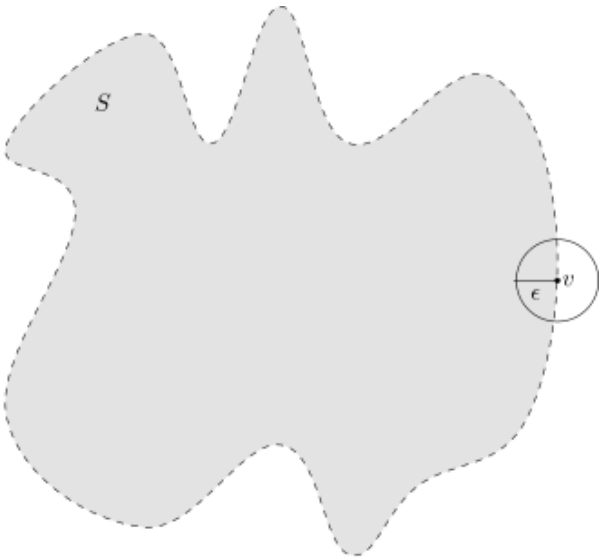
We single out the following very important class of subsets

(5.43) DEFINITION.

A subset $C \subseteq \mathbb{R}^n$ is called compact if it is bounded and closed.

The boundary of a subset

The boundary of a subset S is informally the subset of points barely touching S:



This is made precise in the following definition.

(5.44) DEFINITION.

The boundary ∂S of a subset $S \subseteq \mathbb{R}^n$ is defined as

$$\partial S = \{v \in \mathbb{R}^n \mid \forall \varepsilon > 0 : B(v,\varepsilon) \cap S \neq \emptyset \qquad \textit{and} \qquad B(v,\varepsilon) \cap (\mathbb{R}^n \setminus S) \neq \emptyset\}.$$

(5.45) EXERCISE.

What is the boundary of $[0,1]$? What about $(0,1)$? ♠

The interior of a subset

The interior of a subset consists of the points, which are interior to the subset. More precisely

(5.46) DEFINITION.

The interior of a subset $S \subseteq \mathbb{R}^n$ is defined by

$$S^\circ = \{v \in S \mid \exists \varepsilon > 0 : B(v,\varepsilon) \subseteq S\}.$$

It is a fun little exercise to prove that the interior is an open subset.

(5.47) EXERCISE.

Let $S = [0,1] \subseteq \mathbb{R}$ and $S \times \{1\} = \{(x,1) \mid x \in S\} \subseteq \mathbb{R}^2$. First make a sketch of these two subsets in $\mathbb{R}$ and $\mathbb{R}^2$ respectively. Then find S° and $(S \times \{1\})^\circ$. ♠

5.7 Continuous functions

(5.48) DEFINITION.

A function $f : S \to T$, where $S \subseteq \mathbb{R}^m$ and $T \subseteq \mathbb{R}^n$ is called continuous at $v \in S$ *if for every $\varepsilon > 0$, there exists $\delta > 0$, such that*

$$u \in B(v,\delta) \cap S \implies f(u) \in B(f(v),\varepsilon).$$

for every u. Equivalently,

$$\forall \varepsilon > 0 \exists \delta > 0 \forall u \in S : d(u,v) < \delta \implies d(f(u),f(v)) < \varepsilon. \tag{5.17}$$

The function f is called continuous if it is continuous at every $v \in S$.

LLM

```
I find the definition below quite challenging to understand. Please explain
it to me patiently with lots of examples.
Test me in the end.
'''
A function $f:\rightarrow T$, where $S \subseteq \mathbb{R}^m$ and
$T\subseteq \mathbb{R}^n$ is called \emph{continuous at $v\in S$}
if for every $\epsilon > 0$, there exists $\delta > 0$,
such that
\begin{equation}
\forall \epsilon > 0\, \exists \delta > 0\, \forall u\in S:
d(u, v) < \delta \implies d(f(u), f(v)) < \epsilon.
\end{equation}
'''
```

Definition 5.48 is the formal definition of a continuous function. It is short and sweet, but takes some time to assimilate.

The limit of a function at a point

The definition presented in (5.17) is admittedly a bit long. I will introduce some notation to make it easier. Assume that we have the same setup as in Definition 5.48, but that we do not require that $v \in S$. Then we use the notation

$$\lim_{u \to v} f(u) = w$$

to mean

$$\forall \varepsilon > 0 \exists \delta > 0 \forall u \in S : d(u,v) < \delta \implies d(f(u),w) < \varepsilon$$

and say that $f(u)$ has limit w as u approaches v (inside S). Then (5.17) reads

$$\lim_{u \to v} f(u) = f(v).$$

Here is a little exercise to practice the notation: suppose that

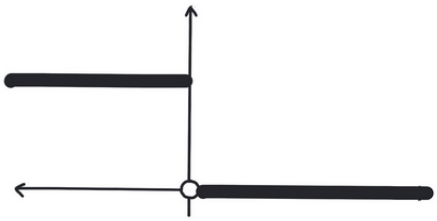
$$\frac{x^2-1}{x-1}.$$

Then f is a function defined on $S = \mathbb{R} \setminus \{1\}$. What is $\lim_{x \to 1} f(x)$?

To get an understanding, you should study the mother of all examples of non-continuous functions given below:

$$f(x) = \begin{cases} 0 & \text{if } x > 0 \\ 1 & \text{if } x \leq 0 \end{cases}. \tag{5.18}$$

This is a function from $S = \mathbb{R}$ to $T = \mathbb{R}$. It is impossible to plot it without lifting the pencil or defining such a beast without using a bracket as in (5.18).



Let me sketch how the formal Definition 5.48 kills any hope of (5.18) being continuous at $v = 0$. To prove this we must prove that the negation of the proposition in (5.17) is true. This reads

$$\exists \varepsilon > 0 \forall \delta > 0 \exists u \in S : d(u,0) < \delta \wedge d(f(u),1) \geq \varepsilon$$

for the function defined in (5.18). You can verify that the above is true by setting $\varepsilon = \frac{1}{2}$ and $u = \frac{\delta}{2}$. For these values,

$$d(u,0) = \frac{\delta}{2} < \delta \qquad \text{and} \qquad d(f(u),1) = d(0,1) = 1 \geq \frac{1}{2}.$$

Almost all functions we encounter will be continuous. The function f above is an anomaly.

Let us stop briefly once more and see Definition 5.48 in action.

(5.49) EXAMPLE.

Let $S = T = \mathbb{R}$ in Definition 5.48. We consider the two functions

$$f(x) = x$$
$$g(x) = c,$$

where $c \in \mathbb{R}$ i.e., f is the identity function and g is a constant function given by the real number c. Both of these functions are continuous. Let us see why.

For the function f, (5.17) reads

$$\forall \varepsilon > 0 \exists \delta > 0 \forall y \in \mathbb{R} : d(y,x) < \delta \implies d(f(y), f(x)) = d(y,x) < \varepsilon.$$

This is certainly true if we pick $\delta = \varepsilon$.

For the function g, (5.17) reads

$$\forall \varepsilon > 0 \exists \delta > 0 \forall y \in \mathbb{R} : d(y,x) < \delta \implies d(f(y), f(x)) = d(c,c) = 0 < \varepsilon.$$

Here $\delta > 0$ can be picked arbitrary, since $0 = d(c,c) < \varepsilon$ is always true. ♠

5.7.1 An elegant way of characterizing a continuous function

Recall the definition of the preimage from Definition 1.112 and the definition of an open subset from Definition 5.33. The following characterization of continuous functions came rather late in the history of mathematics.

(5.50) PROPOSITION.

Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a function. Then f is continuous if and only if $f^{-1}(U)$ is open in $\mathbb{R}^m$ for every open subset $U \subseteq \mathbb{R}^n$.

Proof. Let $U \subseteq \mathbb{R}^n$ be an open subset. Assume first that f is continuous. We wish to prove that $f^{-1}(U)$ is open. Pick $v \in f^{-1}(U)$ and $\varepsilon > 0$ so that $B(f(v), \varepsilon) \subseteq U$. Now use the continuity of f to pick $\delta > 0$ so that (5.17) is satisfied i.e.,

$$u \in B(v, \delta) \implies f(u) \in B(f(v), \varepsilon) \tag{5.19}$$

Since $B(f(v), \varepsilon) \subseteq U$, (5.19) says that

$$B(v, \delta) \subseteq U$$

showing that $f^{-1}(U)$ is an open subset.

Now suppose that $f^{-1}(U)$ is open whenever $U \subseteq \mathbb{R}^n$ is open. For $v \in \mathbb{R}^m$ and $\varepsilon > 0$ we put $V = B(f(v), \varepsilon)$. Since V is an open subset, $f^{-1}(V)$ is open and $v \in f^{-1}(V)$. So we may find $\delta > 0$ so that $B(v, \delta) \subseteq f^{-1}(V)$. But this is exactly the statement that

$$u \in B(v, \delta) \implies f(u) \in B(f(v), \varepsilon)$$

showing that f is continuous. □

The following result is often a very useful tool in showing that a subset is closed.

(5.51) PROPOSITION.

If $F \subseteq \mathbb{R}^n$ is a closed subset and $f : \mathbb{R}^m \to \mathbb{R}^n$ a continuous function, then the preimage

$$f^{-1}(F) = \{v \in \mathbb{R}^m \mid f(v) \in F\}$$

is a closed subset of $\mathbb{R}^m$.

Proof. If $F \subseteq \mathbb{R}^n$ is closed, then $\mathbb{R}^n \setminus F$ is open. Therefore

$$f^{-1}(\mathbb{R}^n \setminus F) = f^{-1}(\mathbb{R}^n) \setminus f^{-1}(F) = \mathbb{R}^m \setminus f^{-1}(F)$$

is open by Proposition 5.50. This implies that $f^{-1}(F)$ is closed. □

(5.52) EXAMPLE.

Let us assume for now that $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x,y) = x^2 + y^2$ is continuous (see Exercise 5.58). Then Proposition 5.51 shows that the subset

$$\begin{aligned}
&\{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \geq 1\} = \\
&\{(x,y) \in \mathbb{R}^2 \mid f(x,y) \geq 1\} = \\
&\{(x,y) \in \mathbb{R}^2 \mid f(x,y) \in [1,\infty)\} = \\
&f^{-1}([1,\infty))
\end{aligned}$$

of $\mathbb{R}^2$ is closed, since $[1,\infty)$ is a closed subset of $\mathbb{R}$ by Proposition 5.42. ♠

(5.53) EXERCISE.

Show formally that the subset

$$\{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 > 1\}$$

is an open subset of $\mathbb{R}^2$. ♠

5.7.2 Working with continuous functions

We give now three important results, which can be used in concrete situations to verify that a given function is continuous. They can be proved without too much hassle. The first result below basically follows from the definition of the norm of a vector (see (5.4)).

(5.54) LEMMA.

The projection functions $\pi_i : \mathbb{R}^n \to \mathbb{R}$ defined in Definition 1.101 are continuous. In general a function $f : S \to T$ is continuous if and only if $\pi_j \circ f : S \to \mathbb{R}$ is continuous for every $j = 1, \ldots, n$, where $S \subseteq \mathbb{R}^m$ and $T \subseteq \mathbb{R}^n$.

(5.55) EXAMPLE.

Lemma 5.54 shows for example that the functions $f(x,y) = x$ and $g(x,y) = y$ are continuous functions from $\mathbb{R}^2$ to $\mathbb{R}$.

Consider the vector function $f : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$f(x,y) = \begin{pmatrix} x^2 + y^2 \\ \sin(xy) \end{pmatrix}$$

as an example. To prove that f is continuous, Lemma 5.54 tells us that it is enough to prove that its coordinate functions $f_1, f_2 : \mathbb{R}^2 \to \mathbb{R}$

$$f_1(x,y) = x^2 + y^2$$
$$f_2(x,y) = \sin(xy)$$

are continuous. ♠

Definition 5.48 also behaves nicely when continuous functions are composed. This is the content of the following

(5.56) PROPOSITION.

Suppose that $g : S \to T$ and $f : T \to R$ are continuous functions, where $S \subseteq \mathbb{R}^n, T \subseteq \mathbb{R}^e$ and $R \subseteq \mathbb{R}^f$. Then the composition

$$(f \circ g) : S \to R$$

is continuous.

To get continuous functions from functions already known to be continuous using arithmetic operations, the result below is useful.

(5.57) PROPOSITION.

Let $f, g : U \to \mathbb{R}$ be functions defined on a subset $U \subseteq \mathbb{R}^n$. If f and g are continuous, then the functions

$$\begin{aligned} (f+g) : U \to \mathbb{R} \quad & \text{given by} \quad (f+g)(x) = f(x) + g(x) \\ (fg) : U \to \mathbb{R} \quad & \text{given by} \quad (fg)(x) = f(x)g(x) \\ (f/g) : V \to \mathbb{R} \quad & \text{given by} \quad (f/g)(x) = f(x)/g(x) \end{aligned}$$

are continuous functions, where $V = \{x \in U \mid g(x) \neq 0\}$ (the last function is defined only if $g(x) \neq 0$).

Proof. This result is a consequence of the definition of continuity and Proposition 5.56. □

(5.58) EXERCISE.

Show in detail that the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y) = x^2 + y^2$$

is continuous by using Proposition 5.57 combined with Lemma 5.54. ♠

(5.59) REMARK.

By combining Example 5.49 with Proposition 5.57, one finds that every polynomial is a continuous function and that

$$h(x) = \frac{f(x)}{g(x)}$$

is continuous for $g(x) \neq 0$, where $f, g \in \mathbb{R}[x]$.

(5.60) EXERCISE.

Verify the claim in Remark 5.59. ♠

(5.61) REMARK.

More advanced (transcendental) functions like $\sin(x)$ and e^x also turn out to be continuous. We will return to this in the next chapter, where differentiable functions are defined.

(5.62) EXERCISE.

Show from scratch (without using Remark 5.59) that

$$g(x) = \frac{a(x)}{b(x)}$$

is a continuous function $g : V \to \mathbb{R}$, where $a(x) = x^2 - 3x + 2$ and $b(x) = x^2 - 4x + 3$ and

$$V = \mathbb{R} \setminus \{1, 3\}.$$

Use Proposition 5.42 and Proposition 5.51 to show that

$$\{x \in \mathbb{R} \mid a(x) \leq 17\}$$

is a closed subset of $\mathbb{R}$.

Hint: Write

$$\{x \in \mathbb{R} \mid a(x) \leq 17\} = a^{-1}(S),$$

where $S \subseteq \mathbb{R}$ is a suitable (closed) interval.

Does

$$\lim_{x \to 3} g(x)$$

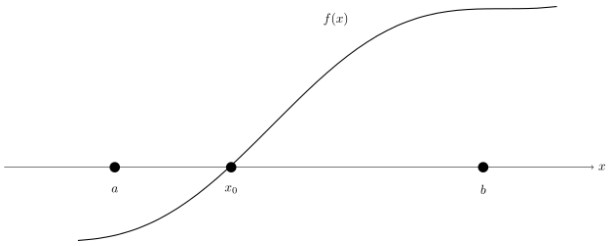
exist? What about

$$\lim_{x \to 1} g(x)?$$

♠

5.8 Important and special results for continuous functions

Below we quote a famous and very intuitive result from 1817 due to Bolzano. This result is also known as the intermediate value theorem.



(5.63) THEOREM.

Let $f:[a,b]\to\mathbb{R}$ be a continuous function, where $a<b$. If $f(a)<0$ and $f(b)>0$, then there exists x_0 with $a<x_0<b$, such that $f(x_0)=0$.

Polynomials are continuous functions. Bolzano's result fits perfectly in the proof of the result below. This result is wrong for polynomials in $\mathbb{Q}[x]$ as witnessed by $f(x)=x^3-2$, which does not have a rational root.

(5.64) EXERCISE.

Use the methods of Example 1.39 to show that there is no $\xi\in\mathbb{Q}$ with $f(\xi)=0$, where $f(x)=x^3-2$. ♠

(5.65) PROPOSITION.

Let

$$f(x)=a_nx^n+\cdots+a_1x+a_0\in\mathbb{R}[x]$$

be a polynomial of odd degree, i.e. n is odd and $a_n\neq 0$. Then f has a root, i.e. there exists $x_0\in\mathbb{R}$, such that $f(x_0)=0$.

Proof. We will assume that $a_n>0$ (if not, just multiply f by -1). Consider $f(x)$ written as

$$f(x)=x^n\left(a_n+\frac{a_{n-1}}{x}+\cdots+\frac{a_1}{x^{n-1}}+\frac{a_0}{x^n}\right).$$

By choosing c negative with $|c|$ extremely big, we have $f(c)<0$, since c^n is negative and

$$a_n+\frac{a_{n-1}}{c}+\cdots+\frac{a_1}{c^{n-1}}+\frac{a_0}{c^n}>0$$

as a_n is positive. Notice here that the terms

$$\frac{a_{n-1}}{c}+\cdots+\frac{a_1}{c^{n-1}}+\frac{a_0}{c^n}$$

are extremely small, when $|c|$ is extremely big.

Similarly by choosing d positive and tremendously big, we have $f(d)>0$. By Theorem 5.63, there exists x_0 with $c<x_0<d$ with $f(x_0)=0$. □

We end this section with a result that might be coined the mathematical cornerstone of optimization (also due to Bolzano, at least for $n = 1$). The result below is called *the extreme value theorem*.

(5.66) THEOREM.

Let C be a compact subset of $\mathbb{R}^n$ and $f : C \to \mathbb{R}$ a continuous function. Then there exists $v_{\min}, v_{\max} \in C$, such that

$$f(v_{\min}) \leq f(v) \qquad \textit{and} \qquad f(v) \leq f(v_{\max})$$

for every $v \in C$.

This is a rather stunning result! You are guaranteed solutions to optimization problems of the type

$$\begin{array}{ll} \text{Minimize} & f(v) \\ \text{with constraint} & \\ & v \in C, \end{array}$$

where C is a compact subset and $f : C \to \mathbb{R}$ a continuous function. Finding the optimal solutions in this setting is another story. It can be extremely hard. For the rest of these notes we will actually dive into methods for computing optimal solutions of optimization problems such as the one above.

(5.67) EXERCISE.

Give two examples, where Theorem 5.66 fails for $n = 1$ if we relax the conditions on C. One, where C is open and another one where C is not bounded. ♠

Chapter 6

Convex functions

In this chapter we will dive deeper into convex functions. The main focus will be on (differentiable) convex functions defined on intervals (convex subsets) of the real numbers i.e. (differentiable) convex functions in just one variable. Along the way, differentiability is formally introduced. I will assume that you are familiar with differentiation in an operational manner.

6.1 Strictly convex functions

Below we strengthen Definition 4.24 of a convex function.

(6.1) DEFINITION.

Let $C \subseteq \mathbb{R}^n$ be a convex subset. A strictly convex function *is a convex function $f : C \to \mathbb{R}$, such that*

$$f((1-t)u+tv) < (1-t)f(u)+tf(v) \tag{6.1}$$

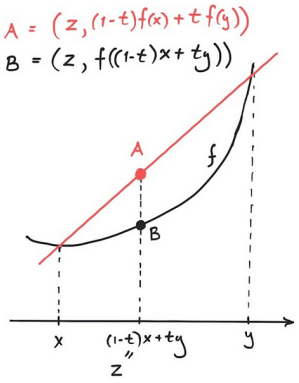
for every number t with $0 < t < 1$ and every $u, v \in C$ with $u \neq v$.

(6.2) REMARK.

The strict inequality in (6.1) collapses to an equality if $u = v, t = 0$ or $t = 1$. For example, if $u = v$, then the left hand side of (6.1) is $f((1-t)u+tu) = f(u)$ and the right hand side is $(1-t)f(u)+tf(u) = f(u)$.

(6.3) FIGURE.

Definition 6.1 is illustrated below for a function $f : \mathbb{R} \to \mathbb{R}$. Here both $u = x$ and $v = y$ are real numbers (that is, $n = 1$ in $\mathbb{R}^n$ in Definition 6.1). The (red) line segment between $(x, f(x))$ and $(y, f(y))$ lies strictly *(<) above the (black) graph of f:*



LLM

```
Please explain patiently the definition below to me. It seems that it is
also valid for functions defined on vectors in the plane ($n=2$).
Give concrete examples of this. Test me with a few questions
in the end.
'''
Let $C \subseteq \mathbb{R}^n$ be a convex subset.
A \emph{strictly convex function} is a convex function
$f: C\rightarrow \mathbb{R}$, such that
\begin{equation}
f((1 - t) u + t v) < (1-t) f(u) + t f(v)
\end{equation}
for every number $t$ with $0 <  t < 1$ and every $u, v\in C$ with $u\neq v$.
```

(6.4) EXAMPLE.

Consider the line (function) $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = ax + b$$

for $a, b \in \mathbb{R}$. This function is convex, since we can formally write for every $t \in \mathbb{R}$:

$$\begin{aligned} f((1-t)x+ty) &= a((1-t)x+ty)+b \\ &= a((1-t)x)+(1-t)b+a(ty)+tb \\ &= (1-t)(ax+b)+t(ay+b) \\ &= (1-t)f(x)+tf(y). \end{aligned} \tag{6.2}$$

However, the computation in (6.2) also shows why there is no chance that $f(x)$ is strictly convex. Intuitively, the graph of convex functions need to bend and curve a bit to be strictly convex. No lines should occur in their graphs. ♠

(6.5) EXERCISE.

Let f be a convex function. Show f is strictly convex if and only if

$$f((1-t)x+ty) = (1-t)f(x)+tf(y)$$

for $0 < t < 1$ implies that $x = y$. ♠

(6.6) EXERCISE.

Give an example of a non-constant convex function $f : \mathbb{R} \to \mathbb{R}$, which is not strictly convex. Show in details that $f(x) = x^2$ is a strictly convex function.

Hint: Look back to the relevant part of Exercise 4.26 for dealing with $f(x) = x^2$. ♠

6.2 Why are convex functions interesting?

We begin this section by giving the following result without proof.

(6.7) THEOREM.

A convex function defined on an open convex subset is continuous.

(6.8) EXERCISE.

Give an example showing that Theorem 6.7 is not true if the convex function is defined on a closed convex subset.

Hint: Try to come up with an example like $f : [0,1] \to \mathbb{R}$. Look at the end point 0.

Hint: Well, try out

$$f(x) = \begin{cases} 1 & \text{if } x = 0 \\ x & \text{if } x > 0 \end{cases}$$

♠

Let us now define precisely what is meant by a local vs a global minimum for a function.

(6.9) DEFINITION.

Let $f : S \to \mathbb{R}$ be a function, where $S \subseteq \mathbb{R}^n$ is an arbitrary subset (not necessarily convex, open or closed). Then $x_0 \in S$ is called a local minimum *for f if*

$$f(x_0) \leq f(x)$$

for every $x \in S$, which is sufficiently close to x_0. Being sufficiently close to means that $x \in S$ satisfies

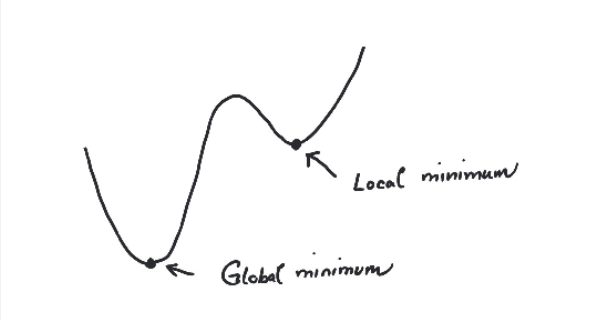
$$|x - x_0| < \varepsilon$$

for some fixed $\varepsilon > 0$.
In a much stronger notion, $x_0 \in S$ is called a global minimum *if*

$$f(x_0) \leq f(x)$$

for every $x \in S$ (not just locally).

(6.10) FIGURE.



Graph of function defined on an interval. This function has a local minimum, which is not a global minimum.

(6.11) EXERCISE.

Give an example of a local minimum that is not a global minimum for a precisely specified function. Also give an example of a global minimum, which is not uniquely defined (again for a precisely specified function). Uniquely defined means that there is precisely one x_0, such that $f(x_0)$ is minimal. ♠

We might as well have talked about maximum instead of minimum above.

(6.12) EXERCISE.

Reformulate Definition 6.9 in order to define a local and a global maximum. ♠

A *local extremum* is a point $x_0 \in S$, which is either a local minimum or a local maximum.

Convex functions $f : C \to \mathbb{R}$ are interesting, because of the local nature of the minimization problem

$$\begin{array}{ll} \text{Minimize} & f(x) \\ \text{with constraint} & \\ & x \in C \end{array} \tag{6.3}$$

If you run into a local minimum in (6.3), then you are sure that it also is a global minimum! This is the content of the result below.

