Mathematical Physics Theory 2003

Martin Plenio
Imperial College London

Version May 27, 2003

Office: Blackett 622

Contents

Introduction

In this course I will provide you with a range of mathematical ideas and tools. It should be clear from the beginning however, that I do not only want to present you with recipes that you learn by heart and then apply. Of course you should simply know certain techniques and concepts. But, much more important than this I would like to imprint on you a mathematical way of thinking. This means that you should be able to think logically and be able to create proofs for hunches that you may have. These proofs have to be written out in an impeccable logical sequence that holds up to standards of proof that we have in theoretical physics. However, proofs in theoretical physics sometimes do not quite conform to the strict requirements of mathematics in that certain things may be taken for granted by a physicist that a mathematician would endeavour to prove. These are the situations where the physicist animatedly waves his or her hands to describe some fluffy not really strict argument. This is always dangerous even for the mathematically competent physicist but even more so for the physicist whose understanding of mathematics is less well developed. There are many potential pitfalls and usually one only learns to handwave correctly once one has made a lot of strict proofs in an area. Only then does one have the experience and feeling to see the right results without proving the strictly. Unfortunately, in England it is rare that physics students actually have to take mathematics lectures and as a consequence there is the danger that they are never exposed to stringent mathematical thinking. To make matters worse, you are then often encouraged to wave your hands. This approach, in my opinion, has a high propensity for leading to disaster and the most dangerous kind of half-knowledge which is the one where the owner thinks he knows but actually doesn't know[1].

This course aims to expose you a little bit more to mathematical thinking. Unfortunately, the time that is available for this course is not sufficient to go into very much depth in the various subjects that I am

[1]Less dangerous are persons that do not know something properly but are aware of this. Even less dangerous are people who know something and are aware of it. The greatest danger for them is to generalize the idea that they know things to other situations where they haven't got a clue.

going to present. However, what I will try to do is to avoid handwaving or to use it only to make a mathematical proof or definition more intuitive. I will not quite achieve the stringency of mathematicians as this would require too much depth and time but I will suggest literature that will go all the way and I suggest to you to have a look at some of these.

The downside of this approach is that it is not easy and may appear dry in places. But I am sure that it is worth it, as I feel that my own exposure to many lectures in pure mathematics (taught by mathematicians) in my old university (I studied in Göttingen) has greatly benefited my work in theoretical physics.

Apart from an appreciation of the beauty of some parts of mathematics, it has helped me to learn to make strict proofs and to develop a reasonably reliable red alert light in my head that flashes when I am waving my hands to vigorously. The instances where it fails I will inevitably be told off by colleagues and PhD sudents who have developed their own alert devices and unveil these mistakes by their critical questions. This brings me to the last point. Feel free to ask questions before, after and during the lectures. Don't worry that some of your questions may turn out to be not terribly deep. It doesn't matter. I do not worry either about making myself a fool in front of my PhD students by asking 'silly' questions.

Ok, thats it. I hope that you will enjoy the lectures.

Chapter 1

Sets, numbers and the concept of infinity

When we describe Nature using mathematics we are generally using natural, rational, real and complex numbers (and maybe quaternions) and manipulate them according to general principles. In this chapter I would like to introduce all these numbers starting from a little bit of set theory to develop the natural numbers and then go over to rational numbers etc. In the process I will introduce ideas such as sequences, convergence etc which are of great use in mathematics and physics. But not only that, these ideas will also force us to consider such weird things as infinities and completeness of numbers. So let us start the journey.

1.1 Sets

In the following I will introduce some very basic concepts of set theory. I do this for two reasons. Firstly, set theory can be used to form the basis for mathematics and in particular arithmetic and it provides a definition of what the natural numbers are. Secondly, it will allow me to give you a glimpse of a rather intriguing concept, namely that of infinity. In this first part of the lecture I will not go into all the depths of detail simply because this would require a full lecture course in itself. If you would like to know more about set theory and the theoretical

foundation of natural numbers, then I wholeheartedly recommend you to have a look at the book 'Naive Set Theory' by Paul R. Halmos which explains the basic ideas of set theory in a very narrative, yet precise, style.