(6.13) THEOREM.

Let $f : C \to \mathbb{R}$ be a convex function defined on a convex subset $C \subseteq \mathbb{R}^n$. If $x_0 \in C$ is a local minimum, then x_0 is a global minimum. If f is strictly convex, then a global minimum for f is unique.

Proof. By the definition of local minimum in Definition 6.9, there exists $\varepsilon > 0$, such that $f(x_0) \leq f(x)$, when $x \in C$ and $|x - x_0| < \varepsilon$. Suppose that x_0 is not a global minimum. Then there exists $x_1 \in C$ with $f(x_1) < f(x_0)$. Consider the point

$$x_t = (1-t)x_0 + tx_1 \in C,$$

where $0 < t < 1$. Then

$$f(x_t) \leq (1-t)f(x_0) + tf(x_1) < (1-t)f(x_0) + tf(x_0) = f(x_0).$$

Since $|x_t - x_0| = t\,|x_1 - x_0|$, we can choose $t > 0$ sufficiently small such that $|x_t - x_0| < \varepsilon$ implying $f(x_0) \leq f(x_t)$, since x_0 is a local minimum. This contradicts that $f(x_t) < f(x_0)$ for every $0 < t < 1$. Let f be strictly convex and let x_0 be a global minimum for f. If $x_1 \in C$, $x_1 \neq x_0$ and $f(x_1) = f(x_0)$, then

$$f((1-\lambda)x_0 + \lambda x_1) < (1-\lambda)f(x_0) + \lambda f(x_1) = f(x_0)$$

for $0 < \lambda < 1$. This would contradict that global minimality of x_0, since $x_0 \neq (1-\lambda)x_0 + \lambda x_1 \in C$ for $0 < \lambda < 1$. □

The following little result turns out to be very useful and also very intuitive and drawable! It is a key component in characterizing convex differentiable functions $f(x)$ in terms of $f''(x)$. We will not give the proof here.

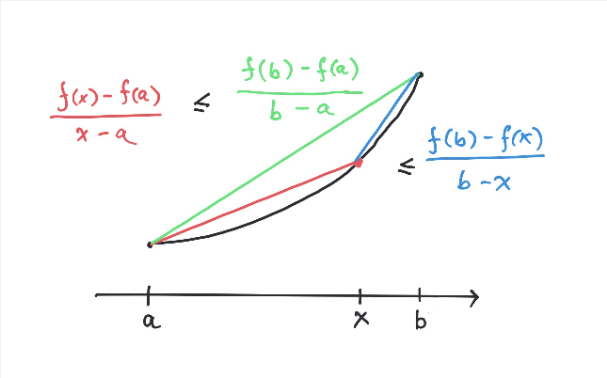
(6.14) LEMMA.

Let $f : [a,b] \to \mathbb{R}$ be a convex function. Then

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a} \leq \frac{f(b) - f(x)}{b - x}$$

for $a < x < b$.

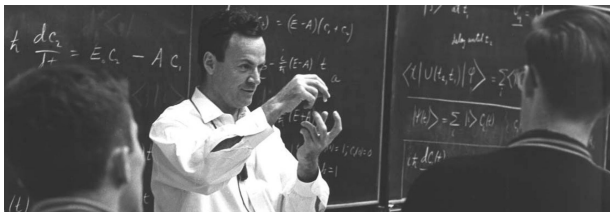
(6.15) FIGURE.



The result in Lemma 6.14 is depicted above. A formal proof can be given from first principles only using Definition 4.24.

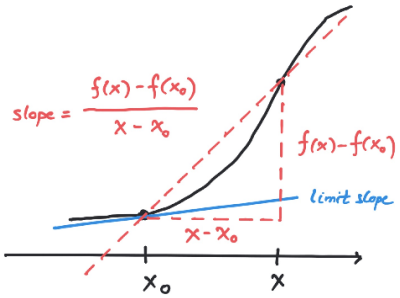
6.3 Differentiable functions

To appreciate the depth of the notion of differentiability, you should read the story (joke, actually) in the second paragraph of section 8-2 in volume I of the famous Feynman Lectures on Physics. Below is a photograph of the master explainer in action.



6.3.1 Definition

Let $f : (a,b) \to \mathbb{R}$ be a function defined on the open interval $(a,b) \subset \mathbb{R}$. The notion of f being differentiable at a point $x_0 \in (a,b)$ can be glanced from the drawing below



where we informally let x approach x_0 and look at the limiting value of the slope. Newton used to say many hundred years ago, that the derivative of f at x_0 is the value of this slope just before x becomes x_0. In modern day mathematical parlance, this translates into the existence of (a slope) c, such that

$$\lim_{h \to 0} \frac{f(x_0+h) - f(x_0)}{h} = c.$$

We will use the equivalent operational definition below in terms of continuous functions ε defined around 0 with $\varepsilon(0) = 0$. This looks difficult, but it is actually a clever way of approaching differentiability (and perhaps more in the spirit of Newton).

(6.16) DEFINITION.

The function $f : (a,b) \to \mathbb{R}$ is differentiable at $x_0 \in (a,b)$ if there exists

(i) $c \in \mathbb{R}$

(ii) $\delta > 0$ *with* $x_0 - \delta, x_0 + \delta \in (a,b)$ *i.e.,* $a + \delta < x_0$ *and* $x_0 < b - \delta$.

(iii) A function $\varepsilon : (-\delta, \delta) \to \mathbb{R}$ *continuous at* 0 *with* $\varepsilon(0) = 0$,

such that

$$f(x_0 + h) - f(x_0) = ch + \varepsilon(h)h \tag{6.4}$$

for every $h \in (-\delta, \delta)$.
The number c is denoted $f'(x_0)$ and called the derivative *of f at x_0; f is called* differentiable *if it is differentiable at every $x_0 \in (a,b)$.*

LLM

```
Please explain the definition of differentiability given below. Illustrate
by a few example and quiz me afterwards.
'''
The function $f: (a, b)\rightarrow \RR$ is differentiable at $x_0\in (a, b)$
if there exists
\begin{enumerate}[(i)]
\item
$c\in \RR$
\item
$\delta > 0$ with $x_0 - \delta, x_0 + \delta\in (a, b)$ i.e.,
$a + \delta < x_0$ and $x_0 < b-\delta$.
\item
A function $\epsilon: (-\delta, \delta) \rightarrow 0$ continuous at $0$
with $\epsilon(0) = 0$,
\end{enumerate}
such that
\begin{equation}\label{operational}
f(x_0 + h) - f(x_0) = c h + \epsilon(h) h
\end{equation}
for every $h\in (-\delta, \delta)$.

The number $c$ is denoted $f'(x_0)$ and called \emph{the derivative} of
$f$ at $x_0$; $f$ is called \emph{differentiable} if
it is differentiable at every $x_0\in (a, b)$.
'''
```

(6.17) REMARK.

If a function $f:(a,b)\to\mathbb{R}$ is differentiable, we get a new function $f':(a,b)\to\mathbb{R}$ giving the (first) derivative at a point as output. We may ask again if this function is differentiable. If this is so, we may define a function $f'':(a,b)\to\mathbb{R}$ given by $f''(x)=(f')'(x)$ called the second derivative. This procedure may be continued. We use the notation $f^{(n)}$ for the n-th derivative.

(6.18) EXAMPLE.

Let us apply Definition 6.16 to the function $f(x)=x^2$ at the point x_0. Here

$$f(x_0+h)-f(x_0)=(x_0+h)^2-x_0^2=2x_0h+h^2.$$

Here you immediately see that $c=f'(x_0)=2x_0$ with $\varepsilon(h)=h$ (and $\delta=\infty$) in Definition 6.16. ♠

(6.19) EXERCISE.

Use Definition 6.16 to formally show that $f'(x)=3x^2$ if $f(x)=x^3$. ♠

A differentiable function is continuous as is shown in the following result.

(6.20) PROPOSITION.

If the function $f:(a,b)\to\mathbb{R}$ is differentiable at $x_0\in(a,b)$, then it is continuous at x_0.

Proof. That f is continuous at x_0 means (recall Definition 5.48) that to every $\varepsilon>0$, we may find $\delta>0$ so that

$$|x-x_0|<\delta \implies |f(x)-f(x_0)|<\varepsilon. \tag{6.5}$$

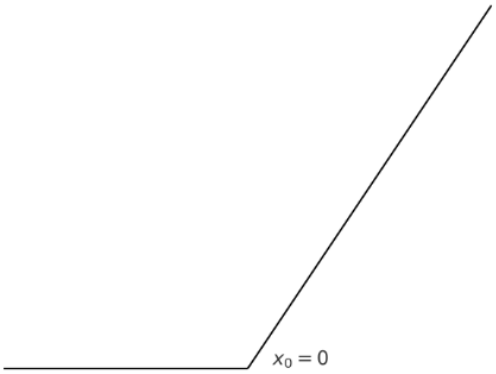
We are assuming that f is differentiable at x_0, so according to Definition 6.16, there exists a number c so that (with $h=x-x_0$)

$$|f(x)-f(x_0)|=|(c+\varepsilon(x-x_0))(x-x_0)|.$$

I will not write every detail out here, but you can see from the formula above that $|f(x)-f(x_0)|<M|x-x_0|$ for some number M, when $|x-x_0|$ is sufficiently small. This gives a δ that can be used in (6.5). □

(6.21) EXAMPLE.

The ReLu function $f(x)=\max(0,x)$ is an example of a function, which is continuous, but not differentiable at $x_0=0$. This is much related to its sharp corner there.



As mentioned in these notes, the ReLu function plays a prominent role as an activation function in neural networks.

♠

(6.22) EXERCISE.

Show precisely that the ReLu function is not differentiable at 0. ♠

6.3.2 Formulas

In operating with differentiable functions you are supposed to draw on your previous knowledge. I have summarized some of this knowledge below (even though we will give hints below as how to prove some of the rules).

1. If $f(x) = ag(x)$, where $a \in \mathbb{R}$, then
$$f'(x) = ag'(x)$$
.
2. If $f(x) = x^n$, where $n \in \mathbb{N}$, then
$$f'(x) = nx^{n-1}.$$
3. If $f(x) = e^x$, then
$$f'(x) = f(x) = e^x.$$
4. If $f(x) = \log(x)$, then
$$f'(x) = 1/x.$$
Here $\log(x)$ denotes the logarithm with base e.
5. If $f(x) = \sin(x)$, then
$$f'(x) = \cos(x).$$
6. If $f(x) = \cos(x)$, then
$$f'(x) = -\sin(x).$$
7. If $f(x)$ and $g(x)$ are differentiable functions, then the derivative of their product is
$$(fg)'(x) = f'(x)g(x) + f(x)g'(x).$$

8. If $f(x)$ and $g(x)$ are differentiable functions, then the derivative of their quotient is

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}.$$

9. If $f(x)$ and $g(x)$ are composable differentiable functions, then the derivative of their composite is

$$(f \circ g)'(x) = f'(g(x))g'(x).$$

(6.23) EXERCISE.

Suppose that $f(x) = \sin(x)$. What is

$$f^{(17)}(x)?$$

♠

6.3.3 The derivative of a product

From high school you know that the derivative of a product of two functions f and g is given by the formula

$$(fg)'(x) = f'(x)g(x) + f(x)g'(x). \tag{6.6}$$

We can use the ε-definition (6.4) to derive the product rule in (6.6). The computation below is a bit cumbersome, but actually quite doable. We assume to begin with that f and g are differentiable at x_0 according to (6.4) i.e.,

$$\begin{aligned} f(x_0+h) &= f(x_0) + f'(x_0)h + \varepsilon_f(h)h \\ g(x_0+h) &= g(x_0) + g'(x_0)h + \varepsilon_g(h)h. \end{aligned}$$

Then we start the computation:

$$\begin{aligned} &(fg)(x_0+h) = f(x_0+h)g(x_0+h) = \\ &(f(x_0) + f'(x_0)h + \varepsilon_f(h)h)\ (g(x_0) + g'(x_0)h + \varepsilon_g(h)h) = \\ &f(x_0)g(x_0) + (f'(x_0)g(x_0) + f(x_0)g'(x_0))h + \varepsilon(h)h, \end{aligned} \tag{6.7}$$

where the function

$$\varepsilon(h) = f(x_0)\varepsilon_g(h) + f'(x_0)g'(x_0)h + f'(x_0)\varepsilon_g(h)h + \varepsilon_f(h)g'(x_0) + \varepsilon_f(h)g'(x_0)h + \varepsilon_f(h)\varepsilon_g(h)h \tag{6.8}$$

is seen to be continuous at $h = 0$ with $\varepsilon(0) = 0$. The end result of this computation shows that fg is differentiable at x_0 with

$$(fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0) \tag{6.9}$$

again according to (6.4).

(6.24) EXERCISE.

Show that the ε function defined in (6.8) satisfies the relevant conditions in Definition 6.16. ♠

The formula for the derivative of a fraction i.e.,

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$$

can be derived using a neat little trick. This is the topic of the following exercise.

(6.25) EXERCISE.

Show how the product rule may be used to derive the rule for finding the derivative of a fraction:

$$\left(\frac{f}{g}\right)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{g(x_0)^2}.$$

Hint:

$$f'(x) = \left(g(x)\left(\frac{f(x)}{g(x)}\right)\right)'.$$

♠

6.3.4 The one variable chain rule

The formula for the derivative of a composite function is given by

$$(f \circ g)'(x_0) = f'(g(x_0))g'(x_0),$$

where $g(x_0)$ is in the domain of f. Let us see how (6.4) applies in showing this.

Suppose that f is differentiable at $g(x_0)$ and g is differentiable at x_0, then we can mess around a bit with the ε-functions for f and g for the composite function $f(g(x))$ around x_0:

$$\begin{aligned} f(g(x_0+h)) &= f(g(x_0) + g'(x_0)h + h\varepsilon_g(h)) \\ &= f(g(x_0)) + f'(g(x_0))g'(x_0)h + \varepsilon(h)h, \end{aligned}$$

where (take a deep breath)

$$\varepsilon(h) = f'(g(x_0))\varepsilon_g(h) + \varepsilon_f(g'(x_0)h + \varepsilon_g(h)h)g'(x_0).$$

Here ε is seen to be continuous at 0 with $\varepsilon(0) = 0$ i.e., the composition $f(g(x))$ is differentiable at x_0 with derivative

$$(f \circ g)'(x_0) = f'(g(x_0))g'(x_0). \tag{6.10}$$

The formula (6.10) is extremely important and useful. We give some applications in the exercises below.

(6.26) EXERCISE.

For the function $f(x) = x^n$ for $n \in \mathbb{N}$, you already know that $f'(x) = nx^{n-1}$. Show that if you define the function $g : \{x \in \mathbb{R} \mid x > 0\} \to \mathbb{R}$ by

$$g(x) = x^a := e^{\log(x)a},$$

for an arbitrary number $a \in \mathbb{R}$, then $g'(x) = ax^{a-1}$. ♠

(6.27) EXERCISE.

Compute the derivative of the function $f : (0, \pi) \to \mathbb{R}$ given by

$$f(x) = \frac{1}{\sqrt{\sin(x)}}$$

using only paper and pencil! You can check your result afterwards using a computer. ♠

(6.28) EXERCISE.

Suppose that g and f are inverse functions i.e.,

$$f(g(x)) = x \qquad \text{and} \qquad g(f(x)) = x.$$

If you know the derivative of f, how can you use the chain rule to get the derivative of g? Illustrate with examples like $f(x) = x^2$ and $g(x) = \sqrt{x}$, $f(x) = e^x$ and $g(x) = \log(x)$. ♠

(6.29) EXERCISE.

Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a convex function. We know that f is continuous, but is f differentiable at every point $x_0 \in \mathbb{R}$?

Hint: Nope. This is wrong. Come up with a convex function f and a point x_0, such that f is not differentiable at x_0. ♠

6.3.5 The Newton-Raphson method for finding roots

We begin this section with a surprising example.

(6.30) EXAMPLE.

Suppose that $a > 0$ and we wish to compute $\sqrt{a}$. To do this we may focus on the quadratic equation $f(x) = x^2 - a = 0$ and attempt to compute an approximate value $x_0 \geq 0$, such that $f(x_0)$ is close to 0. Let me at this point disclose that there is a very effective iterative scheme for doing this. You start by putting $x_0 = a$ and then iterate using the formula

$$x_{i+1} = \frac{1}{2}\left(x_i + \frac{a}{x_i}\right) \tag{6.11}$$

to get better and better approximations $x_0, x_1, x_2, \ldots$ to $\sqrt{a}$.

The formula in (6.11) is derived from

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

where $f(x) = x^2 - a$.

You can try out (6.11) below.

Interactive code not included in static version.

♠

I have been in complete awe of the Newton-Raphson method since my early youth. It is an algorithm, where the notion of differentiability really shines.

The method comes from Definition 6.16 with $h = x - x_0$: we are assuming that x_0 is very close to x, where $f(x) = 0$. Then

$$f(x) - f(x_0) = f'(x_0)(x - x_0) + \text{a very small number.}$$

Ignoring the very small number and solving this equation for x we get

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

In the Sage window below, I have entered the algorithm starting in $x_0 = 0$ running ten iterations for finding a zero for $f(x) = \cos(x) - x$.

Graph:

Interactive code not included in static version.

Interactive code not included in static version.

(6.31) EXERCISE.

Give an example, where the Newton-Raphson method cycles between points and never finds the desired zero. Perhaps a drawing will help here. ♠

The Newton-Raphson converges rapidly in most cases. Of course, it breaks down violently if it runs into a *critical point* i.e., a point x, such that $f'(x) = 0$.

Below is some interactive Sage code for experimenting with Newton's method.

Interactive code not included in static version.

(6.32) EXAMPLE.

The formula (see button in Example 1.84) for the (monthly) payment Y on a (car) loan over N payments with a down payment of P and an interest rate of r (per payment or term) is given by the formula

$$Y = \frac{rP}{1 - \left(\frac{1}{1+r}\right)^N}.$$

There is no explicit formula for calculating r given Y, P and N. Here the Newton-Raphson method is invaluable for estimating r by approximating a zero for the function

$$r(x) = Y - \frac{xP}{1 - \left(\frac{1}{1+x}\right)^N}.$$

♠

(6.33) EXERCISE.

Your bank promises you a loan of 1.000.000 DKK with yearly payments of 45.000 DKK over 30 years. At the same time it claims that its interest rate is very favorable at only 1.0%. Here the bank is wrong! What is the real interest rate? How much money do you save (compared to the original offer from the bank) if you insist that the bank offers you the promised interest rate of 1.0%? ♠

6.3.6 Critical points and extrema

(6.34) DEFINITION.

A critical point for a differentiable function $f : (a,b) \to \mathbb{R}$ is a point $x_0 \in (a,b)$ with

$$f'(x_0) = 0.$$

The crucial result here is the following. It seems to date back to Fermat (see Fermat's theorem).

(6.35) LEMMA.

Let $f : (a,b) \to \mathbb{R}$ be a differentiable function. If x_0 is a local extremum for f, then x_0 is critical point i.e., $f'(x_0) = 0$.

Proof. Suppose that ξ is a local maximum and that

$$f(\xi + h) - f(\xi) = f'(\xi)h + \varepsilon(h)h$$

according to (6.4). If $f'(\xi) > 0$, then we can choose $\delta > 0$ sufficiently small, such that $|\varepsilon(h)| < f'(x_0)$ if $0 \leq h < \delta$, since $\varepsilon(0) = 0$ and ε is continuous in 0. Therefore

$$f(\xi + h) - f(\xi) = (f'(\xi) + \varepsilon(h))h > 0,$$

contradicting that ξ is a local maximum. The proof is similar for $f'(\xi) < 0$ and if ξ is a local minimum. □

(6.36) EXERCISE.

Is the converse of the above lemma true i.e., if $f'(x_0) = 0$ is x_0 a local extremum? ♠

Theorem 6.37 below is called the mean value theorem. It is a consequence of Lemma 6.35 and the extremely important Theorem 5.66 about continuous functions on compact subsets attaining their maxima and minima!

(6.37) THEOREM.

Let $f : [a,b] \to \mathbb{R}$ be continuous and differentiable on (a,b). Then there exists $x_0 \in (a,b)$ such that

$$f'(x_0) = \frac{f(b) - f(a)}{b - a}.$$

6.3.7 Increasing functions

The definition below is much simpler than the definition of differentiability.

(6.38) DEFINITION.

A function $f : S \to \mathbb{R}$ *with* $S \subseteq \mathbb{R}$ *is called* increasing *if*

$$x \leq y \Rightarrow f(x) \leq f(y)$$

and strictly increasing *if*

$$x < y \Rightarrow f(x) < f(y)$$

for $x, y \in S$.

LLM

```
Explain the definition below to me. Give some examples and test me.
\begin{definition}
A function $f:S\rightarrow \mathbb{R}$ with $S\subseteq \mathbb{R}$ is called
  \emph{increasing} if
  \begin{equation*}
    x\leq y\Rightarrow f(x) \leq f(y)
  \end{equation*}
  and \emph{strictly increasing} if
  \begin{equation*}
    x <  y\Rightarrow f(x) < f(y)
  \end{equation*}
  for $x, y\in S$.
\end{definition}
```

(6.39) EXERCISE.

Give an example of an increasing function. Give an example of an increasing function that is not strictly increasing. ♠

The following very important result is a consequence of Theorem 6.37. You probably already know this result from your previous (danish) education (monotoniforhold!).

(6.40) PROPOSITION.

Let $f : (a,b) \to \mathbb{R}$ *be a differentiable function. Then* f *is increasing if and only if* $f'(x) \geq 0$ *for every* $x \in (a,b)$. *If* $f'(x) > 0$ *for every* $x \in (a,b)$, *then* f *is strictly increasing.*

(6.41) EXERCISE.

Which of the properties below are true for the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = x^3 + 2x^2 + x + 1.$$

(a) It is differentiable.

(b) It is continuous.

(c) It has a global minimum.

(d) It has a global maximum.

(e) It has exactly one critical point.

(f) It has a local maximum.

(g) It has a local minimum.

(h) It is increasing.

(i) It has three zeros.

(j) Its derivative has two zeros.

(k) It is convex.

♠

(6.42) EXERCISE.

Show that $f(x) = x^3$ is strictly increasing i.e.,

$$x < y \implies x^3 < y^3.$$

Hint:

$$y^3 - x^3 = (y-x)(y^2+xy+x^2),$$

but why is y^2+xy+x^2 always > 0 except when $x = y = 0$? ♠

(6.43) EXERCISE.

Suppose that $f : [a,b] \to \mathbb{R}$ is a continuous function, such that f is differentiable on the open interval (a,b). Is f increasing on $[a,b]$ if $f'(x) \geq 0$ for every $x \in (a,b)$? ♠

(6.44) EXERCISE.

Is it possible for a strictly increasing function $f : \mathbb{R} \to \mathbb{R}$ to be bounded i.e., does there exist a (positive) number M, such that $|f(x)| \leq M$ for every $x \in \mathbb{R}$?

Hint: Have a look at

$$f(x) = \frac{1}{1+e^{-x}}.$$

♠

6.4 Taylor polynomials

If x_0 is a critical point for f we cannot conclude that x_0 is a local extremum. We know that $f'(x_0) = 0$ and we can get more information out of f by exploring the signs of

$$f''(x_0), f'''(x_0), \ldots$$

Suppose that

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

is a polynomial, then

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n. \tag{6.12}$$

For nice functions like $f(x) = e^x$ we can play this game ad infinitum. In fact in this way we get the beautiful infinite series

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots + \frac{x^n}{n!} + \cdots.$$

If f is an n times differentiable function defined at 0, we call the polynomial in (6.12) the Taylor polynomial about the point 0 of degree n associated with the f. Similarly, one may also define the Taylor polynomial of order n about a point a by

$$f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n.$$

Taylor polynomials can be used to approximate more complicated functions such as $\cos(x)$ and $\sin(x)$ with a well defined error term. This is cool classical mathematics. Unfortunately we do not have time to go deeper into Taylor's theorem, which states this in precise terms.

(6.45) EXERCISE.

Compute the Taylor polynomial for $f(x) = \cos(x)$ up to degree 10. ♠

(6.46) EXERCISE.

Suppose you have a number i that satisfies

$$i^2 = -1.$$

Can you make sense of the formula

$$e^{ix} = \cos(x) + i\sin(x)$$

using Taylor polynomials? ♠

In the context of optimization, the following result becomes important. We will not give the proof, but only notice that Theorem 6.37 also here plays an important role.

(6.47) THEOREM.

Let x_0 be a critical point of an $n+1$ times differentiable function $f:(a,b)\to\mathbb{R}$, such that $f^{(n+1)}$ is a continuous function,

$$\begin{aligned} f''(x_0) &= 0 \\ f'''(x_0) &= 0 \\ &\vdots \\ f^{(n-1)}(x_0) &= 0 \end{aligned}$$

and $f^{(n)}(x_0)\neq 0$. If n is even, then x_0 is a local minimum if $f^{(n)}(x_0)>0$ and a local maximum if $f^{(n)}(x_0)<0$. If n is odd, then x_0 is not a local extremum.

(6.48) EXAMPLE.

Let us apply Theorem 6.47 to the function

$$f(x)=ax^2+bx+c,$$

where $a\neq 0$. Here $f'(x)=2ax+b$ and

$$x_0=-\frac{b}{2a}$$

is a critical point (why?). Since

$$f''(x_0)=2a,$$

we see that x_0 is a local minimum if $a>0$ and a local maximum if $a<0$. ♠

(6.49) EXERCISE.

Have you seen Example 6.48 elsewhere, perhaps in a more geometric setting? What type of curve is the graph of $f(x)$? Here you may consult your previous mathematical knowledge.

What is the outcome, when you apply Theorem 6.47 to the function $f(x)=x^3$ at $x_0=0$? ♠

(6.50) EXERCISE.

Show that $x_0=0$ is a critical point of the function $f:(-\frac{1}{2},\infty)\to\mathbb{R}$ defined by

$$f(x)=e^x+\log(1+2x)-3x.$$

Use Theorem 6.47 in deciding if it is a local maximum or minimum or neither.

♠

6.5 Differentiable convex functions

The following theorem is proved using Lemma 6.14 and Theorem 6.37. It immediately implies Corollary 6.52, which is the result mostly used.

(6.51) THEOREM.

Let $f : (a,b) \to \mathbb{R}$ be a differentiable function. Then f is convex if and only if f' is increasing. If f' is strictly increasing, then f is strictly convex.

Theorem 6.51 leads to the following all important result.

(6.52) COROLLARY.

Let $f : (a,b) \to \mathbb{R}$ be a twice differentiable function. Then f is convex if and only if $f''(x) \geq 0$ for every $x \in (a,b)$. If $f''(x) > 0$ for every $x \in (a,b)$, then f is strictly convex.

(6.53) REMARK.

Wait! Stop! Why did I not write $f''(x) > 0$ if and only if f is strictly convex?

(6.54) EXERCISE.

Which of the properties below are true for the function $f(x) = x^3$?

(a) It is convex on $[0,1]$.

(b) It is strictly convex on $(0,1]$.

(c) It is strictly convex on $[0,1]$.

(d) It is convex on $(-1,1)$.

(e) Since $f'(0) = 0$, it must have a local minimum for $x = 0$.

♠

(6.55) EXERCISE.

You cannot deduce from Corollary 6.52 that the function $g : \mathbb{R} \to \mathbb{R}$ given by $g(x) = x^4$ is a strictly convex function. Why not?

You can deduce from Corollary 6.52 that $f(x) = x^2$ is a strictly convex function. How can $g(x) = f(x)^2$ be used to prove that $g(x)$ is a strictly convex function? ♠

(6.56) EXERCISE.

Show that $f(x) = e^x$ is a strictly convex function $f : \mathbb{R} \to \mathbb{R}$.

Show that $f(x) = -\log(x)$ is a strictly convex function $f : (0,\infty) \to \mathbb{R}$. ♠

(6.57) EXERCISE.

Show that $f : \{x \in \mathbb{R} \mid x \geq 0\} \to \mathbb{R}$ given by

$$f(x) = -\sqrt{x}$$

is a strictly convex function. ♠

Another nice application of Lemma 6.14 (and Theorem 6.51) is the following.

(6.58) THEOREM.

Let $f : (a,b) \to \mathbb{R}$ be a differentiable function. Then f is convex if and only if

$$f(y) \geq f(x) + f'(x)(y-x)$$

for every $x, y \in (a,b)$.

(6.59) EXERCISE.

Suppose that $f : (a,b) \to \mathbb{R}$ is a differentiable convex function and $x_0 \in (a,b)$ is a critical point for f. What can you say about x_0 using Theorem 6.58? ♠

Chapter 7

Several variables

A function of several variables usually refers to a function

$$f : \mathbb{R}^n \to \mathbb{R}, \tag{7.1}$$

where $n > 1$ is a natural number. We have already seen functions of several variables with $n > 1$. In particular, in Chapter 4, we saw linear functions (in connection with linear programming) like

$$f(x_1, x_2) = 3x_1 + 2x_2. \tag{7.2}$$

This is a rather simple function of several variables with $n = 2$ in (7.1). In general functions as in (7.1) can be wildly complicated. One of the main purposes of this chapter is to zero in on the class of differentiable functions in (7.1). In Chapter 6 we defined what it means for a function of one variable to be differentiable. This was inspired by a drawing of the graph of the function. In several variables (for $n > 1$) one has to be a bit clever in the definition of differentiability. The upshot is that the derivative at a point now is a row vector (or more generally a matrix) instead of being a single number. As an example, using notation that we introduce in this chapter, the derivative of the function in (7.2) at $(0,0)$ is

$$\left(\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \right) = (3 \quad 2).$$

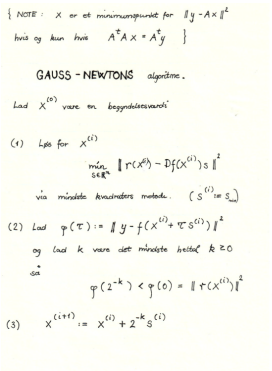
This notation means that partial differentiation with respect to a variable occurs i.e., one fixes the variable and computes the derivative with respect to this variable viewing all the other variables as constants.

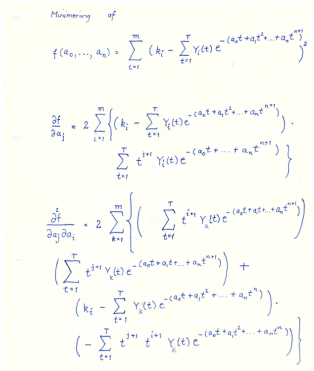
First some treasured memories from the author's past.

7.1 Introduction

Many years ago (1986-89), I had a job as a financial analyst in a bank (now a hotel!) working (often late at night) with a spectacular view of Copenhagen from the office circled below.

This was long before a financial analyst became a quant and machine learning became a buzz word. Digging through my old notes from that time, I found the outlines below.





These were notes I made in connection with modelling the yield curve for zero coupon bonds. I had to fit a very non-linear function in *several variables* to financial data and had to use effective numerical tools (and programming them[1] in APL). Tools that are also used today in machine learning and data science.

Ultimately we are interested in solving optimization problems like

$$\begin{aligned} &\text{Minimize} && f(x_1,\dots,x_n)\\ &\text{with constraint} && \\ & && (x_1,\dots,x_n)\in C, \end{aligned}$$

where $C\subseteq\mathbb{R}^n$ and $f:\mathbb{R}^n\to\mathbb{R}$ is a differentiable (read nice for now) function.

Training neural networks is a fancy name for solving an optimization problem, where usually $C=\mathbb{R}^n$ and f is built just like in the least squares method from some data points. The difference is that in neural networks, f is an incredibly complicated (differentiable) function composed of several intermediate functions. We do not, as in the method of least squares, have an explicit formula for finding a minimum. We have to rely on iterative methods. One such method is called *gradient descent*.

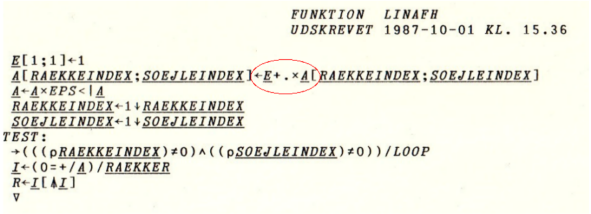
Let me illustrate this in the simplest case, where $n=1$. The general case is conceptually very similar (see Lemma 7.19).

Suppose that f is differentiable at x_0 with $f'(x_0)\neq 0$ and we wish to solve the minimization problem

$$\begin{aligned} &\text{Minimize} && f(x)\\ &\text{with constraint} && \\ & && x\in\mathbb{R}. \end{aligned}$$

1

Programmering i APL



Solving the equation $f'(x) = 0$ (to find potential minima) may be difficult. Instead we try something else.

We know for sure that x_0 is not a local minimum (why?). It turns out that we can move a little bit in the direction[2] of $-f'(x_0)$ and get a better candidate for a minimum than x_0 i.e., for small $\lambda > 0$ and $h = -\lambda f'(x_0)$ we have

$$f(x_0 + h) - f(x_0) < 0.$$

This is a consequence[3] of the definition of f being differentiable at x_0 with $f'(x_0) \neq 0$.

The process is then repeated putting $x_0 := x_0 + h$ until the absolute value of $f'(x_0)$ is sufficiently small (indicating that we are close to a point x with $f'(x) = 0$).

The number $\lambda > 0$ is called the learning rate in machine learning.

(7.1) EXERCISE.

Illustrate the gradient descent method for $f(x) = x^2$. Pay attention to the learning rate $\lambda > 0$. How big is λ allowed to be, when

$$f(x_0 + h) - f(x_0) < 0$$

is required and $h = -\lambda f'(x_0)$? ♠

(7.2) EXERCISE.

This is a hands-on exercise: carry out the gradient descent method numerically for the function

$$f(x) = (x-1)^4 + \sin(x)^2$$

to solve the minimization problem

Minimize	$f(x)$
with constraint	
	$x \in \mathbb{R}$

starting with $x_0 = 1$.

Hint: It is not clear how to choose the step size here. Proceed by letting k be the smallest natural number, such that

$$f(x_0 - 2^{-k} f'(x_0)) < f(x_0).$$

Stop the process, when $|f'(x_0)| < 0.001$.

Helpful code:

Interactive code not included in static version.

Is f a convex function?

[2] Left if $f'(x_0) > 0$ and right if $f'(x_0) < 0$.

[3] If you use the definition of differentiability with $h = -\lambda f'(x_0)$, you will see that

$$f(x_0 + h) - f(x_0) = -\lambda (f'(x_0)^2 + \varepsilon(-\lambda f'(x_0)) f'(x_0)).$$

For small $\lambda > 0$ this shows that $f(x_0 + h) - f(x_0) < 0$, as $f'(x_0)^2 > 0$.

Explain how the Newton-Raphson method[4] may be used to solve the minimization problem and compute the minimum also using this method.

Helpful code:

Interactive code not included in static version.

Interactive code not included in static version.

♠

Recall the definition of a function $f : \mathbb{R} \to \mathbb{R}$ being differentiable at a point $x_0 \in \mathbb{R}$ with derivative $c = f'(x_0)$. Here we measured the change $f(x_0+h) - f(x_0)$ of f in terms of the change h (in x). It had to have the form

$$f(x_0+h) - f(x_0) = ch + \varepsilon(h)h, \tag{7.3}$$

where $\varepsilon : (-\delta, \delta) \to \mathbb{R}$ is a function continuous in 0 with $\varepsilon(0) = 0$ and $\delta > 0$ small. If you divide both sides of (7.3) by h you recover the usual more geometric definition of differentiability as a limiting slope:

$$\lim_{h\to 0} \frac{f(x_0+h) - f(x_0)}{h} = c = f'(x_0). \tag{7.4}$$

We wish to define differentiability at $x_0 \in \mathbb{R}^n$ for a function $f : \mathbb{R}^n \to \mathbb{R}^m$. In this setting the quotient

$$\frac{f(x_0+h) - f(x_0)}{h}$$

in (7.4) does not make any sense. There is no way we can divide a vector $f(x_0+h) - f(x_0) \in \mathbb{R}^m$ by a vector $h \in \mathbb{R}^n$, unless of course $m = n = 1$ as in (7.4), where we faced usual numbers.

The natural thing here is to generalize the definition in (7.3). First let us recall what functions $f : \mathbb{R}^n \to \mathbb{R}^m$ look like.

7.2 Vector functions

We will flesh out the general Definition 1.101 in a special case below.

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ takes a vector $(x_1, \ldots, x_n) \in \mathbb{R}^n$ as input and gives a vector $(y_1, \ldots, y_m) \in \mathbb{R}^m$ as output. This means that every coordinate $y_1, \ldots, y_m$ in the output must be a function of $x_1, \ldots, x_n$ i.e.,

$$y_i = f_i(x_1, \ldots, x_n)$$

for $i = 1, \ldots, m$. So in total, we may write f as

$$f(x_1, \ldots, x_n) = \begin{pmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_m(x_1, \ldots, x_n) \end{pmatrix}. \tag{7.5}$$

Each of the (coordinate) functions f_i are functions from $\mathbb{R}^n$ to $\mathbb{R}$.

[4] This is an iterative method for approximating a zero for a differentiable function $g(x)$. It works by guessing x_0 and then iterating $x_{i+1} = x_i - g(x_i)/g'(x_i)$ to get a sequence $x_0, x_1, \ldots$ approximating a zero z ($g(z) = 0$).

(7.3) EXERCISE.

Look back at Exercise 1.117. Write down precisely the vector function $h : \mathbb{R}^2 \to \mathbb{R}^2$ occuring there. ♠

(7.4) EXERCISE.

The function $f : \mathbb{R}^2 \to \mathbb{R}^2$ is rotating a vector 90 degrees counter clockwise. What are f_1 and f_2 in

$$f(x,y) = \begin{pmatrix} f_1(x,y) \\ f_2(x,y) \end{pmatrix}?$$

Hint: Try rotating some specific vectors like $(1,0),(0,1),(1,1)$ 90 degrees. Do you see a pattern? ♠

7.3 Differentiability

The definition of differentiability for a function $f : \mathbb{R}^n \to \mathbb{R}^m$ mimics (7.3), except that $\varepsilon(h)h$ is replaced by $\varepsilon(h)|h|$. Also the open interval (a,b) is replaced by an open subset U and the (open) interval $(-\delta,\delta)$ is replaced by an open subset O containing 0.

Notice, however, that now the derivate is a matrix!

(7.5) DEFINITION.

Let $f : U \to \mathbb{R}^m$ be a function with $U \subseteq \mathbb{R}^n$ an open subset. Then f is differentiable *at $v_0 \in U$ if there exists*

(i) an $m \times n$ matrix C,

(ii) an open subset $O \subseteq \mathbb{R}^n$ with $0 \in O$, such that $v_0 + h \in U$ for every $h \in O$,

(iii) a function $\varepsilon : O \to \mathbb{R}^m$ continuous at 0 with $\varepsilon(0) = 0$,

such that

$$f(v_0+h) - f(v_0) = Ch + \varepsilon(h)\,|h|\,,$$

In this case, the $m \times n$ matrix C is called the (matrix) derivative of f at x_0 and denoted by $f'(x_0)$. The function f is called differentiable if it is differentiable at every $v \in U$.

How do we compute the matrix derivative C in the above definition? We need to look at the representation of f in (7.5) and introduce the partial derivatives.

7.3.1 Partial derivatives

A function of one variable x has a derivative with respect to x. For a function of several variables $x_1,\ldots,x_n$ we have a well defined derivative with respect to each of these variables. These are called the partial derivatives (if they exist) and they are defined below.

(7.6) DEFINITION.

Let $f : U \to \mathbb{R}$ be a function, where U is an open subset of $\mathbb{R}^n$. Fix a point $v = (a_1, a_2, \ldots, a_n) \in U$ and let

$$p_i = f(a_1, \ldots, a_{i-1}, x, a_{i+1}, \ldots, a_n)$$

for $i = 1, \ldots, n$. If p_i is differentiable at $x = a_i$ according to Definition 7.3, then we say that the partial derivative *of f with respect to x_i exists at $v \in U$ and use the notation*

$$\frac{\partial f}{\partial x_i}(v) := p_i'(a_i).$$

(7.7) REMARK.

The partial derivative with respect to a specific variable is computed by letting all the other variables appear as constants.

To get a feeling for the definition and computation of partial derivatives, take a look at the example below, where we compute using the classical (geometric) definition of the one variable derivative.

(7.8) EXAMPLE.

Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

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

Then

$$\begin{aligned}\frac{\partial f}{\partial x_2}(v) &= \lim_{\delta \to 0} \frac{f(x_1, x_2 + \delta) - f(x_1, x_2)}{\delta} \\ &= \lim_{\delta \to 0} \frac{x_1(x_2 + \delta)^2 + x_1 - (x_1 x_2^2 + x_1)}{\delta} \\ &= x_1 \lim_{\delta \to 0} \frac{(x_2 + \delta)^2 - x_2^2}{\delta} = x_1 \lim_{\delta \to 0} (2x_2 + \delta) = 2x_1 x_2,\end{aligned}$$

where $v = (x_1, x_2)$. This example illustrates that $\frac{\partial f}{\partial x_i}$ can be computed just like in the one variable case, when the other variables ($\neq x_i$) are treated as constants. Notice that

$$\frac{\partial}{\partial x_1}\frac{\partial f}{\partial x_2} = \frac{\partial}{\partial x_2}\frac{\partial f}{\partial x_1} = 2x_2.$$

♠

Partial derivatives behave almost like the usual derivatives of one variable functions. You simply fix one variable that you consider the "real" variable and treat the other variables as constants.

(7.9) EXAMPLE.

$$\frac{\partial}{\partial x}\left(\sin(xy)+x^2y^2+y\right)=y\cos(xy)+2xy^2.$$

♠

Below are examples of Sage code computing partial derivatives. Notice that the variables must be declared first.

Interactive code not included in static version.

The Sage computations above point to a really surprising result. It seems that it makes no difference if you compute the partial derivative with respect to x_1 and then with respect to x_2 or the other way around. You could, just for fun, try this out on the more complicated function

$$f(x_1,x_2)=x_1x_2^2+\cos(\sin(x_1x_2)+\log(x_1^{17}x_2)).B$$

This result is formulated in Theorem 7.13 below.

(7.10) EXERCISE.

Use the Sage window above to verify the computation of the partial derivative in Example 7.9. ♠

The following result tells us how to compute the matrix derivative.

(7.11) PROPOSITION.

Let $f:U\to\mathbb{R}^m$ be a function with $U\subseteq\mathbb{R}^n$ an open subset. If f is differentiable at $x_0\in U$, then the partial derivatives

$$\frac{\partial f_i}{\partial x_j}(x_0)$$

exist for $i=1,\ldots,m$ and $j=1,\ldots,n$ and the matrix C in Definition 7.5 is

$$C=\begin{pmatrix}\frac{\partial f_1}{\partial x_1}(x_0) & \cdots & \frac{\partial f_1}{\partial x_n}(x_0)\\ \vdots & \ddots & \vdots\\ \frac{\partial f_m}{\partial x_1}(x_0) & \cdots & \frac{\partial f_m}{\partial x_n}(x_0)\end{pmatrix}.$$

Proof. The j-th column in C is Ce_j. Putting $h=\delta e_j$ for $\delta\in\mathbb{R}$ in Definition 7.5 gives

$$f(x_0+\delta e_j)-f(x_0)=\delta Ce_j+\varepsilon(\delta e_j)\,|\delta|\,.$$

The i-th coordinate of this identity of m-dimensional vectors can be written

$$f_i(x_0+\delta e_j)-f_i(x_0)=\delta C_{ij}+\tilde{\varepsilon}_i(\delta)\delta \tag{7.6}$$

where

$$\tilde{\varepsilon}_i(\delta)=\begin{cases}\varepsilon_i(\delta e_j)\dfrac{|\delta|}{\delta} & \text{if } \delta\neq 0\\ 0 & \text{if } \delta=0\end{cases}$$

and (7.6) shows that $C_{ij}=\frac{\partial f_i}{\partial x_j}(x_0)$. □

(7.12) EXERCISE.

Compute the matrix derivative of the vector function in Exercise 7.4. ♠

For a function $f : U \to \mathbb{R}$ with $U \subseteq \mathbb{R}^n$ an open subset, the partial derivative, if it exists for every $x \in U$, is a new function

$$\frac{\partial f}{\partial x_j} : U \to \mathbb{R}.$$

We will use the notation

$$\frac{\partial^2 f}{\partial x_i \partial x_j} := \frac{\partial}{\partial x_i}\frac{\partial f}{\partial x_j}$$

for the *iterated (second order) partial derivative*.

The first part of following result is a converse to Proposition 7.11. The second part contains the surprising *symmetry of the second order partial derivatives* under rather mild conditions. We will not go into the proof of this result, which is known as Clairaut's theorem.

(7.13) THEOREM.

Let $f : U \to \mathbb{R}^m$ be a function with $U \subseteq \mathbb{R}^n$ an open subset. If the partial derivatives for f exist at every $x \in U$ with

$$\frac{\partial f_i}{\partial x_j}$$

continuous (for $i = 1, \ldots, m$ and $j = 1, \ldots, n$), then f is differentiable. If the second order partial derivatives exist for a function $f : U \to \mathbb{R}$ and are continuous functions, then

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

for $i, j = 1, \ldots, n$.

(7.14) EXERCISE.

Verify (by hand!) the symmetry of the second order partial derivatives for the function f in Example 7.9 i.e., show that

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}.$$

♠

(7.15) EXERCISE.

Verify that $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x, y) = \frac{x^2 y}{1 + y^2}$$

is a differentiable function by computing

$$\frac{\partial f}{\partial x} \quad \text{and} \quad \frac{\partial f}{\partial y}$$

and applying Theorem 7.13. Check also that

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}.$$

♠

7.4 Newton-Raphson in several variables!

There is a beautiful generalization of the Newton-Raphson method to several variable functions $f : \mathbb{R}^n \to \mathbb{R}^n$. Consider first that you would like to solve the system

$$\begin{aligned} y^2 - x^3 + x &= 0 \\ y^3 - x^2 &= 0 \end{aligned} \tag{7.7}$$

of non-linear equations in the two variables x and y. Notice that we are talking *non-linear* here. This is so much more difficult than the systems of linear equations that you encountered in a previous chapter.

However, just like we used Newton's method in one variable for solving a non-linear equation, Newton's method for finding a zero for a function $f : \mathbb{R}^n \to \mathbb{R}^n$ generalized to the iterative scheme

$$v_{i+1} = v_i - f'(v_i)^{-1} f(v_i) \tag{7.8}$$

provided that the $n \times n$ matrix derivative $f'(v_i)$ is invertible.

The reason that (7.8) works comes again from the powerful definition of differentiability in Definition 7.5 using that

$$f(x) - f(x_0) \qquad \text{is close to} \qquad f'(x_0)(x - x_0) \tag{7.9}$$

provided that $h = x - x_0$ is small. In fact, you get (again) (7.8) from (7.9) by putting $f(x)$ to 0, replacing *is close to* by $=$ and then isolating x.

For the equations in (7.7), the iteration scheme (7.8) becomes

$$\begin{pmatrix} x_{i+1} \\ y_{i+1} \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} -3x_i^2 + 1 & 2y_i \\ -2x_i & 3y_i^2 \end{pmatrix}^{-1} \begin{pmatrix} y_i^2 - x_i^3 + x_i \\ y_i^3 - x_i^2 \end{pmatrix}. \tag{7.10}$$

(7.16) EXERCISE.

Verify the claim in (7.10) by applying (7.8) to

$$f(x,y) = \begin{pmatrix} y^2 - x^3 + x \\ y^3 - x^2 \end{pmatrix}.$$

Carry out sufficiently many iterations starting with the vector $(1,1)$ in (7.10) to see the iteration stabilize. You should do this using a computer, for example by modifying the Sage code in the last half of Example 7.18. ♠

7.5 Local extrema in several variables

For a function $f : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^n$, the derivative $f'(v)$ at $v \in U$ is called *the gradient* for f at v. Classically, it is denoted $\nabla f(v)$ i.e.,

$$\nabla f(v) = \left(\frac{\partial f}{\partial x_1}(v) \dots \frac{\partial f}{\partial x_n}(v) \right).$$

The definition below is inspired by the one variable case (see Definition 6.34).

(7.17) DEFINITION.

Let $f : U \to \mathbb{R}$ be a function, where $U \subseteq \mathbb{R}^n$ is an open subset. Suppose that the partial derivatives exist at $v_0 \in U$. Then v_0 is called a critical point *for f if $\nabla f(v_0) = 0$.*

(7.18) EXAMPLE.

Consider the function $f : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2x + 3\log(y) \\ 3x/y - 3y^2 \end{pmatrix}$$

corresponding to finding critical points for the function

$$g(x,y) = x^2 + 3x\log(y) - y^3. \tag{7.11}$$

You can left click and hold the graph computed below (after it has rendered) and rotate the surface to get a feeling for what (7.11) looks like. Zooming in is also possible.