In mathematics and therefore also in physics we describe properties of assemblies of entities. Usually we call these assemblies simply 'sets' and their constituent parts we call 'elements'. Put it slightly differently, the mathematician and philosopher Bolzano defines a set as 'an embodiment of the idea or concept which we conceive when we regard the arrangement of its parts as a matter of indifference'. So, a set is a collection of things, for example

$$\mathcal{S}_1 = \{\Delta, \square, \diamondsuit, \bigcirc\} \tag{1.1}$$

I have chosen these geometric objects to make clear that elements of a set can take any form. They do not need to be numbers at all, in fact, they may even be sets themselves. Note that I have not yet defined any properties of sets, so its not strictly clear that the above assembly is actually something, that the mathematicians call a set and indeed, we will soon realize that not every collection of elements is admissible as a set[1]. There are some shorthand notation for saying that '$\diamondsuit$ is an element of the set $\mathcal{S}_1$' namely

$$\diamondsuit \in \mathcal{S}_1 \,. \tag{1.2}$$

If '$\diamondsuit$ is not an element of the set $\mathcal{S}_1$' we write

$$\diamondsuit \notin \mathcal{S}_1 \,. \tag{1.3}$$

The first property of sets that one should define, is that of equality of two sets

Definition 1 *Two sets $\mathcal{A}$ and $\mathcal{B}$ are equal, $\mathcal{A} = \mathcal{B}$, if they contain the same elements.*

This is of course a natural definition, but it should nevertheless be made. Of course, you would like to have a rule that allows you to build new sets from a given set. We will need to define which sort of rules are acceptable. You may think that this is trivial, but later on I will show you that in fact it isn't.

[1]Rest assured however that indeed $\mathcal{S}_1$ is a set even according to mathematicians

Definition 2 *(Axiom of selection) For any given set $\mathcal{A}$ and any condition $S(x)$ there is a set $\mathcal{B}$ whose elements are exactly those x which are from $\mathcal{A}$ and for whom $S(x)$ is true.*

Example: Take $\mathcal{A} = \{1, 2, 3, \ldots\}$ to be the set of integers. Furthermore, take the condition $S(x)$ as the condition $x\,is\,even$. Then we obtain the new set $\mathcal{B}$ as

$$\mathcal{B} = \{x \in \mathcal{A} | S(x)\} = \{x \in \mathcal{A} | x\,is\,even\} = \{2, 4, 6, 8, \ldots\} \tag{1.4}$$

From this definition and the assumption that there exists at least one set we can now build new sets. Indeed, the simplest set that you can think of is the empty set which is the set that has no elements in it. This is usually denoted as $\emptyset$. Now let us assume that at least one set $\mathcal{A}$ exists. Then we can define the empty set a bit more formally by writing

$$\emptyset = \{x \in \mathcal{A} | x \neq x\} \tag{1.5}$$

Obviously no element can be in the so defined set, so it is empty. Note that the set that contains the empty set, ie $\{\emptyset\}$, obviously contains an element, namely the empty itself and therefore we have $\emptyset \neq \{\emptyset\}$.

So far this all looks a little bit like Kindergartenmaths. Well, let us look again at the definition of the empty set eq (1.5). In this definition I took great care to assume the existence of the set $\mathcal{A}$. Surely, any old set will do here. Indeed, you could think that I do not need to assume the existence of this set $\mathcal{A}$. Surely, x can be anything you like and there should be no restriction on it. Yes? Well, no. Let us see what happens when we make no restriction at all on x, or in other words, we assume that there is a set that contains everything.

If you don't spot the problem with this immediately, you are in good company. Indeed, in the beginnings of set theory mathematicians have made exactly this assumption without worrying too much. Unfortunately, we will now see that in fact, it would lead to catastrophical structural failure of the theory, ie it would be contradictory. This does not show up in the definition of the empty set but in the following example. Let us define the set