Interactive code not included in static version.

Here

$$f' = \begin{pmatrix} 2 & 3/y \\ 3/y & -3x/y^2 - 6y \end{pmatrix}.$$

In the Sage code below, Newton's method is started at $(1,1)$ and iterated four times.

Interactive code not included in static version.

♠

If v_0 is not a critical point for f we can use the gradient vector to move in a direction making f strictly smaller/larger. This is very important for optimization problems.

(7.19) LEMMA.

Let $f : U \to \mathbb{R}$ be a differentiable function, where $U \subseteq \mathbb{R}^n$ is an open subset. Suppose that $u \in \mathbb{R}^n$ and $\nabla f(v_0)u < 0$ for $v_0 \in U$. Then

$$f(v_0 + \lambda u) < f(v_0)$$

for $\lambda > 0$ small.

Proof. By the differentiability of f,

$$f(v_0 + u) - f(v_0) = \nabla f(v_0)u + \varepsilon(u)|u|,$$

where $\varepsilon : \mathbb{R}^n \to \mathbb{R}$ is a function satisfying $\varepsilon(h) \to 0$ for $h \to 0$. For $\lambda > 0$ with $v_0 + \lambda u \in U$ we have

$$f(v_0 + \lambda u) - f(v_0) = \lambda(\nabla f(v_0)u + \varepsilon(\lambda u)|u|).$$

When λ tends to zero from the right, it follows that $f(v_0 + \lambda u) - f(v_0) < 0$ for small $\lambda > 0$. □

Lemma 7.19 looks innocent, but it is the bread and butter in the training of neural networks. In mathematical terms, training means minimizing a function. In machine learning terms, λ above is called the *learning rate*. One iteration (why do I choose $u = -\nabla f(x)$?)

$$x - \lambda \nabla f(x)$$

of Lemma 7.19 is the central ingredient in an *epoch* in training a neural network.

(7.20) EXAMPLE.

Let us briefly pause and see Lemma 7.19 in action. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y) = x^2 + y^2$$

and $v_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ with $u = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$. In this case $\nabla f(x_0) = \begin{pmatrix} 2 & 2 \end{pmatrix}$ and $\nabla f(v_0)u = -2 < 0$. Therefore we may find a small $\lambda > 0$, such that $f(v_0 + \lambda u) < f(v_0)$. How do we choose λ optimally? If λ is too big we fail and land up in a worse point than x_0. Here

$$f(v_0 + \lambda u) = (1 - \lambda)^2 + 1$$

This is a quadratic polynomial, which is minimal for $\lambda = 1$. Therefor the minimal value reached in the direction of u is 1. The process now continues replacing v_0 by $v_0 + 1 \cdot u$. ♠

The result below is the multi variable generalization of looking for local extrema by putting $f' = 0$ in the one variable case.

(7.21) PROPOSITION.

Let $f : U \to \mathbb{R}$ be a differentiable function, where $U \subseteq \mathbb{R}^n$ is an open subset. If $v_0 \in U$ is a local extremum, then v_0 is a critical point for f.

Proof. Suppose that $\nabla f(v_0) \neq 0$. If v_0 is a local minimum, then we may use $u = -\nabla f(v_0)$ in Lemma 7.19 to deduce that $f(v_0 + \lambda u) < f(v_0)$ for $\lambda > 0$ small. This contradicts the local minimality of v_0. If v_0 is a local maximum we can apply Lemma 7.19 with $-f$ and $u = \nabla f(v_0)$ to reach a similar contradiction. Therefore $\nabla f(v_0) = 0$ and v_0 is a critical point for f. □

(7.22) EXERCISE.

Compute the critical points of

$$f(x,y) = x^3 + xy + y^3.$$

Is $(0,0)$ a local maximum or a local minimum for f?

Hint: Look at

$$f_1(t) = f(t,t)$$
$$f_2(t) = f(t,-t)$$

and $f_1''(0)$ and $f_2''(0)$ (along with Theorem 6.47). ♠

(7.23) EXERCISE.

We will prove later that a differentiable function $f : \mathbb{R}^2 \to \mathbb{R}$ is strictly convex if the socalled Hessian matrix given by

$$\begin{pmatrix} \frac{\partial^2 f}{\partial x^2}(v) & \frac{\partial^2 f}{\partial x \partial y}(v) \\ \frac{\partial^2 f}{\partial y \partial x}(v) & \frac{\partial^2 f}{\partial y^2}(v) \end{pmatrix}$$

is positive definite for every $v \in \mathbb{R}^2$. This is a multivariable generalization of the fact that $g : \mathbb{R} \to \mathbb{R}$ is strictly convex if $g''(x) > 0$ for every $x \in \mathbb{R}$.

Now let

$$f(x,y) = x^2 + y^2 - \cos(x) - \sin(y). \tag{7.12}$$

3D graph: You can left click the surface computed below after it has rendered and rotate or zoom in.

Interactive code not included in static version.

(a) Show that f is strictly convex.

(b) Compute the critical point(s) of f.

Hint: This is a numerical computation! Modify the relevant Sage window for Newton's method in the previous chapter to do it.

(c) For a differentiable convex function $f : \mathbb{R}^2 \to \mathbb{R}$ we have in general that

$$f(v) \geq f(u) + \nabla f(u)(v-u) \tag{7.13}$$

for every $u, v \in \mathbb{R}^2$. This is a multivariable generalization of Theorem 6.58.

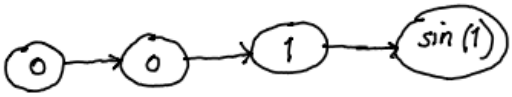
Explain how one can use (7.13) to find a global minimum for the function f in (7.12). Is this minimum unique? Is $f(x,y) \geq -1$ for every $x, y \in \mathbb{R}$?

♠

7.6 The chain rule

Suppose you want to compute the value of the function $h(x) = \sin(\cos(2x))$ for $x = 0$. Then you would start by evaluating the inner function $2x$, then applying cos and finally sin. This computation can be illustrated in the (computational) graph

where you plug $x = 0$ into the leftmost node and fill in each node taking input from its left neighbor



Suppose we want to compute $h'(x)$ for $x = 0$. Can we use the computational graph for this?

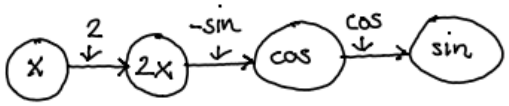
Recall the chain rule for functions of one variable. Here we have functions $f : (a,b) \to \mathbb{R}$ and $g : (c,d) \to \mathbb{R}$, such that $g(x) \in (a,b)$ for $x \in (c,d)$. If g is differentiable at $x_0 \in (c,d)$ and f is differentiable at $g(x_0) \in (a,b)$, the chain rule says that $f \circ g$ is differentiable at x_0 with

$$(f \circ g)'(x_0) = f'(g(x_0))g'(x_0).$$

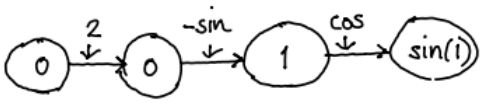
The chain rule tells us that

$$h'(x) = \cos(\cos(2x))(-\sin(2x))2.$$

This expression involves three derivatives corresponding to the three edges in the comptuational graph. We can illustrate the chain rule by labeling each edge with the derivative of its end node:



Then the derivative $h'(0)$ can be computed by as the product of the labels evaluated on their left nodes in the filled in comptutational graph:



i.e., $h'(0) = \cos(1)(-\sin(0))2$. This observation is the basis of the famous backpropagation rule used in training neural networks.

The chain rule for functions of one variable generalizes verbatim to functions of several variables:

$$(f \circ g)'(x_0) = f'(g(x_0))g'(x_0)$$

for compatible multivariate functions f and g when you replace usual multiplication by matrix multiplication.

(7.24) THEOREM.

Let $f : U \to \mathbb{R}^m$ and $g : V \to \mathbb{R}^n$ with $U \subseteq \mathbb{R}^n$, $V \subseteq \mathbb{R}^l$ open subsets and $g(V) \subseteq U$. If g is differentiable at $v_0 \in V$ and f is differentiable at $g(x_0) \in U$, then $f \circ g$ is differentiable at v_0 with

$$(f \circ g)'(v_0) = f'(g(x_0))g'(v_0). \tag{7.14}$$

The proof of the chain rule in this general setting uses Definition 7.5 just as in the one variable case. It is not conceptually difficult, but severely cumbersome. We will not give it here.

7.6.1 Matrix multiplication graphically

To really understand the chain rule, it pays to view the matrix multiplication in (7.14) in a new light (inspired by computer science and neural networks).

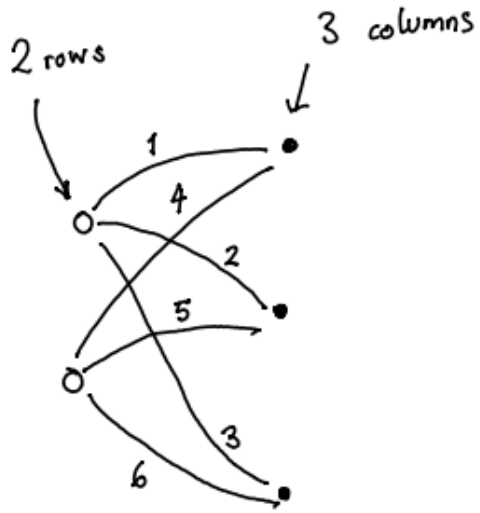
An $m \times n$ matrix (a_{ij}) is a rectangular table with m rows and n columns containing mn numbers. We may also view it as a (bipartite) graph with m left nodes, n right nodes and an edge from the left node i to the right node j with weight a_{ij}. This is best illustrated by an example, which also tells you how matrix multiplication is (beautifully) interpreted in this setting.

(7.25) EXAMPLE.

The 2×3 matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

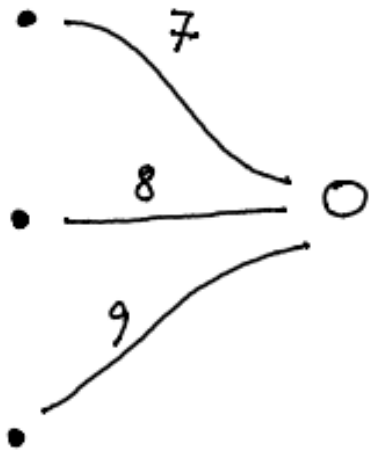
is represented below as a graph



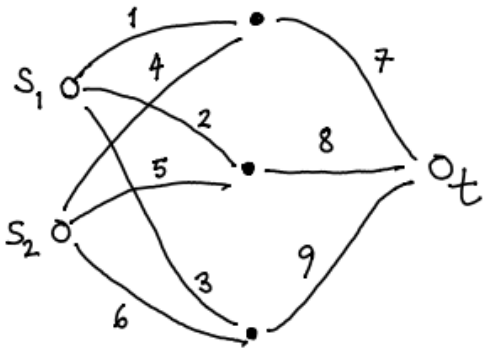
Similarly the 3×1 matrix

$$B = \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix}$$

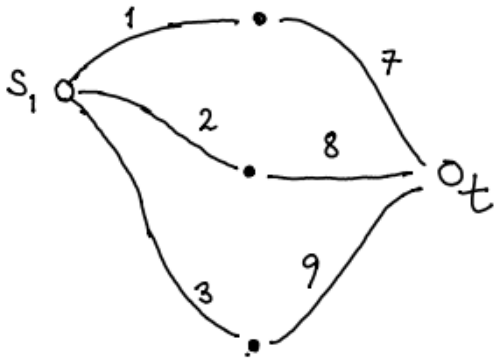
is represented as



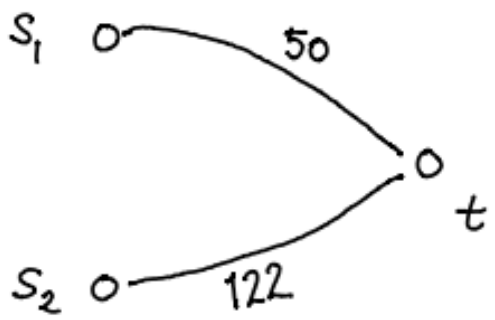
You know that the matrix product AB is a 2×1 matrix. Let us line A and B up graphically:



There are three paths from s_1 to t and three paths from s_2 to t. Here are the three paths from s_1 to t:



Finally, the matrix product AB is represented by the graph



The number 50 on the edge from s_1 to t is gotten by adding the products of the weights on each of the three paths from s_1 to t i.e., $1\cdot 7+2\cdot 8+3\cdot 9$. This is the graphical interpretation of matrix multiplication!

♠

7.6.2 Unpacking the chain rule

The matrix multiplication in (7.14) looks deceivingly simple. Let us write it out. Assume for simplicity that $g:\mathbb{R}^l\to\mathbb{R}^n$ is a function in the variables $x_1,\dots,x_l$ and that $f:\mathbb{R}^n\to\mathbb{R}^m$ is a function in the variables $y_1,\dots,y_n$:

$$g(x_1,\dots,x_l)=\begin{pmatrix} g_1(x_1,\dots,x_l)\\ \vdots\\ g_n(x_1,\dots,x_l)\end{pmatrix}\qquad\text{and}\qquad f(y_1,\dots,y_n))=\begin{pmatrix} f_1(y_1,\dots,y_n)\\ \vdots\\ f_m(y_1,\dots,y_n)\end{pmatrix}.$$

Then $h=(f\circ g):\mathbb{R}^l\to\mathbb{R}^m$ is a function in the variables $x_1,\dots,x_l$:

$$h(x_1,\dots,x_l)=\begin{pmatrix} h_1(x_1,\dots,x_l)\\ \vdots\\ h_m(x_1,\dots,x_l)\end{pmatrix}$$

and we wish to compute $h'(v_0)$, which is an $m\times l$ matrix with entries

$$\frac{\partial h_i}{\partial x_j}(v_0),$$

where $i=1,\dots,m$ represent the rows and $j=1,\dots,l$ the columns. Here (7.14) says that

$$\frac{\partial h_i}{\partial x_j}(v_0)=\frac{\partial f_i}{\partial y_1}(g(v_0))\frac{\partial g_1}{\partial x_j}(v_0)+\cdots+\frac{\partial f_i}{\partial y_n}(g(v_0))\frac{\partial g_n}{\partial x_j}(v_0).$$

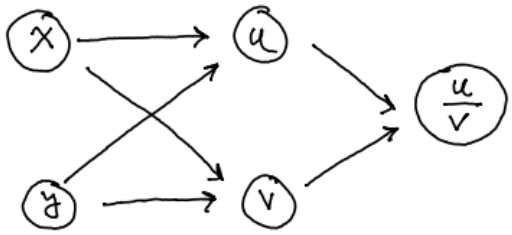
When using the chain rule in computations it pays to use the graphical interpretation of matrix multiplication in subsection 7.6.1 with edges labeled by the derivatives in a computational graph. We illustrate this below.

(7.26) EXAMPLE.

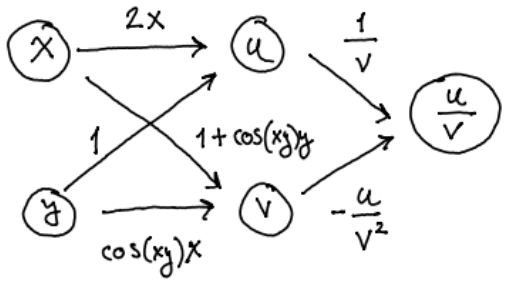
The function $h:\mathbb{R}^2\to\mathbb{R}$ given by

$$h(x,y)=\frac{x^2+y}{x+\sin(xy)}$$

may be evaluated using the computational graph



where u is the function $u(x,y) = x^2 + y$ and v is the function $v(x,y) = x + \sin(xy)$. Similar to the one variable case discussed in the beginning of section 7.6, we label each edge, but now by the partial derivative of the function in its ending node with respect to the variable in its beginning node:



From the graphical interpretation of the matrix product and the chain rule you follow the two paths from the input node x to the output node u/v and conclude that

$$\frac{\partial h}{\partial x} = \frac{2x}{x+\sin(xy)} - \frac{x^2+y}{(x+\sin(xy))^2}(1+\cos(xy)y).$$

♠

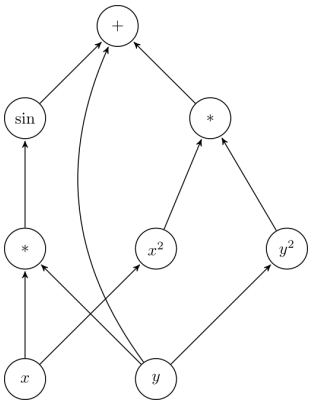
Here is another example of the chain rule in action through a computational graph. In the end you will see an implementation in a famous python library.

(7.27) EXAMPLE.

Consider the example

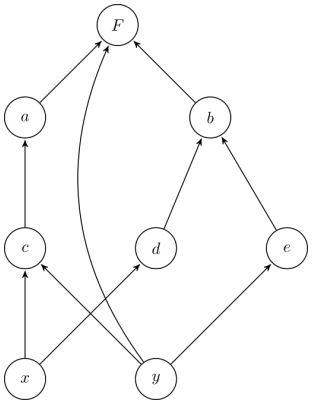
$$f(x,y) = \sin(xy) + x^2y^2 + y$$

from Example 7.9. Even though $f(x,y)$ superficially looks rather simple, it is composed of several smaller functions as displayed in the computational graph



Every node in the above graph, except the input nodes (with no ingoing arrows), represents some function $f : \mathbb{R}^n \to \mathbb{R}^m$. For example the node sin represents a function $f : \mathbb{R} \to \mathbb{R}$ and $*$ represents a function $f : \mathbb{R}^2 \to \mathbb{R}$.

To emphasize that the non-input nodes really are functions we replace them by letters:



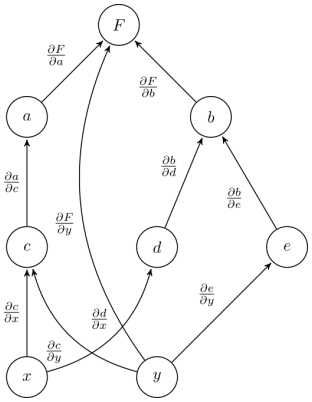
Here we see that

$$f(x,y) = F(a(c(x,y)), y, b(d(x), e(y))),$$

where

$$\begin{aligned} F(a,y,b) &= a+y+b \\ a(c) &= \sin(c) \\ c(x,y) &= xy \\ b(d,e) &= de \\ d(x) &= x^2 \\ e(y) &= y^2 \end{aligned}$$

The gradient is then available from the decorated graph below



by multiplying the decorations on each path from the top to the input variable and the summing up. For example,

$$\frac{\partial F}{\partial x} = \frac{\partial F}{\partial a}\frac{\partial a}{\partial c}\frac{\partial c}{\partial x} + \frac{\partial F}{\partial b}\frac{\partial b}{\partial d}\frac{\partial d}{\partial x}.$$

Computational graphs and the chain rule are important components in machine learning libraries. Below is an example of the computation of $\frac{\partial F}{\partial x}$ in the computational graph above using the `pytorch` library.

Interactive code not included in static version.

♠

(7.28) EXERCISE.

Construct a computational graph for

$$f(x,y) = x^3 + xy + y^3$$

and detail the computation of the gradient ∇f in this context.

Compute the gradient of f at $(x,y) = (1,1)$ using `pytorch`. ♠

(7.29) EXERCISE.

Consider $f : \mathbb{R} \to \mathbb{R}^3$ and $g : \mathbb{R}^3 \to \mathbb{R}$ given by

$$f(t) = \begin{pmatrix} t \\ t^2 \\ t^3 \end{pmatrix} \qquad \text{and} \qquad g\begin{pmatrix} x \\ y \\ z \end{pmatrix} = x^2 + 3y^6 + 2z^5.$$

Compute $(g \circ f)'(t)$ using the chain rule and check the result with an explicit computation of the derivative of $g \circ f : \mathbb{R} \to \mathbb{R}$. ♠

(7.30) EXERCISE.

We wish to show that the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y) = x^2 + y^2$$

is convex. This means that we need to prove that

$$f((1-t)x_0 + tx_1, (1-t)y_0 + ty_1) \leq (1-t)f(x_0,y_0) + tf(x_1,y_1)$$

for every $(x_0.y_0), (x_1,y_1) \in \mathbb{R}^2$ and every t with $0 \leq t \leq 1$. This can be accomplished from the one variable case in the following way. Define

$$g(t) = f((1-t)x_0 + tx_1, (1-t)y_0 + ty_1)$$

and show that g is convex by using the chain rule to show that $g''(t) \geq 0$. Show how the convexity of f follows from this by using that

$$g(t) = g((1-t)\cdot 0 + t \cdot 1).$$

♠

7.7 Logistic regression

The beauty of the sigmoid function is that it takes any value $x \in \mathbb{R}$ and turns it into a probability $0 < \sigma(x) < 1$ by

$$\sigma(x) = \frac{1}{1+e^{-x}},$$

i.e., $\sigma(-\infty) = 0$ and $\sigma(\infty) = 1$.

Graph of the sigmoid function:

Interactive code not included in static version.

(7.31) EXERCISE.

Prove that

$$\sigma'(x) = \sigma(x)(1-\sigma(x))$$

and

$$\log \frac{\sigma(x)}{1-\sigma(x)} = x.$$

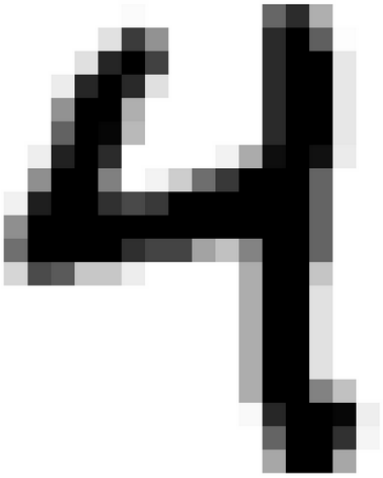
♠

We will not go into all the details (some of which can be traced to introductory probability and statistics), but suppose that we have an outcome E, which may or may not happen.

We have an idea, that the probability of E is dependent on certain parameters $w_0, w_1, \ldots, w_n$ and observations $x_1, \ldots, x_n$ that fit into the sigmoid function as

$$p(x_1, \ldots, x_n) = \sigma(w_0 + w_1 x_1 + \cdots + w_n x_n) = \frac{1}{1+e^{-w_0 - w_1 x_1 - \cdots - w_n x_n}}. \tag{7.15}$$

An example of this could be where $x_1, \ldots, x_{784}$ denote the gray scale of each pixel in a 28×28 image. The event E is whether the image contains the digit 4:



Here $p(x_1, \ldots, x_{784})$ would be the probability that the image contains the digit 4.

7.7.1 Estimating the parameters

Suppose also that we have a table of observations (data set)

$$\begin{array}{cccc} x_{11} & \cdots & x_{1n} & E_1 \\ x_{21} & \cdots & x_{2n} & E_2 \\ \vdots & \ddots & \vdots & \vdots \\ x_{m1} & \cdots & x_{mn} & E_m, \end{array} \tag{7.16}$$

where each row has observations $x_{i1},\dots,x_{in}$ along with a binary variable E_i, which is 1 if E was observed to occur and 0 if not.

Assuming that (7.15) holds, the probability of observing the m observations in (7.16) is

$$\prod_{i=1}^{m} p(x_{i1},\dots,x_{in})^{E_i}(1-p(x_{i1},\dots,x_{in}))^{1-E_i}. \tag{7.17}$$

Notice that (7.17) is a function $L(w_0,\dots,w_n)$ of the parameters $w_0,w_1,\dots,w_n$ for fixed observations $x_1,\dots,x_n$.

We wish to choose the parameters so that $L(w_0,w_1,\dots,w_n)$ is maximized (this is called maximum likelihood estimation). So we are in fact here, dealing with an optimization problem, which is usually solved by gradient descent (for $-L$) or solving the equations

$$\nabla L(w_0,w_1,\dots,w_n)=0.$$

Instead of maximizing $L(w_0,\dots,w_n)$ one usually maximizes the logarithm

$$\begin{aligned}\ell(w_0,w_1,\dots,w_n) &= \log L(w_0,w_1,\dots,w_n)\\ &= \sum_{i=1}^{m} E_i \log p(x_{i1},\dots,x_{in}) + (1-E_i)\log(1-p(x_{i1},\dots,x_{in}))\\ &= \sum_{i=1}^{m} E_i(w_0+w_1x_{i1}+\cdots+w_nx_{in}) - \log(1+e^{w_0+w_1x_{i1}+\cdots+w_nx_{in}}).\end{aligned}$$

Notice that we have used Exercise 7.31 and the logarithm rules $\log(ab)=\log(a)+\log(b)$ and $\log(a/b)=\log(a)-\log(b)$ in the computation above.

(7.32) EXAMPLE.

Suppose that the event E is assumed to be dependent on only one observation x i.e., $n=1$ above. For example, E could be the event of not showing up on a Monday paired with the amount of sleep x in the weekend.

Here

$$p(x)=\sigma(\alpha+\beta x)$$

and

$$\begin{aligned}\ell(\alpha,\beta) &= \sum_{i=1}^{m} E_i \log p(x_i) + (1-E_i)\log(1-p(x_i))\\ &= \sum_{i=1}^{m} E_i(\alpha+\beta x_i) - \log(1+e^{\alpha+\beta x_i}).\end{aligned}$$

♠

(7.33) EXERCISE.

Explain how the end result of the computation of $\ell(\alpha,\beta)$ in Example 7.32 is obtained and compute $\nabla\ell(\alpha,\beta)$.

♠

(7.34) EXAMPLE.

I remember exactly where I was when first hearing about the Challenger[5] disaster in 1986.

[5] See byuistats.github.io for more details on this example

Link to video

This dreadful event was caused by failure of a socalled O-ring. The O-rings had been tested before the launch for failure (=1 below) at different temperatures (in F) resulting in the (partial) table below.

$$\begin{array}{cc} 53.0 & 1 \\ 56.0 & 1 \\ 57.0 & 1 \\ 63.0 & 0 \\ \vdots & \vdots \\ 70.0 & 0 \\ 70.0 & 1 \\ \vdots & \vdots \\ 79.0 & 0 \end{array}$$

At the morning of the launch the outside temperature was (uncharacteristically low for Florida) 31 degrees Fahrenheit. We wish to use logistic regression to compute the probability that the O-ring fails.

Below we have sketched how the logistic regression is carried out using the python library `SciKit-Learn`. The option `solver='lbfgs'` chooses an algorithm for maximizing $\ell(\alpha,\beta)$.

Press the Compute button and see the probability of failure during the launch.

Interactive code not included in static version.

♠

LLM

Explain the function LogicsticRegression in sklearn. In particular, what do the parameters in
model = LogisticRegression(C=25, solver='lbfgs') model.fit(X,y)
mean?

(7.35) EXERCISE.

In the button below is a naive implementation of gradient descent (in fact gradient ascent, because we are dealing with a maximization problem) for the Challenger data set and logistic regression. The implementation is derived from the introduction to gradient descent in this chapter, where we adjusted the step with successive negative powers of 2.

Run experiments with different initial values and number of iterations. Compare with the *official* output from `scikit-learn` in the example above. What is going on?

Also try adjusting the `scikit-learn` output in the example above by removing `C=25` first and then `solver='lbfgs'`. What happens? Compare the quality of the solutions in terms of the gradient (which is available in the output from the Naive code).

Yes, you are allowed (and encouraged) to use generative AI tools here!

Naive code:

Interactive code not included in static version.

♠

7.8 3Blue1Brown

Sit back and enjoy the masterful presentations of neural networks (and the chain rule) by the YouTuber 3Blue1Brown.

7.8.1 Introduction to neural networks

Link to video

7.8.2 Gradient descent

Link to video

7.8.3 Backpropagation and training

Link to video

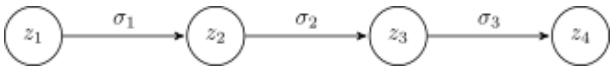
7.8.4 The chain rule in action

Link to video

(7.36) EXERCISE.

Watch the video above before solving this exercise.

Consider the simple neural network



where

$$\begin{aligned} z_2 &= \sigma_1(z_1) = \sigma(a+bz_1) \\ z_3 &= \sigma_2(z_2) = \sigma(c+dz_2) \\ z_4 &= \sigma_3(z_3) = \sigma(e+fz_3), \end{aligned}$$

and σ is the sigmoid function. This neural network has input z_1 and output z_4. Let C be a function of the output z_4. For fixed z_1, we consider C as a function of a,b,c,d,e,f via

$$F\begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \end{pmatrix} = C(\sigma_3(\sigma_2(\sigma_1(z_1)))).$$

Backpropagation for training neural networks is using the chain rule for computing the gradient

$$\nabla F = \left(\frac{\partial F}{\partial a}, \frac{\partial F}{\partial b}, \frac{\partial F}{\partial c}, \frac{\partial F}{\partial d}, \frac{\partial F}{\partial e}, \frac{\partial F}{\partial f}\right).$$

Explain how to do this. ♠

7.9 Lagrange multipliers

The method of Lagrange multipliers is a super classical way of solving optimization problems with non-linear (equality) constraints. We will only consider the special case

$$\begin{aligned} &\text{Maximize/Minimize} && f(x_1,\dots,x_n) \\ &\text{with constraint} && \\ & && g(x_1,\dots,x_n)=0, \end{aligned} \tag{7.18}$$

where both $f:\mathbb{R}^n\to\mathbb{R}$ and $g:\mathbb{R}^n\to\mathbb{R}$ are differentiable functions.

There is a very useful trick for attacking (7.18). One introduces an extra variable λ (a Lagrange multiplier) and the Lagrangian function $L:\mathbb{R}^{n+1}\to\mathbb{R}$ given by

$$L(x_1,\dots,x_n,\lambda)=f(x_1,\dots,x_n)+\lambda g(x_1,\dots,x_n).$$

The main result is the following.

(7.37) THEOREM.

Suppose that $(z_1,\dots,z_n)$ is a local maximum/minimum for (7.18). Then there exists $\lambda\in\mathbb{R}$, such that $(z_1,\dots,z_n,\lambda)$ is a critical point for L.

So to solve (7.18) we simply (well, this is not always so simple) look for critical points for L. This amounts to solving the $n+1$ (non-linear) equations coming from $\nabla L=0$ i.e.,

$$\begin{aligned} g(x_1,\dots,x_n)&=0\\ \frac{\partial f}{\partial x_1}(x_1,\dots,x_n)+\lambda\frac{\partial g}{\partial x_1}(x_1,\dots,x_n)&=0\\ &\vdots\\ \frac{\partial f}{\partial x_n}(x_1,\dots,x_n)+\lambda\frac{\partial g}{\partial x_n}(x_1,\dots,x_n)&=0 \end{aligned} \tag{7.19}$$

For $n=2$ we can quickly give a sketch of the idea behind the proof. The (difficult) fact is that we may find a differentiable function $x(t)$ in one variable t, such that

$$g(t,x(t))=0$$

and the local minimum has the form $v_0=(t_0,x(t_0))$.

Once we have this, the chain rule does its magic. We consider the one variable functions

$$\begin{aligned} F(t)&=f(t,x(t))\\ G(t)&=g(t,x(t)) \end{aligned} \tag{7.20}$$

For both of these we have $F'(t_0)=G'(t_0)=0$ (why?). The chain rule now gives a non-zero vector orthogonal to $\nabla f(v_0)$ and $\nabla g(v_0)$. This is only possible if they are parallel as vectors i. e. , there exists λ, such that

$$\nabla f(v_0)=\lambda\nabla g(v_0).$$

(7.38) EXAMPLE.

Consider the minimization problem

$$\begin{array}{ll} \text{Minimize} & x+y \\ \text{with constraint} & \\ & x^2+y^2=1. \end{array}$$

First of all, why does this problem have a solution at all? We write the non-linear equations

$$\begin{aligned} 1+2x\lambda &= 0 \\ 1+2y\lambda &= 0 \\ x^2+y^2-1 &= 0 \end{aligned}$$

up coming from the critical points of the Langrange function. Now we know that these can be solved and that amongst our solutions there is a minimum! ♠

(7.39) EXERCISE.

Computing the distance from the line $y=x+1$ to the point $(1,1)$ gives rise to the minimization problem

$$\begin{array}{ll} \text{Minimize} & (x-1)^2+(y-1)^2 \\ \text{with constraint} & \\ & y=x+1. \end{array}$$

Solve this minimization problem using Theorem 7.37. ♠

(7.40) EXERCISE.

Use Theorem 7.37 to maximize x^2+y^2 subject to $x^2+xy+y^2=4$.

Hint: Here you end up with the system

$$\begin{aligned} (2\lambda+2)x+\lambda y &= 0 \\ \lambda x+(2\lambda+2)y &= 0 \end{aligned}$$

of linear equations in x and y, where you regard λ as a constant. Use Gaussian elimination to solve this system in order to derive a (nice) quadratic equation in λ coming from

$$-\frac{\lambda}{2\lambda+2}\lambda y+(2\lambda+2)y=0,$$

where you assume that $y\neq 0$. Handle the case $y=0$ separately.

Consider the subset $C=\{(x,y)\in\mathbb{R}^2 \mid x^2+xy+y^2=4\}$. Why is C a closed subset? Why is C bounded?

Hint: To prove that C is bounded you can keep y fixed in

$$x^2+yx+y^2-4=0 \tag{7.21}$$

and solve for x. A last resort is using the plot in Sage in the Hint button below, but that does not give any real insight unless you explain how Sage makes the plot from the equation (7.21).

How does this relate to Theorem 5.66?

Does the optimization problem have a geometric interpretation?

Hint:

Interactive code not included in static version.

♠

(7.41) EXERCISE.

A rectangular box has side lengths x, y and z. What is its maximal volume when we assume that (x,y,z) lies on the plane

$$\frac{x}{a}+\frac{y}{b}+\frac{z}{c}=1$$

for $a,b,c>0$. ♠

(7.42) EXERCISE.

A company is planning to produce a box with volume 2 m^3. For design reasons it needs different materials for the sides, top and bottom. The cost of the materials per square meter is 1 dollar for the sides, 1.5 dollars for the bottom and the top. Find the measurements of the box minimizing the production costs.

Hint: Let x,y and z be the measurements. Use $xyz=2$ to rewrite the Lagrange equations so that y and z are expressed in terms of x. ♠

(7.43) EXERCISE.

Maximize	$-p_1\log_2(p_1)-\cdots-p_n\log_2(p_n)$
with constraint(s)	
	$p_1+\cdots+p_n=1.$
	$p_1>0,\ldots,p_n>0$

The sum

$$H(p_1,\ldots,p_n)=-p_1\log_2(p_1)-\cdots-p_n\log_2(p_n)$$

is called the (Shannon) entropy of the discrete probability distribution $p_1,\ldots,p_n$. One may use Jensen's inequality applied to the convex function $-\log_2(x)$ to prove that

$$H(p_1,\ldots,p_n)\leq\log_2(n).$$

♠

7.10 Optimization using the interior and boundary of a subset

Suppose that $C \subseteq \mathbb{R}^n$ is a closed subset and $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function. Recall (see Theorem 5.66) that the optimization problem

$$\begin{array}{ll} \text{Optimize} & f(v) \\ \text{with constraint} & \\ & v \in C \end{array}$$

always has a solution if C in addition to being closed is also bounded. To solve such an optimization problem, it often pays to decompose C as

$$C = \partial C \cup C^o,$$

where ∂C is the boundary of C (recall Definition 5.44) and C^o the interior of C (recall Definition 5.46). The strategy is then to look for an optimal solution both in ∂C and C^o and then compare these. In some sense we are making a "recursive" call to a lower dimensional optimization problem for the boundary ∂C. This is illustrated by the basic example: $f(x) = x^2 - 5x + 6$ and $C = [0,4]$. Here $\partial C = \{0,4\}$ and $C^o = (0,4)$. Notice that ∂C is finite here.

If v_0 is an element of C^o, then there exists an open subset $U \subseteq C$, such that $v_0 \in U$. Therefore the following proposition holds, when you take Proposition 7.21 into account.

(7.44) PROPOSITION.

Consider an optimization problem

$$\begin{array}{ll} \textit{Optimize} & f(x) \\ \textit{with constraint} & \\ & x \in C, \end{array} \tag{7.22}$$

where $C \subseteq \mathbb{R}^n$ is a subset, $f : \mathbb{R}^n \to \mathbb{R}$ a differentiable function and v_0 an optimal solution to (7.22). If $v_0 \in C^o$, then v_0 is a critical point of f.

Basically, to solve an optimization problem like (7.22) one needs to consider the boundary and interior as separate cases. For points on the boundary we cannot use the critical point test in Proposition 7.21. This test only applies to the interior points.

Usually the boundary cases are of smaller dimension and easier to handle as illustrated in the example below.

(7.45) EXAMPLE.

Consider the minimization problem

$$\begin{array}{ll} \text{Minimize} & x+y \\ \text{with constraint} & \\ & x^2+y^2 = 1. \end{array}$$

from Example 7.38. Let us modify it to

$$\begin{array}{ll} \text{Minimize} & x+y \\ \text{with constraint} & \\ & (x,y) \in C, \end{array} \tag{7.23}$$

where

$$C = \{(x,y) \in \mathbb{R}^2 \mid x^2+y^2 \leq 1\}.$$

We are now minimizing not only over the unit circle, but the whole unit disk. Here

$$\partial C = \{(x,y) \in \mathbb{R}^2 \mid x^2+y^2 = 1\} \quad \text{and} \quad C^o = \{(x,y) \in \mathbb{R}^2 \mid x^2+y^2 < 1\}.$$

Proposition 7.44 guides us first to look for optimal points in C^o. Here we use Proposition 7.21 to show that there can be no optimal points in C^o, because the gradient of the function $f(x,y) = x+y$ is

$$\nabla f = (1,1).$$

Therefore the boundary needs to be analyzed and the usual technique (as was implicit in Lagrange multipliers) is to find a parametrization for the points (x,y) satisfying $x^2+y^2 = 1$. There are two of those (one for the upper unit circle and one for the lower unit circle):

$$\left(t, \sqrt{1-t^2}\right)$$
$$\left(t, -\sqrt{1-t^2}\right),$$

where $t \in [-1,1]$. This means that the optimization problem for the boundary ∂C turns into the two simpler optimization problems of minimizing

$$t+\sqrt{1-t^2} \qquad \text{and} \qquad t-\sqrt{1-t^2}$$

subject to $t \in [-1,1]$. These can as one variable optimization problems be solved the usual way. ♠

The exercises below are taken from an older Calculus course at Aarhus.

(7.46) EXERCISE.

Solve the two optimization problems

Maximize/Minimize	$x^2-2xy+2y$
with constraint	
	$(x,y) \in C,$

where $C = \{(x,y) \in \mathbb{R}^2 \mid 0 \le x \le 3, 0 \le y \le 2\}$. But first give a reason as to why they both are solvable.

Hint: First find ∂C and C^o. Then try with Proposition 7.44 supposing that a maximal point really is to be found in C^o and not on ∂C. ♠

(7.47) EXERCISE.

Solve the two optimization problems

Maximize/Minimize	$1+4x-5y$
with constraint	
	$(x,y) \in C,$

where $C = \{(x,y) \in \mathbb{R}^2 \mid 0 \le x, 0 \le y, 3x+2y \le 6\}$. But first give a reason as to why they both are solvable. ♠

(7.48) EXERCISE.

Solve the two optimization problems

$$\begin{array}{ll} \text{Maximize/Minimize} & 3+xy-x-2y \\ \text{with constraint} & \\ & (x,y)\in C, \end{array}$$

where C is the triangle with vertices in $(1,0),(5,0)$ and $(1,4)$. But first give a reason as to why they both are solvable. ♠

(7.49) EXERCISE.

Use Proposition 7.44 to give all the minute details in applying Theorem 7.37 to solve Exercise 7.42.

Hint: First rewrite to the problem, where you minimize $6/y+4/x+2xy$ subject to $x>0, y>0$ by using $xyz=2$. Then explain why this problem may be solved by restricting with upper and lower bounds on x and y. The minimum $(6\sqrt[3]{6})$ is attained in a critical point and not on the boundary. For $N\in\mathbb{N}\setminus\{0\}$ one may optimize over the compact subset

$$C_N=\{(x,y)\mid \frac{1}{N}\le x\le N, \frac{1}{N}\le y\le N\}$$

and analyze what happens when $N\to\infty$. ♠

Chapter 8

The Hessian

In Chapter 6 we exploited the second derivative $f''(x)$ of a one variable real function $f : (a,b) \to \mathbb{R}$ to analyze convexity along with local minima and maxima.

In this chapter we introduce an analogue of the second derivative for real functions $f : \mathbb{R}^n \to \mathbb{R}$ of several variables. This will be an $n \times n$ matrix. The important notion of a matrix being positive (semi-) definite introduced in Section 3.7 will now make its appearance.

8.1 Introduction

In Section 6.4 the Taylor expansion for a one variable differentiable function $f : \mathbb{R} \to \mathbb{R}$ centered a x_0 with step size $h = x - x_0$ was introduced as

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{1}{2!}f''(x_0)h^2 + \cdots \tag{8.1}$$

Recall that the second derivative $f''(x_0)$ contains a wealth of information about the function. Especially if $f'(x_0) = 0$, then we might glean from $f''(x_0)$ if x_0 is a local maximum or minimum or none of these (see Theorem 6.47 and review Exercise 6.50).

We also noticed that gradient descent did not work so well only descending along the gradient. We need to take the second derivative into account to get a more detailed picture of the function.

8.2 Several variables

Our main character is a differentiable function $F : \mathbb{R}^n \to \mathbb{R}$ in several variables. We already know that

$$F(v_0 + h) = F(v_0) + \nabla F(v_0)h + \varepsilon(h)\,|h|\,,$$

where v_0 and h are vectors in $\mathbb{R}^n$ (as opposed to the good old numbers in (8.1)). Take a look back at Definition 7.5 for the general definition of differentiability.

We wish to have an analogue of the Taylor expansion in (8.1) for such a function of several variables. To this end we introduce the function $g : \mathbb{R} \to \mathbb{R}$ given by

$$g(t) = F(v_0 + th). \tag{8.2}$$

Notice that

$$g(t) = (F \circ A)(t),$$

where $A : \mathbb{R} \to \mathbb{R}^n$ is the function given by $A(t) = x_0 + th$. In particular we get

$$g'(t) = F'(v_0 + th)h = \nabla F(v_0 + th)h \tag{8.3}$$

by using the chain rule (see Theorem 7.24).

(8.1) EXERCISE.

Explain how the chain rule is applied to get (8.3). ♠

The derivative $g'(t)$ is also composed of several functions and again we may compute $g''(t)$ by using the chain rule:

$$\begin{aligned} g''(t) &= (C \circ B \circ A)'(t) \\ &= (C \circ B)'(A(t))A'(t) \\ &= C'(B(A(t)))B'(A(t))A'(t), \end{aligned} \tag{8.4}$$

where $B : \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$B(v) = \nabla F(v)^\top$$

and $C : \mathbb{R}^n \to \mathbb{R}$ by

$$C(v) = v^\top h.$$

(8.2) DEFINITION.

The Hessian matrix *of F at the point $v \in \mathbb{R}^n$ is defined by*

$$\nabla^2 F(v) := \begin{pmatrix} \frac{\partial^2 F}{\partial x_1 \partial x_1}(v) & \cdots & \frac{\partial^2 F}{\partial x_1 \partial x_n}(v) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 F}{\partial x_n \partial x_1}(v) & \cdots & \frac{\partial^2 F}{\partial x_n \partial x_n}(v) \end{pmatrix}.$$

A very important observation is that $\nabla^2 F(v)$ above is a symmetric matrix if F satisfies the condition in the last part of Theorem 7.13.

(8.3) EXAMPLE.

Suppose that $f : \mathbb{R}^2 \to \mathbb{R}$ is given by

$$f(x,y) = \sin(xy) + x^2 y^2 + y.$$

Then the gradient

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)$$

and the Hessian

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}$$

of f are computed in the Sage window below.

Interactive code not included in static version.

See the further documentation for Calculus functions in Sage. ♠

(8.4) EXERCISE.

Verify (just this once) by hand the computations done by Sage in Example 8.3.

Also, experiment with a few other functions in the Sage window and compute their Hessians. ♠

By applying Proposition 7.11 it is not too hard to see that the Hessian matrix fits nicely into the framework above, since

$$B'(v) = \nabla^2 F(v). \tag{8.5}$$

The full application of the chain rule then gives

$$g''(t) = h^\top \nabla^2 F(v_0 + th)h. \tag{8.6}$$

(8.5) EXERCISE.

Give a detailed explanation as to why (8.5) holds. ♠

8.3 Newton's method for finding critical points

We may use Newton's method for computing critical points for a function $F : \mathbb{R}^n \to \mathbb{R}$ of several variables. Recall that a critical point is a point $v_0 \in \mathbb{R}^n$ with $\nabla F(v_0) = 0$. By (7.8) and (8.5) the computation in Newton's method becomes

$$v_1 = v_0 - \big(\nabla^2 F(v_0)\big)^{-1} \nabla F(v_0). \tag{8.7}$$

In practice the (inverse) Hessian appearing in (8.7) is often a heavy computational burden. This leads to the socalled quasi-Newton methods, where the inverse Hessian in (8.7) is replaced by other matrices.

(8.6) EXAMPLE.

We will return to the logistic regression in Example 7.34 about the Challenger disaster. Here we sought to maximize the function

$$\ell(\alpha, \beta) = \sum_{i=1}^{m} E_i(\alpha + \beta x_i) - \log(1 + e^{\alpha + \beta x_i}). \tag{8.8}$$

In order to employ Newton's method we compute the gradient and the Hessian of (8.8)

$$\begin{aligned}
\frac{\partial \ell}{\partial \alpha} &= \sum_{i=1}^{m} E_i - \sigma(\alpha + \beta x_i) \\
\frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^{m} E_i x_i - x_i \sigma(\alpha + \beta x_i) \\
\frac{\partial^2 \ell}{\partial \alpha^2} &= \sum_{i=1}^{m} -\sigma'(\alpha + \beta x_i) \\
\frac{\partial^2 \ell}{\partial \beta \partial \alpha} &= \sum_{i=1}^{m} -\sigma'(\alpha + \beta x_i) x_i \\
\frac{\partial^2 \ell}{\partial \beta^2} &= \sum_{i=1}^{m} -\sigma'(\alpha + \beta x_i) x_i^2,
\end{aligned} \tag{8.9}$$

where

$$\sigma(t) = \frac{1}{1+e^{-t}}$$

is the sigmoid function.

Notice the potential problem in using Newton's method here: the formula for the second order derivatives in (8.9) show that if the $\alpha + \beta x_i$ are just mildly big, say ≥ 50, then the Hessian is extremely close to the zero matrix and therefore Sage considers it non-invertible and (8.7) fails.

In the code below we have nudged the initial vector so that it works, but you can easily set other values and see its failure. Optimization is not just mathematics, it also calls for some good (engineering) implementation skills (see for example details on the quasi Newton algorithms).

In the instance below we do, however, get a gradient that is practically $(0,0)$.

Code for Newton's method:

Interactive code not included in static version.

♠

8.3.1 Transforming data for better numerical performance

The numerical problems with Newton's method in Example 8.6 can be prevented by transforming the input data. It makes sense to transform data from large numbers to smaller numbers around 0. There is a rather standard way of doing this.

Suppose in logistic regression we have a set of data

$$x_1, x_2, \ldots, x_n \tag{8.10}$$

associated with outcomes $E_1, \ldots, E_n$. Then the function

$$\ell(\alpha,\beta) = \sum_{i=1}^{m} E_i(\alpha + \beta x_i) - \log(1 + e^{\alpha+\beta x_i}).$$

from Example 7.32 becomes much more manageable if we first transform the data according to

$$x_i' = \frac{x_i - \bar{x}}{\sigma}$$

and instead optimize the function

$$\ell'(\alpha,\beta) = \sum_{i=1}^{m} E_i(\alpha + \beta x_i') - \log(1 + e^{\alpha+\beta x_i'}).$$

Here

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

is the mean value and

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

the variance of the data in (8.10).

Now if α' and β' is an optimum for ℓ', then

$$\begin{aligned} \alpha &= \alpha' - \frac{\bar{x}}{\sigma}\beta' \\ \beta &= \frac{\beta'}{\sigma} \end{aligned} \tag{8.11}$$

is an optimum for ℓ, since

$$\ell'(\alpha,\beta) = \ell\left(\alpha - \beta\frac{\bar{x}}{\sigma}, \frac{\beta}{\sigma}\right).$$

(8.7) EXERCISE.

Why is the claim/trick alluded to in (8.11) true?

Below is a snippet of Sage code implementing the trick in (8.11). The function `test` takes as input `x0` (an initial vector like `[0,0]`) and `noofits` (the number of iterations of Newton's method). You execute this in the Sage window by adding for example

Interactive code not included in static version.

and then pressing `Compute`.