$$\mathcal{T} = \{x | x \notin x\} \tag{1.6}$$

Note that now I allow myself to chose any x that I like and test whether the condition $x \notin x$ is satisfied or not. This may look a bit weird in the first instance as we are treating the elements of the set as sets themselves. But that's not unnatural. After all you could imagine the set of all those sets that contain exactly one element. But the real trouble starts when you try to decide the question whether $\mathcal{T} \in \mathcal{T}$ or $\mathcal{T} \notin \mathcal{T}$. One of the two assumption should be correct. So, lets check them then. If $\mathcal{T} \in \mathcal{T}$ then $\mathcal{T}$ has to satisfy the defining condition for being a member of $\mathcal{T}$, namely $\mathcal{T} \notin \mathcal{T}$, but that's a contradiction. Equally, when we assume that $\mathcal{T} \notin \mathcal{T}$ then $\mathcal{T}$ satisfies the condition for being in $\mathcal{T}$ and should therefore be a member of $\mathcal{T}$. Another contradiction. So what has happened. Obviously something has gone wrong, but what? Well, it turns out that we made a mistake that all mathematicians had done in the early years of the development of set theory. Namely, we took something for granted which is not quite so natural. In this case it was the assumption that there is a set that contains everything. Such a set does not exist, meaning that it leads to logical contradictions. So, what we should write is the following

$$\mathcal{T} = \{x \in \mathcal{A} | x \notin x\} \tag{1.7}$$

for some set $\mathcal{A}$. Now we can run the same argument as above. If we try to assume that $\mathcal{T} \in \mathcal{T}$ then we obtain a contradiction. So, as a consequence $\mathcal{T} \notin \mathcal{T}$ and we realize that for every set there is something that is not included in it. In other words, **there is no set that contains everything**, ie the 'set of everything' does not exist. It took people many years to realize this and indeed, it was Bertrand Russell in 1894, and independently Zermelo, who found the above example and communicated it to Frege who was just finishing off a book in which he used set theory as the logical basis for mathematics. Large part of his work were invalidated by this paradox and it took mathematicians quite a while to put things right (see Halmos for details).

This first taste of the non-triviality of set theory teaches you two things. Firstly, don't take things for granted in mathematics without checking and of course it should also convince you that set theory is not quite so trivial. Now let us continue to define properties of sets. Firstly let us define the notion of subsets.

Definition 3 *A set $\mathcal{A}$ is called subset of a set $\mathcal{B}$, written $\mathcal{A} \subset \mathcal{B}$ if for all $x \in \mathcal{A}$ we have that $x \in \mathcal{B}$.*

Clearly you would like to combine different sets with each other. Of course this can be done and the idea is essentially based on the axiom of selection.

There are a range of operations that you can do and I present three of the basic and most important operations.

1. Union of sets: The union of two sets $\mathcal{A}$ and $\mathcal{B}$ is also a set which is written $\mathcal{A} \cup \mathcal{B}$. It is defined as

$$\mathcal{A} \cup \mathcal{B} = \{x | x \in \mathcal{A} \text{ or } x \in \mathcal{B})\} \tag{1.8}$$

Figure 1.1: The Venn diagram that illustrates the union of two sets $\mathcal{A}$ and $\mathcal{B}$.

2. Intersection of two sets: The intersection of two sets $\mathcal{A}$ and $\mathcal{B}$ is also a set written $\mathcal{A} \cap \mathcal{B}$. It is defined as

$$\mathcal{A} \cap \mathcal{B} = \{x | x \in \mathcal{A} \text{ and } x \in \mathcal{B})\} \tag{1.9}$$

3. Difference between sets: The difference between two sets $\mathcal{A}$ and $\mathcal{B}$ is also a set written $\mathcal{A} \setminus \mathcal{B}$. It is defined as

$$\mathcal{A} \setminus \mathcal{B} = \{x | x \in \mathcal{A} \text{ and } x \notin \mathcal{B})\} \tag{1.10}$$

Figure 1.2: The Venn diagram that illustrates the intersection of two sets $\mathcal{A}$ and $\mathcal{B}$.

Figure 1.3: The Venn diagram that illustrates the subtraction of two sets $\mathcal{A}$ and $\mathcal{B}$.

Note that in all these definitions I make a statement of the form '... the union of two sets is also a set ...' which is a statement of existence. Again this is necessary as it helps us to define which sets are admissible and which ones aren't. After all, we do not know the properties of sets, but we have to define them in a way that appears natural. The above definitions appear natural but nevertheless they are definitions and not fundamental truths. Before you dismiss this as a triviality you should always remember the fact that a little bit earlier we discovered that the set that contains everything does not exist.

To get you familiarized with the operations of forming the 'union, intersection and subtraction' of sets let me give you a range of examples.