Experiment and compare with the official output from Example 7.34. Also, compute the gradient of the output below for the original non-transformed problem.

Transformed code:

Interactive code not included in static version.

♠

8.4 The Hessian and critical points

Now we are in a position to state at least the first terms in the Taylor expansion for a differentiable function $F : \mathbb{R}^n \to \mathbb{R}$. The angle of the proof is to reduce to the one-dimensional case through the function $g(t)$ defined in (8.2). Here one may prove that

$$g(t) = g(0) + g'(0)t + \frac{1}{2}g''(0)t^2 + \varepsilon(t)t^2, \tag{8.12}$$

where $\varepsilon(0) = 0$ with ε continuous at 0, much like in the definition of differentiability except that we also include the second derivative.

Now (8.12) translates into

$$F(v_0 + th) = F(v_0) + (\nabla F(v_0)h)t + \frac{1}{2}\left(h^\top \nabla^2 F(v_0)h\right)t^2 + \varepsilon(t)t^2 \tag{8.13}$$

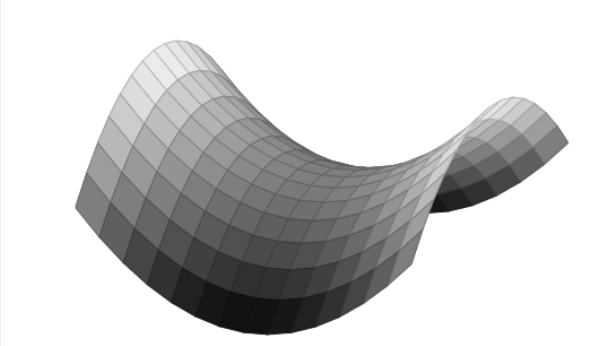
by using (8.3) and (8.6).

(8.8) DEFINITION.

A critical point v_0 is called a saddle point *for F if there exists two vectors $u, v \in \mathbb{R}^n$, such that*

$$\begin{aligned} t = 0 \quad & \textit{is a local minimum for the function} \quad F(v_0 + tu) \\ t = 0 \quad & \textit{is a local maximum for the function} \quad F(v_0 + tv) \end{aligned}$$

as illustrated in the graphics below.



Now go back and recall the definition of positive definite matrices in Section 3.7. We call a symmetric A matrix *negative definite* if $-A$ is positive definite. One more concept (related to the definition of saddle point above):

(8.9) DEFINITION.

A symmetric $n \times n$ matrix A is called indefinite *if there exists $u, v \in \mathbb{R}^n$ with*

$$\begin{aligned} u^\top A u > 0 \quad & \textit{and} \\ v^\top A v < 0. & \end{aligned}$$

So an indefinite matrix is mixed up in the sense that it is neither positive definite, nor negative definite.

(8.10) EXAMPLE.

$$\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \quad \text{is positive definite}$$

$$\begin{pmatrix} -2 & 0 \\ 0 & -3 \end{pmatrix} \quad \text{is negative definite}$$

$$\begin{pmatrix} 2 & 0 \\ 0 & -3 \end{pmatrix} \quad \text{is indefinite}$$

$$\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{is neither positive definite, negative definite, nor indefinite}$$

♠

We have the following addition to Proposition 3.41 (with a similar proof).

(8.11) PROPOSITION.

Let A be a symmetric $n \times n$ matrix and B an invertible $n \times n$ matrix. Then A is indefinite (positive definite, negative definite) if and only if

$$B^\top AB$$

is indefinite (positive definite, negative definite).

From (8.13) one can prove the following nice criterion, which may be viewed as a several variable generalization of Theorem 6.47.

(8.12) THEOREM.

Let v_0 be a critical point for $F : \mathbb{R}^n \to \mathbb{R}$. Then

(i) v_0 is a local minimum if $\nabla^2 F(v_0)$ is positive definite.

(ii) v_0 is a local maximum if $\nabla^2 F(v_0)$ is negative definite.

(iii) v_0 is a saddle point if $\nabla^2 F(v_0)$ is indefinite.

(8.13) EXAMPLE.

Consider, with our new technology in Theorem 8.12, Exercise 7.22 once again. Here we analyzed the point $v_0 = (0,0)$ for the function

$$f(x,y) = x^3 + xy + y^3$$

and showed (by a trick) that v_0 is neither a local maximum nor a local minimum for f. The Hessian matrix for $f(x,y)$ at v_0 is

$$H = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Now Theorem 8.12iii shows that v_0 is a saddle point, since

$$\begin{pmatrix} x & y \end{pmatrix} H \begin{pmatrix} x \\ y \end{pmatrix} = 2xy$$

and

$$u^\top Hu > 0 \qquad \text{for } u = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$
$$v^\top Hv < 0 \qquad \text{for } v = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Interactive code not included in static version.

♠

(8.14) EXERCISE.

Try plotting the graph for different values of `a`[1] in the Sage window in Example 8.13. What do you observe for the point v_0 with respect to the function? Does `a` have to be a number? Could it be a symbolic expression in the variables `x` and `y` like `a = -10*cos(x)*sin(y)`? ♠

(8.15) EXERCISE.

Check the computation of the Hessian matrix H in Example 8.13 by showing that the Hessian matrix for f at the point (x,y) is

$$\begin{pmatrix} 6x & 1 \\ 1 & 6y \end{pmatrix}.$$

♠

(8.16) EXERCISE.

What about u and v in Example 8.13? How do they relate to the hint given in Exercise 7.22? ♠

(8.17) EXERCISE.

Give an example of a function $F : \mathbb{R}^2 \to \mathbb{R}$ having a local minimum at x_0, where $\nabla^2 F(x_0)$ is not positive definite. ♠

(8.18) EXERCISE.

The following exercise is a `sci2u` exercise from the Calculus book.

(i) The point $\left(0, \frac{\sqrt{3}}{3}\right)$ is a critical point for

$$f(x,y) = x^3 + y^3 - y.$$

What does Theorem 8.12 say about this point?

(ii) The point $\left(\frac{1}{3}, \frac{1}{3}\right)$ is a critical point for

$$f(x,y) = -x^3 - x^2 + x - y^3 + 2y^2 - y.$$

What does Theorem 8.12 say about this point?

(iii) The point $(0,1)$ is a critical point for

$$f(x,y) = x^3 - x^2 + y^3 - y^2 - y.$$

What does Theorem 8.12 say about this point?

[1] `a=4` shows the saddle point clearly.

♠

(8.19) EXERCISE.

Consider the function

$$f(x,y) = x^4y^2 + x^2y^4 - 3x^2y^2.$$

Compute its critical points and decide on their types according to Theorem 8.12. Try to convince yourself that

$$f(x,y) \geq -1$$

for every $x, y \in \mathbb{R}$.

Interactive code not included in static version.

Hint:

Look at the minimization problem

$$\min f(x,y)$$

subject to

$$(x,y) \in C = \{(x,y) \mid -M \leq x \leq M,\ -M \leq y \leq M\},$$

where M is a big number. ♠

(8.20) EXERCISE.

Give an example of a function $f : \mathbb{R} \to \mathbb{R}$ that has a local maximum, but where there exists $x \in \mathbb{R}$ with $f(x) > M$ for any given (large) number M. ♠

8.5 Differential convex functions of several variables

Below is the generalization of Theorem 6.58 to several variables. You have already seen this in Exercise 6.59, right?

(8.21) THEOREM.

Let $f : U \to \mathbb{R}$ be a differentiable function, where $U \subseteq \mathbb{R}^n$ is an open convex subset. Then f is convex if and only if

$$f(x) \geq f(x_0) + \nabla f(x_0)(x - x_0) \tag{8.14}$$

for every $x, x_0 \in U$.

Proof: Suppose that (8.14) holds and let $x_t = (1-t)x_0 + tx$ with $0 \leq t \leq 1$, where $x_0, x \in U$. To prove that f is convex we must verify the inequality

$$f(x_t) \leq (1-t)f(x_0) + tf(x). \tag{8.15}$$

Let $\xi = \nabla f(x_t)$. Then

$$f(x) \geq f(x_t) + \xi(1-t)(x-x_0)$$
$$f(x_0) \geq f(x_t) - \xi t(x-x_0)$$

by (8.14). If you multiply the first inequality by t, the second by $1-t$ and then add the two, you get (8.15).

Suppose on the other hand that f is a convex function. Let $x_0, x \in U$. Since U is an open subset, it follows that $(1-t)x_0 + tx \in U$ for $t \in I = (-\delta, 1+\delta)$, where $\delta > 0$ is sufficiently small. Now define the function $g : I \to \mathbb{R}$ by

$$g(t) = f((1-t)x_0 + tx) = f(x_0 + t(x-x_0)).$$

Being the composition of two differentiable functions, g is differentiable. Suppose that $0 \leq \alpha \leq 1$ and $t_1, t_2 \in I$. Then

$$\begin{aligned} g((1-\alpha)t_1 + \alpha t_2) &= f(x_0 + ((1-\alpha)t_1 + \alpha t_2)(x-x_0)) \\ &= f((1-\alpha)(x_0 + t_1(x-x_0)) + \alpha(x_0 + t_2(x-x_0))) \\ &\leq (1-\alpha) f(x_0 + t_1(x-x_0)) + \alpha f(x_0 + t_2(x-x_0)) \\ &= (1-\alpha) g(t_1) + \alpha g(t_2) \end{aligned}$$

showing that g is a convex function. By Theorem 6.58,

$$g(1) \geq g(0) + g'(0),$$

which translates into

$$f(x) \geq f(x_0) + \nabla f(x_0)(x-x_0)$$

by using the chain rule in computing $g'(0)$.

(8.22) EXERCISE.

Prove that a bounded convex differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is constant. ♠

The following is the generalization of Corollary 6.52.

(8.23) THEOREM.

Let $f : U \to \mathbb{R}$ be a differentiable function with continuous second order partial derivatives, where $U \subseteq \mathbb{R}^n$ is a convex open subset. Then f is convex if and only if the Hessian $\nabla^2 f(x)$ is positive semidefinite for every $x \in U$. If $\nabla^2 f(x)$ is positive definite for every $x \in U$, then f is strictly convex.

Proof: We have done all the work for a convenient reduction to the one variable case. Suppose that f is convex. Then the same reasoning as in the proof of Theorem 8.21 shows that

$$g(t) = f(x+tv)$$

is a convex function for every $x \in U$ and every $v \in \mathbb{R}^n$ from an open interval $(-\delta, \delta)$ to $\mathbb{R}$ for suitable $\delta > 0$. Therefore $g''(0) = v^\top \nabla^2 f(x) v \geq 0$ by Theorem 6.51. This proves that the matrix $\nabla^2 f(x)$ is positive semidefinite for every $x \in U$. Suppose on the other hand that $\nabla^2 f(x)$ is positive semidefinite for every $x \in U$. Then Theorem 6.51 shows that $g(t) = f(x+t(y-x))$ is a convex function from $(-\delta, 1+\delta)$ to $\mathbb{R}$ for $\delta > 0$ small and $x, y \in U$, since

$$g''(\alpha) = (y-x)^\top \nabla^2 f(x+\alpha(y-x))(y-x) \geq 0$$

for $0 \leq \alpha \leq 1$. Therefore f is a convex function, since

$$\begin{aligned} f((1-t)x+ty) =& g((1-t)\cdot 0 + t\cdot 1) \\ & \leq (1-t)g(0)+tg(1) = (1-t)f(x)+tf(y). \end{aligned}$$

The same argument (using the last part of Theorem 6.51 on strict convexity), shows that g is strictly convex if $\nabla^2 f(x)$ is positive definite. It follows that f is strictly convex if $\nabla^2 f(x)$ is positive definite for every $x \in U$.

(8.24) EXERCISE.

Prove that

$$f(x,y) = x^2 + y^2$$

is a strictly convex function from $\mathbb{R}^2$ to $\mathbb{R}$. Also, prove that

$$\{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$$

is a convex subset of $\mathbb{R}^2$. ♠

(8.25) EXERCISE.

Is $f(x,y) = \cos(x) + \sin(y)$ strictly convex on some non-empty open convex subset of the plane? ♠

(8.26) EXERCISE.

Show that $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y) = \log(e^x + e^y)$$

is a convex function. Is f strictly convex? ♠

(8.27) EXERCISE.

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$f(x,y) = ax^2 + by^2 + cxy,$$

where $a, b, c \in \mathbb{R}$.

(i) Show that f is a strictly convex function if and only if $a > 0$ and $4ab - c^2 > 0$.

Hint: This is a hint for the only if part. If H is the Hessian for f, then

$$f(v) = \frac{1}{2} v^\top H v,$$

where $v = (x,y)^\top$ - this is seen by a matrix multiplication computation. We know that H is positive semidefinite. If H was not positive definite, there would exist $v \neq 0$ with $f(v) = 0$. Now use $f(tv) = t^2 f(v)$ to complete the proof that H is positive definite by looking at $f((1-t)\cdot 0 + t \cdot v)$.

(ii) Suppose now that $a > 0$ and $4ab - c^2 > 0$. Show that $g(x,y) = f(x,y) + x + y$ has a unique global minimum and give a formula for this minimum in terms of a, b and c.

♠

8.6 How to decide the definiteness of a matrix

In this section we will outline a straightforward method for deciding if a matrix is positive definite, positive semidefinite, negative definite or indefinite.

Before proceeding it is a must that you do the following exercise.

(8.28) EXERCISE.

Show that a diagonal matrix

$$\begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}$$

is positive definite if and only if $\lambda_1 > 0, \dots, \lambda_n > 0$, positive semidefinite if and only if $\lambda_1 \geq 0, \dots, \lambda_n \geq 0$ and indefinite if and only if there exists $i \neq j$ with $\lambda_i > 0$ and $\lambda_j < 0$.

♠

The crucial ingredient is the following result.

(8.29) THEOREM.

Let A be a real symmetric $n \times n$ matrix. Then there exists an invertible matrix B, such that $B^\top AB$ is a diagonal matrix.

The proof contains an algorithm for building B by different steps. We will supply examples afterwards illustrating these. An operational procedure implementing the steps is outlined in section 8.7.

Proof: Suppose that $A = (a_{ij})$. If A has a non-zero entry in the upper left hand corner i.e., $a_{11} \neq 0$, then

$$B_1^\top AB_1 = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & c_{11} & \cdots & c_{1,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & c_{n-1,1} & \cdots & c_{n-1,n-1}, \end{pmatrix}$$

where $C = (c_{ij})$ is a real symmetric matrix and B_1 is the invertible $n \times n$ matrix

$$\begin{pmatrix} 1 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

By induction on n we may find an invertible matrix $(n-1) \times (n-1)$ matrix B_2 such that

$$B_2^\top CB_2 = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n-1} \end{pmatrix}.$$

Putting

$$B = B_1 \begin{pmatrix} 1 & 0 \\ 0 & B_2 \end{pmatrix},$$

it follows that

$$B^\top AB = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n-1} \end{pmatrix}.$$

We now treat the case of a zero entry in the upper left hand corner i.e., $a_{11} = 0$. Suppose first that $a_{jj} \neq 0$ for some $j > 1$. Let P denote the identity matrix with the first and j-th rows interchanged. The operation $A \mapsto AP$ amounts to interchanging the first and j-th columns in A. Similarly $A \mapsto P^\top A$ is interchanging that first and j-th rows in A. The matrix P is invertible and $P^\top AP$ is a symmetric matrix with $(P^\top AP)_{11} = a_{jj} \neq 0$ and we have reduced to the case of a non-zero entry in the upper left hand corner.

If $a_{ii} = 0$ for every $i = 1, \ldots, n$ we may assume that $a_{1j} \neq 0$ for some $j > 1$. Let B denote the identity matrix where the entry in the first column and j-th row is 1. The operation $A \mapsto AB$ amounts to adding the j-th column to the first column in A. Similarly $A \mapsto B^\top A$ is adding the j-th row to the first row in A. All in all we get $(B^\top AB)_{11} = 2a_{1j} \neq 0$, where we have used that $a_{ii} = 0$ for $i = 1, \ldots, n$. Again we have reduced to the case of a non-zero entry in the upper left hand corner.

(8.30) EXAMPLE.

Consider the 3×3 real symmetric matrix.

$$A = (a_{ij}) = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 8 & 4 \\ 3 & 4 & 16 \end{pmatrix}.$$

Here $a_{11} = 1 \neq 0$. Therefore the fundamental step in the proof of Theorem 8.29 applies and

$$\begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix} A \begin{pmatrix} 1 & -2 & -3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & -2 \\ 0 & -2 & 7 \end{pmatrix}$$

and again

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & -2 \\ 0 & -2 & 7 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 6 \end{pmatrix}.$$

Summing up we get

$$B = \begin{pmatrix} 1 & -2 & -3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -2 & -4 \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}.$$

You are invited to check that

$$B^\top AB = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 6 \end{pmatrix}.$$

♠

(8.31) EXAMPLE.

Let

$$A = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & 3 \\ 1 & 2 & 1 & 4 \\ 1 & 3 & 4 & 0 \end{pmatrix}.$$

Here $a_{11} = a_{22} = 0$, but the diagonal element $a_{33} \neq 0$. So we are in the second step of the proof of Theorem 8.29. Using the matrix

$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

we get

$$P^\top AP = \begin{pmatrix} 1 & 2 & 1 & 4 \\ 2 & 0 & 0 & 3 \\ 1 & 0 & 0 & 1 \\ 4 & 3 & 1 & 0 \end{pmatrix}.$$

As argued in the proof, this corresponds to interchanging the first and third columns and then interchanging the first and third rows. In total you move the non-zero a_{33} to the upper left corner in the matrix. ♠

(8.32) EXAMPLE.

Consider the symmetric matrix

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}.$$

We have zero entries in the diagonal. As in the third step in the proof of Theorem 8.29 we must find an invertible matrix B_1, such that the upper left corner in $B_1^\top AB_1$ is non-zero. In the proof it is used that every diagonal element is zero: if we locate a non-zero element in the j-th column in the first row, we can add the j-th column to the first column and then the j-th row to the first row obtaining a non-zero element in the upper left corner. For A above we choose $j = 2$ and the matrix B_1 becomes

$$B_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

so that

$$B_1^\top AB_1 = \begin{pmatrix} 2 & 1 & 2 & 2 \\ 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix}.$$

♠

(8.33) EXERCISE.

Let A be any matrix. Show that

$$A^\top A$$

is positive semidefinite. ♠

(8.34) EXERCISE.

Find inequalities defining the set

$$\left\{ (a,b) \in \mathbb{R}^2 \,\middle|\, \begin{pmatrix} 2 & 1 & a \\ 1 & 1 & 1 \\ a & 1 & b \end{pmatrix} \text{ is positive definite} \right\}.$$

Same question with positive semidefinite. Sketch and compare the two subsets of the plane $\{(a,b) \mid a,b \in \mathbb{R}\}$. ♠

(8.35) EXERCISE.

Let $f : \mathbb{R}^3 \to \mathbb{R}$ denote the function given by

$$f(x,y,z) = x^2 + y^2 + z^2 + axy + xz + yz,$$

where $a \in \mathbb{R}$. Let H denote the Hessian of f in a point $(x,y,z) \in \mathbb{R}^3$.

(i) Compute H.

(ii) Show that $f(v) = v^\top Av$ for $v = (x,y,z) \in \mathbb{R}^3$ and $A = \frac{1}{2}H$.

(iii) Compute a non-zero vector $v \in \mathbb{R}^3$, such that $Hv = 0$ in the case, where $a = 2$. Is H invertible in this case?

(iv) Show that f is strictly convex if $-1 < a < 2$.

(v) Is f strictly convex if $a = 2$?

Hint: Consider the line segment between 0 and a suitable vector $u \neq 0$, where $f(u) = 0$.

♠

(8.36) EXERCISE.

Why is the subset given by the inequalities

$$\begin{aligned} x &\geq 0 \\ y &\geq 0 \\ xy - z^2 &\geq 0 \end{aligned}$$

a convex subset of $\mathbb{R}^3$? ♠

8.7 A schematic procedure for transforming symmetric matrices

Suppose that A is a symmetric $n \times n$ matrix. We wish to find an invertible matrix B and a diagonal matrix D so that

$$B^\top AB = D.$$

Every step in the algorithm in the proof of Theorem 8.29 involve an operation on the columns of A followed by a similar operation on the rows. These steps can be carried out systematically by transforming the extended $(2n) \times n$ matrix

$$\begin{pmatrix} I_n \\ A \end{pmatrix} \qquad \text{into} \qquad \begin{pmatrix} B \\ D \end{pmatrix}. \tag{8.16}$$

The recipe is: every column operation (on A) is carried out on the full $(2n) \times n$ matrix, whereas every row operation is only carried out on the lower $n \times n$ matrix in (8.16).

Here is how this plays out for the examples above.

(8.37) EXAMPLE.

Here is the schematic procedure applied to Example 8.30:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 2 & 3 \\ 2 & 8 & 4 \\ 3 & 4 & 16 \end{pmatrix} \to \begin{pmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 3 \\ 2 & 4 & 4 \\ 3 & -2 & 16 \end{pmatrix} \to \begin{pmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 3 \\ 0 & 4 & -2 \\ 3 & -2 & 16 \end{pmatrix} \to \begin{pmatrix} 1 & -2 & -3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 4 & -2 \\ 3 & -2 & 7 \end{pmatrix} \to$$

$$\begin{pmatrix} 1 & -2 & -3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 4 & -2 \\ 0 & -2 & 7 \end{pmatrix} \to \begin{pmatrix} 1 & -2 & -4 \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & -2 & 6 \end{pmatrix} \to \begin{pmatrix} 1 & -2 & -4 \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 6 \end{pmatrix}.$$

♠

(8.38) EXAMPLE.

Here is the schematic procedure applied to Example 8.31:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & 3 \\ 1 & 2 & 1 & 4 \\ 1 & 3 & 4 & 0 \end{pmatrix} \to \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 2 & 0 & 0 & 3 \\ 1 & 2 & 1 & 4 \\ 4 & 3 & 1 & 0 \end{pmatrix} \to \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 2 & 1 & 4 \\ 2 & 0 & 0 & 3 \\ 1 & 0 & 0 & 1 \\ 4 & 3 & 1 & 0 \end{pmatrix}.$$

♠

(8.39) EXAMPLE.

Here is the schematic procedure applied to Example 8.32:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 2 & 1 & 2 & 2 \\ 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix}.$$

♠

Chapter 9

Convex optimization

In this last chapter we will deal exclusively with convex optimization problems.

Recall that a convex optimization problem has the form

$$\begin{array}{ll} \text{Minimize} & f(x_1,\ldots,x_n) \\ \text{with constraint} & \\ & (x_1,\ldots,x_n) \in C, \end{array}$$

where $C \subseteq \mathbb{R}^n$ is a *convex subset* (see Definition 4.19) and $f : C \to \mathbb{R}$ a *convex function* (see Definition 4.24). We will mainly deal with the case, where f is differentiable defined on all of $\mathbb{R}^n$ in addition to just being convex defined on C. Also recall that convex optimization problems are very well behaved in the sense that local minima are global (see Theorem 6.13).

(9.1) EXAMPLE.

Below is an example of a convex optimization problem in the plane $\mathbb{R}^2$.

$$\begin{array}{ll} \text{Minimize} & x^2+y^2 \\ \text{with constraint} & \\ & (x,y) \in C, \end{array}$$

where C is the subset of points (x,y) in $\mathbb{R}^2$ satisfying

$$\begin{aligned} x+y &\geq 2 \\ y &\leq 2 \\ x &\leq 3 \\ y &\geq 1. \end{aligned}$$

♠

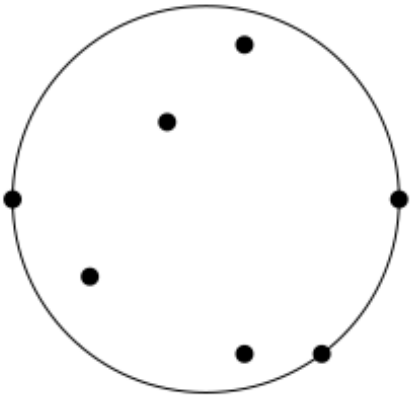
(9.2) EXERCISE.

Sketch the subset C in Example 9.1. Show that Example 9.1 really is a convex optimization problem and solve it. ♠

Below is an example of a convex optimization problem modelling the real life problem of placing a fire station (center of circle) so that the maximal distance to the surrounding houses (points to be enclosed) is minimal.

(9.3) EXAMPLE.

Given n points $(x_1,y_1),\ldots,(x_n,y_n) \in \mathbb{R}^2$, what is the center and radius of the smallest circle containing these points?