1. $\mathcal{A} \cup \emptyset = \mathcal{A}$
   Proof: We have to check that $x \in \mathcal{A} \cup \emptyset \Leftrightarrow x \in \mathcal{A}$. This can be seen via

(i) $x \in \mathcal{A} \cup \emptyset \Rightarrow x \in \mathcal{A}$ or $x \in \emptyset \Rightarrow x \in \mathcal{A}$
(ii) $x \in \mathcal{A} \quad \Rightarrow x \in \mathcal{A}$ or $x \in \emptyset \Rightarrow x \in \mathcal{A} \cup \emptyset$
So we satisfy both criteria for equality of two sets and we are finished.

2. $\mathcal{A} \cap \emptyset = \emptyset$
3. $\mathcal{A} \cup \mathcal{B} = \mathcal{B} \cup \mathcal{A}$
4. $\mathcal{A} \cap \mathcal{B} = \mathcal{B} \cap \mathcal{A}$
5. For any pair of sets we have $\mathcal{A} \subset \mathcal{A} \cup \mathcal{B}$
6. If $\mathcal{A} \subset \mathcal{B}$ and $\mathcal{B} \subset \mathcal{A}$ then $\mathcal{A} = \mathcal{B}$.
7. $\mathcal{A} \cup (\mathcal{B} \cap \mathcal{A}) = \mathcal{A}$
Proof: We have to check that $x \in \mathcal{A} \cup (\mathcal{B} \cap \mathcal{A}) \Leftrightarrow x \in \mathcal{A}$. This follows from
(i) $x \in \mathcal{A} \cup (\mathcal{B} \cap \mathcal{A}) \Leftrightarrow x \in \mathcal{A}$ or $(x \in \mathcal{A}$ and $x \in \mathcal{B})$. Clearly, this implies either that $x \in \mathcal{A}$ or that we have $(x \in \mathcal{A}$ and $x \in \mathcal{B})$ which again implies $x \in \mathcal{A}$. Therefore $x \in \mathcal{A} \cup (\mathcal{B} \cap \mathcal{A}) \Rightarrow x \in \mathcal{A}$. As a consequence we have $\mathcal{A} \cup (\mathcal{B} \cap \mathcal{A}) \subset \mathcal{A}$
(ii) On the other hand we have for any set $\mathcal{X}$ that $\mathcal{A} \subset \mathcal{A} \cup \mathcal{X}$. In particular if we set $\mathcal{X} = (\mathcal{B} \cap \mathcal{A})$ we have $\mathcal{A} \subset \mathcal{A} \cup (\mathcal{B} \cap \mathcal{A})$.
8. $\mathcal{A} \cap (\mathcal{B} \cup \mathcal{A}) = \mathcal{A}$
9. $(\mathcal{A} \cup \mathcal{B}) \setminus \mathcal{C} = (\mathcal{A} \setminus \mathcal{C}) \cup (\mathcal{B} \setminus \mathcal{C})$
10. $(\mathcal{A} \cap \mathcal{B}) \setminus \mathcal{C} = (\mathcal{A} \setminus \mathcal{C}) \cap (\mathcal{B} \setminus \mathcal{C})$
11. $\mathcal{A} \cup (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cup \mathcal{B}) \cup \mathcal{C}$
12. $\mathcal{A} \cup (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \cup \mathcal{B}) \cap (\mathcal{A} \cup \mathcal{C})$
13. $\mathcal{A} \cap (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{C})$
14. $\mathcal{A} \setminus (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \setminus \mathcal{B}) \cap (\mathcal{A} \setminus \mathcal{C})$
15. $\mathcal{A} \setminus (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \setminus \mathcal{B}) \cup (\mathcal{A} \setminus \mathcal{C})$
16. $\mathcal{A} \setminus (\mathcal{B} \setminus \mathcal{C}) = (\mathcal{A} \setminus \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{C})$

As an additional exercise, try to visualize the proofs by drawing Venn diagrams that show the equality as well.