We can write this optimization problem as

$$\begin{array}{ll} \text{Minimize} & r \\ \text{with constraint} & \\ & (x,y) \in C, \end{array} \tag{9.1}$$

where

$$C = \{(x,y) \in \mathbb{R}^2 \mid (x-x_i)^2 + (y-y_i)^2 \leq r^2 \quad \text{for} \quad i = 1,\ldots,n\}.$$

Upon rewriting this turns into the optimization problem

$$\begin{array}{ll} \text{Minimize} & x^2+y^2+\lambda \\ \text{with constraint} & \\ & (x,y) \in C' \end{array} \tag{9.2}$$

where

$$C' = \{(x,y,\lambda) \in \mathbb{R}^3 \mid x_i^2 + y_i^2 \leq 2x_i x + 2y_i y + \lambda \quad \text{for} \quad i = 1,\ldots,n\}$$

and $\lambda = r^2 - x^2 - y^2$. ♠

(9.4) EXERCISE.

Prove that (9.1) and (9.2) both are convex optimization problems. Explain how (9.1) is rewritten into (9.2). ♠

(9.5) EXERCISE.

Prove that (9.1) and (9.2) both are convex optimization problems. Explain how (9.1) is rewritten into (9.2).

Hint: Expand

$$(x-x_i)^2 + (y-y_i)^2 \leq r^2$$

and put $\lambda = r^2 - x^2 - y^2$. ♠

9.1 Finding the optimal hyperplane separating data

In section 5.3.2 we were presented with labeled data

$$(v_1, y_1), \ldots, (v_m, y_m), \tag{9.3}$$

where $v_i \in \mathbb{R}^n$ and $y_i = \pm 1$. The task at hand was to separate differently labeled data by a hyperplane $\alpha^\top v + \beta = 0$, such that

$$\begin{aligned} \alpha^\top v_i + \beta > 0 \quad \text{if} \quad y_i = 1 \\ \alpha^\top v_i + \beta < 0 \quad \text{if} \quad y_i = -1 \end{aligned} \tag{9.4}$$

for $i = 1, , \ldots, m$. Please browse back to Definition 4.38 for the definition of a hyperplane in $\mathbb{R}^n$.

To make this more real, consider the points $(1,1), (-1,-1)$ with label $+1$ and the points $(-1,1), (1,-1)$ with label -1 (as in Exercise 5.9 and above it). Here a hyperplane satisfying (9.4) cannot exist: suppose that $\alpha = (\alpha_1, \alpha_2)^\top$. Then (9.4) is tantamount to the following inequalities

$$\begin{aligned} \alpha_1 + \alpha_2 + \beta > 0 \\ -\alpha_1 - \alpha_2 + \beta > 0 \\ -\alpha_1 + \alpha_2 + \beta < 0 \\ \alpha_1 - \alpha_2 + \beta < 0. \end{aligned}$$

But these inequalities are unsolvable in α_1, α_2 and β (why?).

If the data in (9.3) can be separated according to (9.4), we may find a hyperplane $(\alpha^*)^\top x + \beta^* = 0$, such that

$$\begin{aligned} (\alpha^*)^\top v_i + \beta^* \geq 1 \quad \text{if} \quad y_i = 1 \\ (\alpha^*)^\top v_i + \beta^* \leq -1 \quad \text{if} \quad y_i = -1. \end{aligned} \tag{9.5}$$

(9.6) EXERCISE.

How do you go from (9.4) to (9.5)?

Hint: Suppose that

$$\begin{aligned} \alpha^\top v_i + \beta > 0 \quad \text{if} \quad y_i = 1 \\ \alpha^\top v_i + \beta < 0 \quad \text{if} \quad y_i = -1. \end{aligned}$$

Let

$$\begin{aligned} N = \min\{\alpha^\top v_i + \beta \mid y_i = 1\} \\ M = \max\{\alpha^\top v_i + \beta \mid y_i = -1\}. \end{aligned}$$

Show that $N > 0$ and $M < 0$. How can N and M be applied in constructing α^* and β^*?

Hint: Consider

$$\begin{aligned} \alpha^\top v_i + \beta \geq N > 0 \quad \text{if} \quad y_i = 1 \\ \alpha^\top v_i + \beta \leq M < 0 \quad \text{if} \quad y_i = -1. \end{aligned}$$

What is special about $m = \min(N, |M|)$?

♠

What does optimal hyperplane mean in this setting? This is the one maximizing the width of the strip between the two labeled clusters.

(9.7) FIGURE.

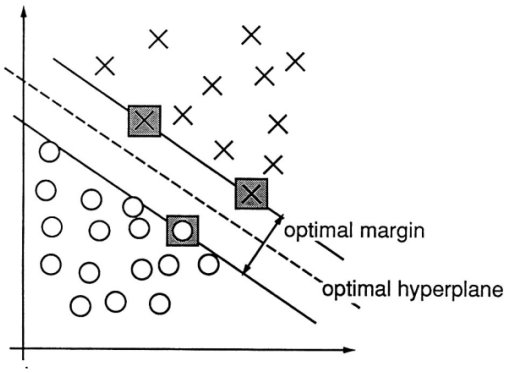


Figure from the Cortes and Vapnik paper: Support vector networks, Machine Learning, 1995.

A rather crucial insight (not explicitly mentioned in the paper by Cortes and Vapnik) is that such a hyperplane is given as $H = \{v \in \mathbb{R}^d \mid \alpha^\top v + \beta = 0\}$ satisfying (9.5) **and** with maximal distance to all of the data points. Therefore the function to maximize (with respect to $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$) is

$$\min\left\{\frac{|\alpha^\top v_i + \beta|}{|\alpha|}\,\middle|\, i = 1,\dots,m\right\},$$

since the distance from H to a point u is (see Exercise 9.8)

$$\frac{|\alpha^\top u + \beta|}{|\alpha|}.$$

The conditions in (9.5) may be written as $|\alpha^\top v_i + \beta| \geq 1$ for $i = 1,\dots,m$. If $m = \min\{|\alpha^\top v_i + \beta| \mid i = 1,\dots,m\} > 1$, then we multiply α and β by $1/m$, so that we may assume

$$\min\left\{\frac{|\alpha^\top v_i + \beta|}{|\alpha|}\,\middle|\, i = 1,\dots,m\right\} = \frac{1}{|\alpha|}.$$

(9.8) EXERCISE.

Let H be the hyperplane in $\mathbb{R}^n$ given by $\alpha^\top v + \beta = 0$ and let $u \in \mathbb{R}^n$. The point closest to u in H can be found by solving the optimization problem

$$\begin{array}{ll} \text{Minimize} & |v-x|^2 \\ \text{with constraint} & \\ & x \in H. \end{array} \tag{9.6}$$

Explain why (9.6) is a convex optimization problem.

Show how Theorem 7.37 can be used to solve this optimization problem by first deducing the equations

$$\begin{aligned} -2(u-v) + \lambda\alpha &= 0 \\ \alpha^\top v + \beta &= 0 \end{aligned} \tag{9.7}$$

for the Lagrange multiplier λ. Notice here that $-2(u-v)+\lambda\alpha=0$ above really contains n equations, whereas $\alpha^\top v+\beta=0$ is only one equation in $x_1,\ldots,x_n$, where $v=(x_1,\ldots,x_n)^\top$. Solve the equations (9.7) for $v\in\mathbb{R}^n$ and $\lambda\in\mathbb{R}$. How can we be sure that v really is a minimum in (9.6)?

Finally show that the distance from H to u is given by the formula

$$\left|\frac{\alpha^\top u+\beta}{|\alpha|}\right|.$$

♠

We have therefore proved the following result (notice that maximizing $1/|\alpha|$ is the same as minimizing $|\alpha|^2$)

(9.9) THEOREM.

Let the points $v_1,\ldots,v_m\in\mathbb{R}^n$ be labeled by $y_1,\ldots,y_m\in\{\pm1\}$. Then the optimal hyperplane $H=\{v\in\mathbb{R}^n \mid \alpha^\top v+\beta=0\}$ separating the points is given by the optimization problem

$$\begin{aligned}&\textit{Minimize} && |\alpha|^2\\ &\textit{with constraints}\\ &&& \alpha^\top v_i+\beta\geq 1 \quad \textit{if} \quad y_i=1\\ &&& \alpha^\top v_i+\beta\leq -1 \quad \textit{if} \quad y_i=-1\end{aligned}$$

for $i=1,\ldots,m$.

The vectors v_i among the data points $v_1,\ldots,v_m$ satisfying $\alpha^\top v_i+\beta=1$ or $\alpha^\top v_i+\beta=-1$ are called *support vectors*.

(9.10) EXAMPLE.

Let us explicitly write up the optimization problem In Theorem 9.9 in a very simple situation: finding the best line $y=ax+b$ separating the points $(1,1)$ and $(2,2)$. In the notation of (9.5), we have (without the stars on α and β)

$$\alpha=\begin{pmatrix}a\\-1\end{pmatrix} \qquad \text{and} \qquad \beta=b$$

so that

$$\alpha^\top\begin{pmatrix}x\\y\end{pmatrix}+b=ax-y+b=0.$$

The points are

$$v_1=\begin{pmatrix}1\\1\end{pmatrix} \qquad \text{and} \qquad v_2=\begin{pmatrix}2\\2\end{pmatrix},$$

where $y_1=1$ and $y_2=-1$.

Therefore the optimization problem in Theorem 9.9 becomes

$$\begin{aligned}&\text{Minimize} && 1+a^2\\ &\text{with constraints}\\ &&& a+b\geq 2\\ &&& 2a+b\leq 1\end{aligned} \tag{9.8}$$

♠

(9.11) EXERCISE.

Solve the optimization problem (9.8) and verify that the best line from the optimization problem is the one we expect it to be. Also, check how WolframAlpha solves this optimization problem.

Hint: You could maybe use Fourier-Motzkin elimination to show that

$$\begin{aligned} a+b &\geq 2 \\ 2a+b &\leq 1 \end{aligned}$$

implies $a \leq -1$. ♠

Notice that the optimization problem in Theorem 9.9 has number of constraints equal to the number of points to be separated. For an extended (soft margin) optimization problem, when the data at hand cannot be separated we refer to section 3 of the Cortes and Vapnik paper.

Usually one does not use the optimization problem in Theorem 9.9, but rather its socalled (Lagrange) dual for finding the optimal hyperplane. This dual optimization problem uses that the normal vector α is a linear combination

$$\alpha = \lambda_1 v_1 + \cdots + \lambda_m v_m \tag{9.9}$$

of the support vectors. It is an optimization problem in $\Lambda^\top = (\lambda_1, \ldots, \lambda_m)$ from (9.9) and looks like

$$\begin{aligned} &\text{Maximize} && \lambda_1 + \cdots + \lambda_m - \frac{1}{2}\Lambda^\top D \Lambda \\ &\text{with constraints} \\ &&& \Lambda \geq 0 \\ &&& \Lambda^\top Y = 0, \end{aligned} \tag{9.10}$$

where $Y = (y_1, \ldots, y_m)^\top$ is the vector of labels attached to the points $v_1, \ldots, v_m$ and D is the symmetric $m \times m$ matrix given by

$$D_{ij} = y_i y_j v_i^\top v_j = y_i y_j v_i \cdot v_j. \tag{9.11}$$

(9.12) REMARK.

Notice that the dual optimization problem is an optimization problem in $\mathbb{R}^d$ with $d = m$, where m is the number of data points. This can be in stark contrast to the original optimization problem in Theorem 9.9, which is an optimization problem in $\mathbb{R}^n$, where n is the dimension of the data points. Sometimes the data points are high dimensional and it pays to solve the dual optimization problem.

(9.13) EXAMPLE.

Let us write down the dual optimization problem for the points in Example 9.10. Here

$$\Lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}, \quad Y = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} 2 & -4 \\ -4 & 8 \end{pmatrix}$$

so that the dual optimization problem becomes

$$\begin{aligned} &\text{Maximize} && \lambda_1+\lambda_2-\lambda_1^2-4\lambda_2^2+4\lambda_1\lambda_2 \\ &\text{with constraints} \\ &\quad \lambda_1 \geq 0 \\ &\quad \lambda_2 \geq 0 \\ &\quad \lambda_1-\lambda_2=0. \end{aligned}$$

This reduces to the optimization problem of maximizing $2\lambda-\lambda^2$ subject to $\lambda \geq 0$, which has the solution $\lambda=1$. Therefore the optimal hyperplane has normal vector

$$\alpha=\lambda_1\begin{pmatrix}1\\1\end{pmatrix}+\lambda_2\begin{pmatrix}2\\2\end{pmatrix}=\begin{pmatrix}3\\3\end{pmatrix}.$$

♠

The dual optimization problem (9.10) can be derived formally from the original optimization problem in Theorem 9.9. This is, however, beyond the scope of this course (see section 2.1 of the Cortes and Vapnik paper).

(9.14) EXAMPLE.

Quadratic optimization problems, such as the one in Theorem 9.9 can in fact be handled by Sage (well, python in this case). See CVXOPT for further information. Note that the code below needs to be executed as `Python` code (choose Python in the pull down). It attempts (in general) to solve the optimization problem

$$\begin{aligned} &\text{Minimize} && \frac{1}{2}x^\top Qx+p^\top x \\ &\text{with constraints} \\ &&& Gx \leq h \\ &&& Ax=b. \end{aligned}$$

In the Sage window below the optimization problem

$$\begin{aligned} &\text{Minimize} && 2x_1^2+x_2^2+x_1x_2+x_1+x_2 \\ &\text{with constraints} \\ &&& x_1 \geq 0 \\ &&& x_2 \geq 0 \\ &&& x_1+x_2=1 \end{aligned}$$

has been entered.

Interactive code not included in static version.

What happens if you remove

Interactive code not included in static version.

from the code above? ♠

(9.15) EXERCISE.

Take a look at the input format in Example 9.14. Can you tell which optimization problem in this chapter is solved below? Also, the code below seems to report some errors after pressing the compute button. Can you make it run smoothly by making a very, very small change?

Interactive code not included in static version.

♠

9.1.1 Separating by non-linear functions

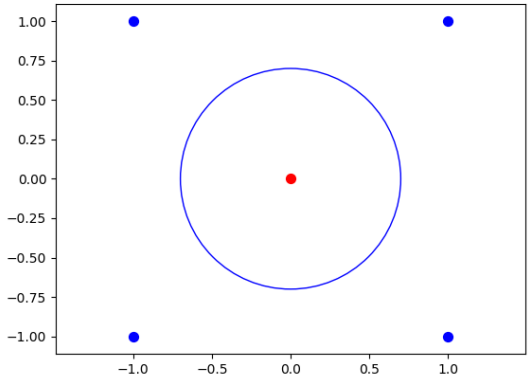
Sometimes one needs more complex separating curves than just a line. Consider the five points

$$(-1,1), \quad (1,1), \quad (1,-1), \quad (-1,-1), \quad \text{and} \quad (0,0),$$

where we wish to separate $(0,0)$ from the other points. This is impossible using a line, but certainly doable by a circle

$$x^2+y^2=r^2, \tag{9.12}$$

where $0<r<\sqrt{2}$:



The circle (9.12) may be a circle in two dimensions, but viewed in three dimensions it turns into a hyperplane in the following way.

By using the function $\varphi : \mathbb{R}^2 \to \mathbb{R}^3$ given by

$$\varphi(x,y) = (x^2,y^2,1) \in \mathbb{R}^3 = \{(x_1,y_1,z_1) \mid x_1,y_1,z_1 \in \mathbb{R}\},$$

points lying on (9.12) map to points lying on the hyperplane in $\mathbb{R}^3$ given by

$$x_1+y_1=r^2z_1$$

in $\mathbb{R}^3$.

(9.16) EXAMPLE.

An even simpler example is given in dimension one. Consider the points (or numbers)

$$-1,1,2 \in \mathbb{R} \tag{9.13}$$

with respective labels $1,-1,1$. These cannot be separated by a hyperplane $ax+b=0$. Things change dramatically if we use the function $\varphi(x)=(x,x^2)$ to embed the points in $\mathbb{R}^2$. Using φ, the points (9.13) map to

$$(-1,1),(1,1),(2,4)\in\mathbb{R}^2$$

with the (same) labels $1,-1,1$. Here they can be separated by the hyperplane $y=x+1$. This means that the original numbers can be separated by the non-linear function $x^2=x+1$ or rather $f(x)=x^2-x-1$ so that $f(-1)>0, f(1)<0$ and $f(2)>0$. ♠

The general trick is to find a suitable map

$$\varphi:\mathbb{R}^d\to\mathbb{R}^N,$$

such that the transformed data

$$(\varphi(x_1),y_1),\ldots,(\varphi(x_m),y_m)$$

becomes linearly separable. Since,

$$\alpha=\lambda_1\varphi(x_1)+\cdots+\lambda_m\varphi(x_m)$$

for suitable $\lambda_1,\ldots,\lambda_m\in\mathbb{R}$, the (dual) optimization problem in (9.10) becomes

$$\begin{aligned}&\text{Maximize}\qquad \lambda_1+\cdots+\lambda_m-\frac{1}{2}\Lambda^\top D\Lambda\\ &\text{with constraints}\\ &\Lambda\geq 0\\ &\Lambda^\top Y=0,\end{aligned}$$

where $Y=(y_1,\ldots,y_m)^\top$ is the vector of labels attached to the points $\varphi(x_1),\ldots,\varphi(x_m)$ and D is the symmetric $m\times m$ matrix given by

$$D_{ij}=y_iy_j\,\varphi(x_i)\cdot\varphi(x_j). \tag{9.14}$$

The beauty of the dual problem (as mentioned in Remark 9.12) is that we do not have to care about the (sometimes astronomical, even "infinite") dimension of $\mathbb{R}^N$. The optimization problem is situated in $\mathbb{R}^m$, where m is the number of data points (or more precisely the number of constraints). We only need a clever way of getting our hands on $\varphi(x_i)\cdot\varphi(x_j)$ in (9.14). Here an old concept from pure mathematics called kernel functions helps us.

9.1.2 Kernel functions

A kernel function is a function $k:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$, that is a hidden dot product in the following sense: there exists a function

$$\varphi:\mathbb{R}^n\to\mathbb{R}^N,$$

such that

$$k(u,v)=\varphi(u)\cdot\varphi(v). \tag{9.15}$$

(9.17) EXAMPLE.

Let $k:\mathbb{R}^2\times\mathbb{R}^2\to\mathbb{R}$ be given by

$$k(u,v)=(u^\top v+1)^2.$$

Then

$$\begin{aligned}k((x_1,y_1),(x_2,y_2))&=(x_1x_2+y_1y_2+1)^2\\&=x_1^2x_2^2+y_1^2y_2^2+2x_1x_2y_1y_2+2x_1x_2+2y_1y_2+1.\end{aligned} \tag{9.16}$$

One gleans from (9.16) that k is a kernel function, since (9.15) is satisfied for $\varphi : \mathbb{R}^2 \to \mathbb{R}^6$ given by

$$\varphi(a,b) = (a^2, b^2, \sqrt{2}ab, \sqrt{2}a, \sqrt{2}b, 1).$$

♠

Once we have a kernel function for φ we can replace the matrix in (9.14) by

$$D_{ij} = y_i y_j k(x_i, x_j)$$

and proceed to solve the optimization problem without worrying about the sometimes insurmountable size of $\mathbb{R}^N$.

9.1.3 The kernel perceptron algorithm

Recall the stunningly simple perceptron algorithm from section 5.3.2. This algorithm can be modified to handle non-linear separation too by using kernel functions. In fact, this modification was one of the inspirations for the development of the support vector machines described above.

After having mapped a set of vectors $x_1, \ldots, x_m \in \mathbb{R}^d$ with labels $y_1, \ldots, y_m \in \{\pm 1\}$ to $\varphi(x_1), \ldots, \varphi(x_m)$, via $\varphi : \mathbb{R}^d \to \mathbb{R}^N$, we are looking for a vector $w \in \mathbb{R}^N$, such that

$$\begin{aligned} w^\top \varphi(x_i) &> 0 \quad \text{if} \quad y_i = 1 \\ w^\top \varphi(x_i) &< 0 \quad \text{if} \quad y_i = -1. \end{aligned} \tag{9.17}$$

Such a vector is expressible as

$$w = \lambda_1 \varphi(x_1) + \cdots + \lambda_m \varphi(x_m).$$

The (dual) perceptron algorithm works adjusting the coefficients $\lambda_1, \ldots, \lambda_m$ successively as follows: if $w^\top$ is wrong about the placement of $\varphi(x_j)$ in (9.17) i.e., if $y_j w^\top \varphi(x_j) < 0$, then let

$$\lambda_j := \lambda_j + y_j.$$

If we have a kernel function k for φ, then

$$w^\top \varphi(x_j) = w \cdot \varphi(x_j) = \sum_{i=1}^{m} \lambda_i \varphi(x_i) \cdot \varphi(x_j) = \sum_{i=1}^{m} \lambda_i k(x_i, x_j)$$

and we can use the kernel function in the algorithm without resorting to computing φ and the inner product in $\mathbb{R}^N$.

(9.18) EXERCISE.

Use the kernel function in Example 9.17 and the kernel perceptron algorithm to separate

$$((-1,-1),-1), \quad ((-1,1),-1), \quad ((1,-1),-1), \quad ((0,0),1), \quad ((1,1),1).$$

Sketch the points and the separating curve. ♠

9.2 Logarithmic barrier functions

We need an algorithm for solving optimization problems like (9.9). There is a very nice trick (probably going back to von Neumann) for solving constrained optimization problems of the form

$$\begin{array}{ll} \text{Minimize} & f(x_1,\ldots,x_n) \\ \text{with constraint} & \\ & (x_1,\ldots,x_n) \in C, \end{array} \tag{9.18}$$

where C is defined by the differentiable functions $g_i : \mathbb{R}^n \to \mathbb{R}$ as

$$C = \{x \in \mathbb{R}^n \mid g_1(x) \leq 0, \ldots, g_m(x) \leq 0\}.$$

The functions g_i define the boundary (or barrier) of C. We use them to define the logarithmic *barrier function*

$$B(x) = -\sum_{i=1}^{m} \log(-g_i(x))$$

defined on the interior

$$C^o = \{x \in \mathbb{R}^n \mid g_1(x) < 0, \ldots, g_m(x) < 0\}.$$

The boundary of C is

$$\partial C = \{x \in C \mid g_1(x) = 0 \vee \cdots \vee g_m(x) = 0\}.$$

You can see that the logarithmic barrier function explodes (becomes unbounded), when a vector $x \in C^o$ approaches ∂C, since $-\log(t)$ is unbounded as $t \to 0$ for $t > 0$.

The cool idea is to consider the function

$$f_\varepsilon(x) = f(x) + \varepsilon B(x) \tag{9.19}$$

for $\varepsilon > 0$. This function has a global minimum $x_\varepsilon \in C^o$.

(9.19) EXERCISE.

Prove that f_ε is a convex function if f and $g_1, \ldots, g_m$ are convex functions.

Hint: Prove and use that if f is a decreasing convex function (in one variable) and g is a convex function, then $f(-g(x))$ is a convex function, where we assume the composition makes sense. ♠

The upshot is that $x_\varepsilon \to x_0$ as $\varepsilon \to 0$. This is the content of the following theorem, which we will not prove.

(9.20) THEOREM.

Let x_ε be a point in C^o with

$$f_\varepsilon(x_\varepsilon) = \min\{f_\varepsilon(x) | x \in C^o\}$$

for $\varepsilon > 0$ and $f^* = \min\{f(x) | x \in C\}$. Then

$$0 \leq f(x_\varepsilon) - f^* \leq \varepsilon m$$

and $f(x_\varepsilon) \to f^*$ as $\varepsilon \to 0$. If (9.18) has a unique optimum x^*, then by using $\varepsilon = \frac{1}{n}$ we obtain a sequence $x_{\frac{1}{n}} \to x^*$ as $n \to \infty$.

We move on to give concrete examples of Theorem 9.20 in action.

9.2.1 Quadratic function with polyhedral constraints

A much used setup in optimization is minimizing a quadratic functions subject to polyhedral constraints. This is the optimization problem

$$\begin{array}{ll} \text{Minimize} & x^T Qx + c^T x \\ \text{with constraint} & \\ & Ax \leq b, \end{array} \tag{9.20}$$

where Q is an $n \times n$ matrix, A is an $m \times n$ matrix, $c \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$.

Certainly the constraints $Ax \leq b$ define a convex subset of $\mathbb{R}^n$, but the function $x^\top Qx + c^\top x$ is not strictly convex unless Q is positive definite. If Q is not positive semidefinite (9.20) is difficult.