Now, as promised I introduce the natural numbers from the concept of sets. Again, we have to ask ourselves what are the basic properties of natural numbers and we would like to use them to make a construction that we then call 'natural numbers'. Well, the main property of natural numbers is that we use them for counting, ie for every number there is a successor and, of course, there is a smallest (first) natural number. We will define the natural numbers starting from only a single type of sets, namely the empty set and we build the remaining numbers via the union of sets. We start by defining

$$0 = \emptyset \,. \tag{1.11}$$

Furthermore, we say that the successor x^+ of a number is defined as

$$x^+ = x \cup \{x\} \,. \tag{1.12}$$

For example

$$\begin{aligned} 1 &\equiv 0^+ = 0 \cup \{0\} = \{0\} = \{\emptyset\} && (1.13)\\ 2 &\equiv 1^+ = 1 \cup \{1\} = \{0,1\} = \{\emptyset, \{\emptyset\}\} && (1.14)\\ 3 &\equiv 2^+ = 2 \cup \{2\} = \{0,1,2\} && (1.15)\\ &\vdots \end{aligned}$$

The set of all the so defined numbers is then given by

$$\omega = \{0, 1, 2, \ldots\} \,. \tag{1.16}$$

This is in the smallest set that contains with every element also its successor. The size of this set is the natural definition for infinity as it clearly does not contain finitely many elements. Note however, that we actually have to define that this is an admissible set. After all so far we have really only dealt with sets that have finitely many elements and what tells you that you are not running into problems or contradictions when you consider sets with infinitely elements. The axiom that the set ω is admissible is sometimes called the axiom of infinity in set theory, for obvious reasons.

1.2 From counting to infinities

In the previous section I have introduced the natural numbers by formalizing the idea of counting and even got a first idea of how to define an infinite number. Now let us assume that the set of natural numbers is given and let us call it

$$\mathbb{N} = \{1, 2, 3, 4, \ldots\} \tag{1.17}$$

Note that from now on I exclude the 0 from the set of natural numbers. I do this because in the following I will talk about counting and very few of you will count "zero, one, two, ..." but rather "one, two, three, ...". In a moment we will see that this does not change the concept of infinity, ie that I will show you in a moment that infinity and infinity + 1 are just the same.[2].

In order to study these questions carefully, we need to be able to compare the size of sets and in particular to compare the size of sets to the size of ω also commonly known as counting. How does this go? Let us consider a set $\mathcal{A}$ whose size we wish to compare with $\mathbb{N}$. We count by picking the first element of $\mathbb{N}$ and associate with it an element of the set $\mathcal{A}$, namely a_1. Then we pick the next element of $\mathbb{N}$, namely 1, and associate with it one of the remaining elements of $\mathcal{A}$, ie an element of $\mathcal{A} \setminus \{a_1\}$[3]. What we are doing here is to make a one-to-one correspondence between the two sets in such a way that no element of the two sets are left out. Mathematicians say that we construct a bijective map between the two sets. What is this? Here is the formal definition

Definition 4 *A map f that maps a set $\mathcal{A}$ to a set $\mathcal{B}$, ie to every element $x \in \mathcal{A}$ it associates another element $f(x) \in \mathcal{B}$, is called bijective if*

[2]Natural numbers are the most basic concept of numbers and is often regarded as the most pure mathematical concept, or to quote Leopold Kronecker (the one from the Kronecker symbol and many other things) "Die ganzen Zahlen hat der liebe Gott gemacht, alles andere ist Menschenwerk" ("The natural numbers have been created by god, all the rest are human constructions") implying the imperfection of any concept different from natural numbers

[3]Note that this prescription somehow requires that the two sets are ordered, indeed this is essentially the requirement, that each subset of the set contains a 'smallest' element. As a consequence, one can order the elements according to their size.

(i) the map f is injective, ie for any pair $x, y \in \mathcal{A}$ with $x \neq y$ we have that $f(x) \neq f(y)$
and
(ii) the map f is surjective, ie for any pair $z \in \mathcal{B}$ there is an $x \in \mathcal{A}$ such that $f(x) = z$.

Now we can give the formal definition that two sets have the same size, namely

Definition 5 *Two sets $\mathcal{A}$ and $\mathcal{B}$ are said to have the same size, $\mathcal{A} \sim \mathcal{B}$, if there is a bijective map f between them.*
We say that $\mathcal{A}$ is at most as large as $\mathcal{B}$, $\mathcal{A} \preceq \mathcal{B}$, is there is an injective map f from $\mathcal{A}$ to $\mathcal{B}$.
We say that $\mathcal{A}$ is at least as large as $\mathcal{B}$, $\mathcal{B} \preceq \mathcal{A}$, is there is a surjective map f from $\mathcal{A}$ to $\mathcal{B}$.

The last two statement can also be formulated in a slightly different way. If a set $\mathcal{A}$ has the same size as a proper subset of $\mathcal{B}$, ie a set that does not contain all the elements of $\mathcal{B}$, then we say it is of smaller or equal extent [4] and vice versa and write $\mathcal{A} \preceq \mathcal{B}$.

Now let us get a better grip on the concept of infinity by first saying that the size of the set $\mathbb{N}$ is infinite. We will denote the size of the set of natural numbers by $\aleph_0$ (this letter is Greek and is pronounced aleph) . Now we can start to consider other sets and compare their size to that of $\mathbb{N}$, ie we will count these other sets. This approach will then allow us to find out some basic properties of infinities such as $\aleph_0$ and it will furthermore help us later on to illuminate the question whether there are different degrees of infinity. Below I write a few sets.

$$\mathcal{S}_1 = \{n \in \mathbb{N} | n \neq n\} \tag{1.18}$$
$$\mathcal{S}_2 = \{n \in \mathbb{N} |\, \exists \text{natural number } m \text{ such that } 120 = mn\} \tag{1.19}$$
$$\mathcal{S}_3 = \{n \in \mathbb{N} |\, \exists \text{natural number } m \text{ such that } m^2 = n\}. \tag{1.20}$$

Clearly, the first set is actually the empty set $\emptyset$ and has 0 elements. The second set $\mathcal{S}_2 = \{1, 2, 3, 4, 5, 6, 8, 10, 12, 15, 20, 24, 30, 40, 60, 120\}$ has 16 elements which is meant to say that it has the same number of

[4] Note that we do not conclude that the set $\mathcal{A}$ is truly smaller than $\mathcal{B}$. Indeed in a moment we will see that this is not necessarily true.

elements as the set $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16\}$. Both sets are finite. We count them by picking the first natural number 1 and associate with it one element of the set. Then we take number 2 and associate with it another element of the set. We continue with this until we have completed the numbering. For the set $\mathcal{S}_2$ for example this means

$\mathcal{S}_2$	$\mathbb{N}$
1	1
2	2
3	3
4	3
5	5
6	6
8	7
10	8
12	9
15	10
20	11
24	12
30	13
40	14
60	15
120	16

The procedure that we have presented here is an example for counting as I explained at the end of the last section.

Note that I did not make any requirement concerning the size of the two sets. They can be finite or infinite. An example for an infinite set is $\mathcal{S}_3$. Let us see how its size compares to the one of $\mathbb{N}$. To this end we try to establish a one-to-one correspondence between the elements with the following table

$\mathcal{S}_3$	$\mathbb{N}$
1	1
4	2
9	3
16	4
25	5
36	6
49	7
64	8
$\vdots$	$\vdots$

Here we see that to every element of $\mathcal{S}_3$ we can associate one element of $\mathbb{N}$ in a one-to-one manner without leaving out a single element of each set. More formally we define the map f from $\mathbb{N}$ to $\mathcal{S}_3$ via $f(n) = n^2$. Clearly this map is injective on $\mathbb{N}$ because for $m \neq n$ we have $m^2 \neq n^2$. So $\mathbb{N} \preceq \mathcal{S}_3$. We could now proceed to show that the map is also surjective, to show that the two sets have the same size, but here I proceed slightly differently. I rather show that we can define a map g from $\mathcal{S}_3$ to $\mathbb{N}$ via $g(x) = +\sqrt{x}$ where $x \in \mathcal{S}_3$. Again, it is clear straightaway, that this map is injective, so that we have $\mathcal{S}_3 \preceq \mathbb{N}$. Given that we have $\mathbb{N} \preceq \mathcal{S}_3$ and $\mathcal{S}_3 \preceq \mathbb{N}$ we have that the two sets have the same size, ie $\mathcal{S}_3 \sim \mathbb{N}$.[5]

Any set for which we can find such an association between the elements of the set and those of the natural numbers we call *countable*. Sometimes we will see that we do not need all the elements of $\mathbb{N}$ or more precisely we need only a finite number of them. That is when we say that the set is finitely countable. The sets $\mathcal{S}_1$ and $\mathcal{S}_2$ are of that type. The set $\mathcal{S}_3$ is different however. Here we really need *all* the elements of $\mathbb{N}$. For such a set we say that it is countably infinite and its size is denoted by $\aleph_0$. The way I counted the elements of $\mathcal{S}_3$ is by writing a table but it is more convincing to write an analytical connection between the natural numbers and the elements of $\mathcal{S}_3$. If I call the elements of $\mathcal{S}_3$ that I wish to correspond to the natural number