If Q is positive semidefinite, the interior point method outlined above usually works well.

(9.21) EXAMPLE.

The optimization problem (9.2) has the form (9.20), when we put

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$c = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}$$

$$A = \begin{pmatrix} -2x_1 & -2y_1 & -1 \\ \vdots & \vdots & \vdots \\ -2x_n & -2y_n & -1 \end{pmatrix}$$

$$b = \begin{pmatrix} -x_1^2 - y_1^2 \\ \vdots \\ -x_n^2 - y_n^2 \end{pmatrix}$$

♠

(9.22) EXAMPLE.

The optimization problem in Theorem 9.9 has the form (9.20), when we put

$$Q = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

$$c = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$A = \begin{pmatrix} -y_1 x_1 & -y_1 \\ \vdots & \vdots \\ -y_n x_n & -y_n \end{pmatrix}$$

$$b = \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix}$$

Here Q is a $(d+1)\times(d+1)$ matrix, A is an $n\times(d+1)$ matrix and $b\in\mathbb{R}^n$. ♠

Optimization of a quadratic function as in (9.20) is implemented below using the interior point method and exact line search. See Section 10.5.1 of my book Undergraduate Convexity for further details. Only `python` with `numpy` is used.

Interactive code not included in static version.

(9.23) EXAMPLE.

Below are samples of output running the interior point algorithm on the enclosing circle problem in Example 9.3.

$$\varepsilon = 1,\quad 0.5,\quad 0.1,\quad 0.05,\quad 0.01,\quad 0.005,\quad 0.001,\quad 0.0005,\quad 0.0001$$

in the barrier function $f_\varepsilon(x)$ in (9.19). We are attempting to compute the center of the smallest enclosing circle of the points

$$(0,0),\quad (2,2),\quad (-3,2),\quad (1,0),\quad (-2,1),\quad (-1,3),\quad \text{and}\quad (0,4).$$

Interactive code not included in static version.

The first two coordinates of the output are the x- and y-coordinates of the center. The third is λ from (9.2). ♠

(9.24) EXERCISE.

Try out the code in the Sage window above on the Exercises 7.46, 7.47 and 7.48. Check the output of the code by actually solving these exercises. ♠

(9.25) EXERCISE.

Compute the best line separating the labeled data

Interactive code not included in static version.

♠

9.3 A geometric optimality criterion

Consider the general optimization problem

$$\begin{array}{ll} \text{Minimize} & f(x_1,\ldots,x_n) \\ \text{with constraint} & \\ & (x_1,\ldots,x_n)\in C, \end{array} \tag{9.21}$$

where C is a subset of $\mathbb{R}^n$.

(9.26) PROPOSITION.

Suppose that $C \subseteq \mathbb{R}^n$ is a convex subset and $f : \mathbb{R}^n \to \mathbb{R}$ a differentiable function in (9.21). If $v_0 \in C$ is an optimal solution of (9.21), then

$$\nabla f(v_0)(v - v_0) \geq 0 \qquad \text{for every } v \in C. \tag{9.22}$$

If f in addition is a convex function, then (9.22) implies that v_0 is optimal.

Proof: If v_0 is an optimal solution and $x \in C \setminus \{v_0\}$, then

$$\begin{aligned} 0 \leq f((1-t)v_0 + tv) - f(v_0) &= f(v_0 + t(v - v_0)) - f(v_0) \\ &= t\,(\nabla f(v_0)(v - v_0) + \varepsilon(t(v - v_0))\,|v - v_0|) \end{aligned}$$

for every t with $0 \leq t \leq 1$, where ε denotes the epsilon function in the definition of differentiability (see Definition 7.5). Therefore

$$\nabla f(v_0)(v - v_0) + \varepsilon(t(v - v_0))\,|v - v_0| \geq 0$$

for $0 \leq t \leq 1$. This is only possible if $\nabla f(v_0)(v - v_0) \geq 0$. We have silently applied the convexity of C and the differentiability of f at v_0.

If f in addition is convex and (9.22) holds, then Theorem 8.21 shows that v_0 is an optimal solution.

A nice application of Proposition 9.26 is where the constraints in C are linear. Then you can test if v_0 is an optimal solution by solving the linear program

$$\min \nabla f(v_0)v$$

for $v \in C$. If the optimal value is $\geq \nabla f(v_0)v_0$, then v_0 is an optimal solution.

(9.27) EXAMPLE.

A nice application of Proposition 9.26 is for example to the optimization problem

$$\begin{array}{ll} \text{Minimize} & (x+1)^2 + (y+1)^2 \\ \text{with constraint} & \\ & x^2 + 3y^2 \leq 1 \end{array}$$

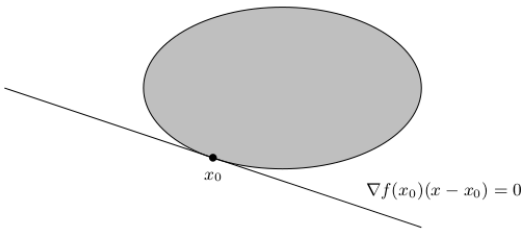
Here Proposition 9.26 shows that $v_0 = \left(-\frac{1}{2}, -\frac{1}{2}\right)$ is optimal, since the hyperplane

$$\nabla f(v_0)v = \nabla f(v_0)v_0$$

touches the boundary of

$$C = \{(x,y) \in \mathbb{R}^2 \mid x^2 + 3y^2 \leq 1\}$$

as shown below.



♠

(9.28) EXERCISE.

Sketch how Proposition 9.26 applies to show that an optimum in a linear programming problem

$$\begin{array}{ll} \text{Minimize} & cx+dy \\ \text{with constraint} & \\ & A\begin{pmatrix} x \\ y \end{pmatrix} \leq b \end{array}$$

in the plane $\mathbb{R}^2$ always can be found in a vertex. ♠

(9.29) EXERCISE.

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a differentiable convex function and

$$S = \{(x,y) | -1 \leq x \leq 2, -1 \leq y \leq 1\}.$$

Suppose that $\nabla f(v_0) = (1,0)$ for $v_0 = (-1, \frac{1}{2})$. Prove that v_0 is a minimum for f defined on S. ♠

(9.30) EXERCISE.

Guess the solution to the optimization problem

$$\min\left\{(x-5)^2 + (y-5)^2 \,\middle|\, x \geq 0, y \geq 0, x^2+y^2 \leq 25\right\}.$$

Show that your guess was correct! ♠

9.4 KKT

The KKT in the title of this section is short for Karush-Kuhn-Tucker.

We will limit ourselves to a convex optimization problem of the form

$$\begin{array}{ll} \text{Minimize} & f(x_1,\ldots,x_n) \\ \text{with constraint} & \\ & (x_1,\ldots,x_n) \in C, \end{array} \tag{9.23}$$

where C is defined by the differentiable convex functions $g_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1,\ldots,m$ as

$$C = \{v \in \mathbb{R}^n \mid g_1(v) \leq 0, \ldots, g_m(v) \leq 0\}$$

and $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function.

To the optimization problem (9.23) we associate the (famous) Karush-Kuhn-Tucker (KKT) conditions:

$$\begin{aligned}
&\lambda_1 \geq 0, \ldots, \lambda_m \geq 0 \\
&g_1(v_0) \leq 0, \ldots, g_m(v_0) \leq 0 \\
&\lambda_1 g_1(v_0) = 0, \ldots, \lambda_m g_m(v_0) = 0 \\
&\nabla f(v_0) + \lambda_1 \nabla g_1(v_0) + \cdots + \lambda_m \nabla g_m(v_0) = 0.
\end{aligned} \tag{9.24}$$

Notice that the KKT conditions consist of $2m$ inequalities and $m+n$ equations in the $m+n$ unknowns $\lambda_1, \ldots, \lambda_m, v_0 = (x_1, \ldots, x_n)$. The KKT conditions form a surprising theoretical foundation for optimization problems of the type in (9.23). You should take a peek back to the theory of Lagrange multipliers in section 7.9 and compare with (9.24).

(9.31) EXAMPLE.

The KKT conditions associated with the convex optimization problem in (9.1) are

$$\begin{aligned}
\lambda_1, \lambda_2, \lambda_3, \lambda_4 &\geq 0 \\
-x - y + 2 &\leq 0 \\
y - 2 &\leq 0 \\
x - 3 &\leq 0 \\
-y + 1 &\leq 0 \\
\lambda_1(-x - y + 2) &= 0 \\
\lambda_2(y - 2) &= 0 \\
\lambda_3(x - 3) &= 0 \\
\lambda_4(-y + 1) &= 0 \\
2x - \lambda_1 + \lambda_3 &= 0 \\
2y - \lambda_1 + \lambda_2 - \lambda_4 &= 0.
\end{aligned}$$

♠

(9.32) EXERCISE.

Verify that the KKT conditions of the optimization problem in (9.1) are the ones given in Example 9.31. ♠

To state our main theorem we need a definition.

(9.33) DEFINITION.

The optimization problem (9.23) *is called* strictly feasible *if there exists* $z_0 \in \mathbb{R}^n$ *with*

$$\begin{aligned}
g_1(z_0) &< 0 \\
&\vdots \\
g_m(z_0) &< 0.
\end{aligned}$$

Below is the main result in our limited convex setting. We will not go into the proof, which can be found in my book Undergraduate Convexity.

(9.34) THEOREM.

(i) Let v_0 be an optimal solution of (9.23). If (9.23) is strictly feasible, then the KKT conditions are satisfied at v_0 for suitable $\lambda_1, \ldots, \lambda_m$.

(ii) If the KKT conditions are satisfied at $z \in \mathbb{R}^n$ for some $\lambda_1, \ldots, \lambda_m$, then z is an optimal solution to (9.23).

(9.35) EXAMPLE.

Let us now touch base with a rather simple example. Consider the optimization problem

$$\begin{array}{ll} \text{Minimize} & x \\ \text{with constraint} & \\ & x \in [1,2]. \end{array} \tag{9.25}$$

Here $f(x) = x, g_1(x) = -x+1$ and $g_2(x) = x-2$ in (9.23). Therefore the KKT conditions in (9.24) are

$$\begin{aligned} \lambda_1 &\geq 0 \\ \lambda_2 &\geq 0 \\ -x+1 &\leq 0 \\ x-2 &\leq 0 \\ \lambda_1(-x+1) &= 0 \\ \lambda_2(x-2) &= 0 \\ 1-\lambda_1+\lambda_2 &= 0. \end{aligned} \tag{9.26}$$

Before even thinking about moving on to the next section, you should attempt to find a solution x, λ_1, λ_2 to the above KKT conditions (inequalities) and then verify using Theorem 9.34ii that x is optimal. Also, try only using Theorem 9.34i and (9.26) to show that $x = 2$ is not a solution to (9.25).

♠

(9.36) EXERCISE.

Give an example of a convex optimization problem as in (9.23), which is not strictly feasible and with an optimal solution v_0 that does not satisfy the KKT conditions. Such an example shows that strict feasibility is necessary in Theorem 9.34i. ♠

9.5 Computing with KKT

9.5.1 Strategy

A general strategy for finding solutions to the KKT conditions in (9.24) is zooming in on (the Lagrange multipliers) $\lambda_1, \ldots, \lambda_m$ testing each of them for the two cases $\lambda_i = 0$ and $\lambda_i > 0$.

One important point, which you can read from (9.24), is that $g_i(v_0) = 0$ if $\lambda_i > 0$. To further elaborate, if $\lambda_i > 0$, then an optimal solution must satisfy $g_i(v_0) = 0$.

(9.37) EXERCISE.

So where exactly in (9.24) is the above claim verified? ♠

The condition $\lambda_i = 0$ simplifies the equations

$$\nabla f(v_0) + \lambda_1 \nabla g_1(v_0) + \cdots + \lambda_m \nabla g_m(v_0) = 0$$

in (9.24).

In principle to solve the KKT conditions, one has to try out all the 2^m possibilities coming from $\lambda_i = 0$ or $\lambda_i > 0$ for $i = 1, \ldots, m$.

(9.38) EXERCISE.

Why 2^m possibilities above? ♠

(9.39) EXERCISE.

How do you solve the optimization problem (or decide there is no solution) if $\lambda_1 = \cdots = \lambda_m = 0$? ♠

9.5.2 Example

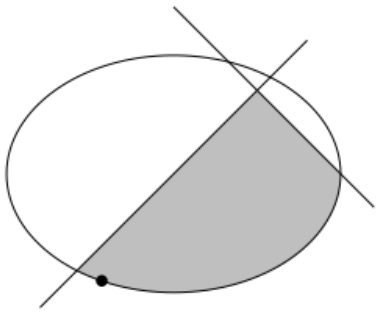
Let C denote the set (see Figure 9.40) of points $(x, y) \in \mathbb{R}^2$ with

$$\begin{aligned} x^2 + 2y^2 &\leq 1 \\ x + y &\leq 1 \\ y &\leq x. \end{aligned}$$

We will illustrate the mechanics of solving the KKT conditions in finding an optimal solution for

$$\begin{array}{ll} \text{Minimize} & x + 3y \\ \text{with constraint} & \\ & (x, y) \in C. \end{array} \tag{9.27}$$

(9.40) FIGURE.



The convex set C with optimal solution for (9.27) marked.

Putting

$$\begin{aligned} g_1(x,y) &= x^2+2y^2-1 \\ g_2(x,y) &= x+y-1 \\ g_3(x,y) &= y-x \end{aligned}$$

and $f(x,y)=x+3y$, we are in a position to apply Theorem 9.34, since g_1,g_2,g_3 are convex functions and $g_1(z_0)<0, g_2(z_0)<0, g_3(z_0)<0$ for $z_0=(0,-\frac{1}{2})$. This means that an optimal solution of (9.27) satisfies the KKT conditions. The same theorem tells us that the x_0 in a solution of the KKT conditions is an optimal solution (here we also use that f is a convex function). The full set of KKT conditions in $x,y,\lambda_1,\lambda_2,\lambda_3 \in \mathbb{R}$ are

$$\begin{aligned} x^2+2y^2-1 &\leq 0 \\ x+y-1 &\leq 0 \\ y-x &\leq 0 \\ \lambda_1,\lambda_2,\lambda_3 &\geq 0 \\ \lambda_1(x^2+2y^2-1) &= 0 \\ \lambda_2(x+y-1) &= 0 \\ \lambda_3(-x+y) &= 0 \\ 1+2\lambda_1 x+\lambda_2-\lambda_3 &= 0 \\ 3+4\lambda_1 y+\lambda_2+\lambda_3 &= 0. \end{aligned}$$

A strategy for finding a solution to the KKT conditions is trying (the eight) different combinations of strict inequalities in $\lambda_1,\lambda_2,\lambda_3 \geq 0$. You can see from the last two equations that $\lambda_1=0$ is impossible. The condition $\lambda_1>0$ shows that an optimal solution has to occur on the lower arc in Figure 9.40. If $\lambda_3>0$, then $x=y$ and $\lambda_2=1+3\lambda_3>0$ by the last two equations. This implies $x=y=\frac{1}{2}$ violating $x^2+2y^2-1=0$. Therefore $\lambda_3=0$. If $\lambda_2>0$, then $y=1-x$ and $5+4\lambda_1+3\lambda_2=0$ by $\lambda_3=0$ and the last two equations. Therefore $\lambda_2=0$. So we are left with the case $\lambda_1>0$ and $\lambda_2=\lambda_3=0$ giving

$$x=-\frac{1}{2\lambda_1} \quad \text{and} \quad y=-\frac{3}{4\lambda_1}.$$

Inserting this into $x^2+2y^2-1=0$ we end up with (see Figure 9.40)

$$\lambda_1=\frac{\sqrt{11}}{2\sqrt{2}}, \quad x=-\sqrt{\frac{2}{11}} \quad \text{and} \quad y=-\frac{3}{\sqrt{22}}.$$

Theorem 9.34 is beautiful mathematics. Going through the KKT conditions as above can be quite lengthy if not impossible in practice. As we have seen, there are other methods for (at least) approximating an optimal solution.

9.6 Optimization exercises

Below are some exercises especially related to the KKT conditions. In some of the exercises the minimization problem

$$\begin{aligned} &\text{Minimize} && f(x_1,\ldots,x_n) \\ &\text{with constraint} && \\ & && (x_1,\ldots,x_n)\in C, \end{aligned} \tag{9.28}$$

is denoted

$$\min\{f(x_1,\ldots,x_n) \mid (x_1,\ldots,x_n)\in C\}.$$

This should cause no confusion.

(9.41) EXERCISE.

Consider the optimization problem

$$\begin{array}{ll} \text{Minimize} & -x+y \\ \text{with constraints} & \\ & x^2+y^2 \le 1 \\ & (x+1)^2+(y-1)^2 \le 1 \end{array} \tag{9.29}$$

(a) Show that (9.29) is a convex optimization problem.

(b) Sketch the set of constraints in $\mathbb{R}^2$ and show that $\left(-\frac{1}{2},\frac{1}{2}\right)$ cannot be an optimal solution to (9.29).

(c) Write up the KKT conditions for (9.29) and explain theoretically (without actually solving them) why they must have a solution.

(d) Now solve (9.29). Is the solution unique?

♠

(9.42) EXERCISE.

Consider the function $f:\mathbb{R}^3\to\mathbb{R}$ given by

$$f(x_1,x_2,x_3)=2x_1^2+3x_2^2+4x_3^2.$$

(a) Show that f is strictly convex.

(b) Let $S\subseteq\mathbb{R}^3$ denote the subset of points $(x_1,x_2,x_3)\in\mathbb{R}^3$ satisfying

$$\begin{aligned} x_1+x_2+x_3 &\ge 1 \\ x_1+2x_2+3x_3 &\le 5 \end{aligned}$$

Show that S is a closed convex subset.

(c) Solve the optimization problem

$$\begin{array}{ll} \text{Minimize} & f(x_1,x_2,x_3) \\ \text{with constraints} & \\ & (x_1,x_2,x_3)\in S \end{array} \tag{9.30}$$

♠

(9.43) EXERCISE.

Let $S\subseteq B\subseteq\mathbb{R}^3$, where

$$\begin{aligned} S &= \{(x,y,z)\in\mathbb{R}^3 \mid x^2+y^2+z^2=1\} \qquad \text{and} \\ B &= \{(x,y,z)\in\mathbb{R}^3 \mid x^2+y^2+z^2\le 1\}. \end{aligned}$$

(a) Why does the optimization problem

$$\begin{array}{ll} \text{Minimize} & x+2y+z \\ \text{with constraints} & \\ & (x,y,z) \in S \end{array} \tag{9.31}$$

have a solution?

(b) Find all optimal solutions to (9.31).

(c) Let $a,b,c \in \mathbb{R}$, where at least one of a,b,c is non-zero. Show that an optimal solution to

$$\begin{array}{ll} \text{Minimize} & ax+by+cz \\ \text{with constraints} & \\ & (x,y,z) \in B \end{array}$$

belongs to S.

♠

(9.44) EXERCISE.

Let

$$S = \left\{ (x,y) \in \mathbb{R}^2 \middle| \begin{array}{rcrcl} -x & - & y & \leq & 0 \\ 2x & - & y & \leq & 1 \\ -x & + & 2y & \leq & 1 \end{array} \right\}.$$

1. Use the KKT conditions to solve the minimization problem

$$\min\{-x-4y | (x,y) \in S\}.$$

2. Use the KKT conditions to solve the minimization problem

$$\min\{x+y | (x,y) \in S\}.$$

♠

(9.45) EXERCISE.

Solve the optimization problem

$$\min\left\{ x^2+2y^2+3z^2-2xz-xy \middle| \begin{array}{ccc} 2x^2+y^2+z^2 & \leq & 4 \\ 1 & \geq & x+y+z \end{array} \right\}.$$

♠

(9.46) EXERCISE.

Let $S = \left\{(x,y) \middle| 2x^2+y^2 \leq 3, x^2+2y^2 \leq 3\right\}$ and $f(x,y) = (x-4)^2+(y-4)^2$.

1. State the KKT conditions for $\min\{f(x,y) | (x,y) \in S\}$ for $(x,y) = (1,1)$.

2. Suppose now that $g(x,y) = (x-a)^2 + (y-b)^2$. For which a and b does $\min\{g(x,y)|(x,y) \in S\}$ have optimum in $(1,1)$? State the KKT conditions when $(a,b) = (1,1)$.

♠

(9.47) EXERCISE.

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$f(x,y) = (x-1)^2 + (y-1)^2 + 2xy.$$

1. Show that f is a convex function.
2. Find $\min\left\{f(x,y)\middle|(x,y) \in \mathbb{R}^2\right\}$. Is this minimum unique? Is f a strictly convex function.

   Let

   $$S = \left\{(x,y) \in \mathbb{R}^2\middle|x+y \leq 0, \quad x-y \leq 0\right\}.$$

3. Apply the KKT-conditions to decide if $(-1,-1)$ is an optimal solution to

   $$\min\{f(x,y)|(x,y) \in S\}.$$

4. Find

   $$m = \min\{f(x,y)|(x,y) \in S\}$$

   and

   $$\left\{(x,y) \in \mathbb{R}^2\middle|f(x,y) = m\right\}.$$

♠

(9.48) EXERCISE.

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$f(x,y) = x^2 + y^2 - e^{x-y-1}$$

and let

$$C = \{(x,y)|x-y \leq 0\}.$$

1. Show that $f : \mathbb{R}^2 \to \mathbb{R}$ is not a convex function.
2. Show that f is a convex function on the open subset

   $$\left\{(x,y) \in \mathbb{R}^2\middle|x-y < \tfrac{1}{2}\right\}$$

   and conclude that f is convex on C.
3. Show that $v = (0,0)$ is an optimal solution for the optimization problem $\min\{f(v)|v \in C\}$. Is v a unique optimal solution here?

♠

(9.49) EXERCISE.

Let $f : \mathbb{R}^4 \to \mathbb{R}$ be given by

$$f(x_1,x_2,x_3,x_4) = (x_1-x_3)^2 + (x_2-x_4)^2$$

and $C \subseteq \mathbb{R}^4$ by

$$C = \left\{(x_1,x_2,x_3,x_4) \in \mathbb{R}^4\middle|x_1^2 + (x_2-2)^2 \leq 1, x_3 - x_4 \geq 0\right\}.$$

1. Show that f is a convex function. Is f strictly convex?
2. Show that C is a convex subset of $\mathbb{R}^4$.
3. Does there exist an optimal point $v = (x_1, x_2, x_3, x_4) \in \mathbb{R}^4$ for the minimization problem

$$\min_{v \in C} f(v)$$

with $x_3 = x_4 = 0$?

4. Does there exist an optimal point $v = (x_1, x_2, x_3, x_4) \in \mathbb{R}^4$ for the minimization problem

$$\min_{v \in C} f(v)$$

with $x_3 = x_4 = 1$?

♠

(9.50) EXERCISE.

Let

$$f(x,y) = (x-1)^2 + y^2$$

and

$$C = \left\{(x,y) \in \mathbb{R}^2 \middle| -1 \leq x \leq 0,\ -1 \leq y \leq 1\right\}.$$

Solve the optimization problem

$$\min\left\{f(x,y) | (x,y) \in C\right\}.$$

♠

(9.51) EXERCISE.

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$f(x,y) = \tfrac{1}{2}x^2 + y^2 - 2y + 2.$$

Below, the minimization problem

$$\min\left\{f(x,y) | (x,y) \in S\right\} \tag{9.32}$$

is analyzed for various subsets $S \subseteq \mathbb{R}^2$.

1. Show that f is a convex function
2. Let

$$S = \left\{(x,y) \in \mathbb{R}^2 \middle| -x + 2y \leq 1\right\}.$$

Show that $(-1,0) \in S$ cannot be an optimal solution to (9.32). Find an optimal solution to (9.32).

3. Find an optimal solution in (9.32) for

$$S = \left\{(x,y) \in \mathbb{R}^2 \middle| -x + 2y \geq 1\right\}.$$

4. Are the optimal solutions in 2 and 3 unique?

♠

(9.52) EXERCISE.

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$f(x,y) = 2x^2 + 3x + y^2 + y.$$

1. Show that f is a convex function and solve the minimization problem $\min\{f(x,y)|x,y \in \mathbb{R}\}$.
   Now let
   $$S = \left\{ (x,y) \in \mathbb{R}^2 \middle| \begin{array}{rcl} x^2 + (y+1)^2 & \leq & 1 \\ y - x & \leq & 0 \end{array} \right\}$$
   and consider the minimization problem (P) given by
   $$\min\{f(x,y)|(x,y) \in S\}.$$
2. Show using the KKT conditions that $(0,0)$ is not optimal for (P).
3. Find an optimal solution for (P). Is it unique?

♠