[5]Note that in a more stringent formulation of set theory than the one I presented here, this line of argument actually needs to be proven and is called the theorem of Schröder and Bernstein. Namely, two sets $\mathcal{A}$ and $\mathcal{B}$ have the same size exactly if $\mathcal{A} \preceq \mathcal{B}$ and $\mathcal{B} \preceq \mathcal{A}$.

n by a_n, then I find $a_n = n^2$. Now it becomes clear that there is a one-to-one correspondence between the elements of $\mathcal{S}_3$ and $\mathbb{N}$.

Quite obviously, finitely countable sets are rather boring. Therefore, let us explore a few more examples of infinite sets to learn some of the computational rules for infinities. In fact, let us consider the set $\mathbb{N}$ and let us add a single element to it, namely 0. The new set is then $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Now the natural expectation is that the second set $\mathbb{N}_0$ has more elements than $\mathbb{N}$. Certainly, it does not have less elements, ie $\mathbb{N} \preceq \mathbb{N}_0$. But does it really have more elements? To decide this question, let us count them. When there is a one-to-one correspondence between $\mathbb{N}_0$ and $\mathbb{N}$ then they have the same number of elements. Let us assume the elements of the two sets are nicely ordered and $a_n \in \mathbb{N}$ and $b_n \in \mathbb{N}_0$. Clearly, $a_1 = 1$ and $b_1 = 0$. In general we have $a_n = b_n + 1$ so that indeed we have a one-to-one correspondence and the sets have the same (infinite) number of elements. In an equation this statement reads

$$1 + \aleph_0 = \aleph_0 \tag{1.21}$$

As another example let me introduce the integers which is the set

$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\} \tag{1.22}$$

Can you find a way to enumerate the integers, ie can you show that the size of this set is again just as large as the natural numbers? The solution in tabulated form is

$\mathcal{S}_3$	$\mathbb{N}$
0	1
1	2
-1	3
2	4
-2	5
3	6
-3	7
4	8
-4	9
5	10
-5	11
$\vdots$	$\vdots$

from which it is self-evident that the two sets are equal in size. Drawing the table is a first step to a proper proof which amounts to writing the rule for associating natural numbers one-to-one with integers. Formally we find

$$a_1 = 1 \quad a_{2n} = n \quad a_{2n+1} = -n \tag{1.23}$$

ie we have defined a sequence that runs through all the integers[6]. So we have a one-to-one mapping between the set of integers and the natural numbers.

As we doubled the set and added one more element (the 0) this corresponds to the equation

$$2 \cdot \aleph_0 + 1 = \aleph_0 \tag{1.24}$$

Now, let us make the set of natural numbers smaller by leaving out every second element, i.e. let us consider the set of even numbers

$$\mathcal{S}_{even} = \{n | \exists m \in \mathbb{N} \,\text{such that}\, n = 2m\}^7. \tag{1.25}$$

Quite clearly $\mathcal{S}_{even} \subset \mathbb{N}$ so you would say that it has less elements than the natural numbers. So let us count the set $\mathcal{S}_{even}$.

$\mathcal{S}_3$	$\mathbb{N}$
2	1
4	2
6	3
8	4
10	5
12	6
14	7
16	8
18	9
20	10
22	11
⋮	⋮

[6]I could have written this in a more compact form as $a_m = (-1)^m \lfloor \frac{m}{2} \rfloor$ where the special brackets $\lfloor x \rfloor$ mean that one choses the largest integer smaller than x. For example $\lfloor 1.5 \rfloor = 1$ and $\lfloor -1.2 \rfloor = -2$.

[7]The symbol $\exists$ means 'there exists'

So we see that there is a one-to-one correspondence for each of the elements and as a consequence both sets have the same number of elements which you should turn into an arithmetic expression, namely $n \mapsto a_n = 2n$. This can be summed up in the equation

$$\frac{\aleph_0}{2} = \aleph_0 \tag{1.26}$$

In general multiplying or dividing $\aleph_0$ by a finite natural number and adding a finite amount gives as a result $\aleph_0$ again.

Now we know that multiplying $\aleph_0$ by a finite number, we get $\aleph_0$ again, so no change. A less trivial question is that of the result of

$$\aleph_0 \cdot \aleph_0 = ??? \tag{1.27}$$

To make sense of this equation we first have to make sense of what it means to multiply by $\aleph_0$. If we multiply by 2 we add for every existing element of the set another one. To see a systematic way of doing this let us consider sets of the form

$$\mathcal{T}_1 = \{(1,1),(2,1),(3,1),(4,1),\ldots\} \tag{1.28}$$

where the objects in the round brackets such as $(3,1)$ form a single element of the set and the set is continued following the obvious rule. Obviously this is an infinite set and it has the same size as the natural numbers, ie $\mathbb{N} \sim \mathcal{T}_1$. Now let us double the set. One way to do this is

$$\mathcal{T}_2 = \{(1,1),(1,2),(2,1),(2,2),(3,1),(3,2),(4,1),(4,2),\ldots\} \tag{1.29}$$

This set has the size $2 \cdot \aleph_0$. Now let us write down the set which has the size $\aleph_0^2$. This is quite intuitively

$$\mathcal{T}_\infty = \left\{ \begin{array}{cccc} (1,1), & (1,2), & (1,3), & \ldots \\ (2,1), & (2,2), & (3,3), & \ldots \\ (3,1), & (3,2), & (3,3), & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{array} \right\}, \tag{1.30}$$

where I have ordered the elements in a quadratic scheme to make it more intuitive. Of course, this does not affect the set itself. So, the big question is now whether this set is still the same size as the natural numbers. The answer is yes, because you can still enumerate the set $\mathcal{T}_\infty$. You do this in the following way

$\mathcal{T}_\infty$	$\mathbb{N}$
(1,1)	1
(2,1)	2
(1,2)	3
(3,1)	4
(2,2)	5
(1,3)	6
(4,1)	7
(3,2)	8
(2,3)	9
(4,1)	10
$\vdots$	$\vdots$

ie you first enumerate the elements (k,l) where $k+l=2$, then you enumerate the elements where $k+l=3$ and so on. This works because for $k+l=r$ there are $r-1$ elements, ie a finite number. More precisely we make the association

$$(k,l) \leftrightarrow \frac{1}{2}(k+l-2)(k+l-1)+l \tag{1.31}$$

Exercise: Prove that the analytical form of this correspondence and convince yourself that this is indeed a one-to-one correspondence.

As a consequence we have

$$\aleph_0 \cdot \aleph_0 = \aleph_0 \tag{1.32}$$

Now the set $\mathcal{T}_\infty$ that I introduced here looks like a big array of elements (k,l) and this is surely one way of looking at it. But there is another way. Let me make the correspondence

$$(k,l) \leftrightarrow \frac{k}{l} \tag{1.33}$$

where the righthand-side is a rational number. So we see that the set

$$\mathcal{Q}_+ = \left\{ \begin{array}{cccc} \frac{1}{1}, & \frac{1}{2}, & \frac{1}{3}, & \cdots \\ \frac{2}{1}, & \frac{2}{2}, & \frac{2}{3}, & \cdots \\ \frac{3}{1}, & \frac{3}{2}, & \frac{3}{3}, & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{array} \right\} \tag{1.34}$$

and now you see that this is also the set of the positive rational numbers. Note that you may say that this is not exactly the set of rational numbers because for example we have $\frac{4}{2} = \frac{2}{1}$ and as a consequence there may be a case for identifying the two fractions $\frac{4}{2}$ and $\frac{2}{1}$. This would define a set that is slightly different, namely the set of all fractions where numerator and denominator have no common factor. This is of course possible, but it does not tell us something new because clearly that set is smaller than $\mathcal{Q}_+$ for which we have already verified that it has the same size as $\mathbb{N}$. As both sets are infinite then they are both countable.

Now you could start to ask the question whether there is a different degree of inifnity, ie is there a number that is larger than $\aleph_0$? This question is equivalent to the questions whether there is a set that cannot be enumerated. This is not a straightforward question at all, and in fact 150 years ago it was believed that all infinities are equal. It took a brilliant mathematician, Georg Cantor, to realize that infinities are really more complicated than one would have thought at first. Indeed, there are sets that are truly bigger than the natural numbers and we will get to know them soon.

1.3 Basic logic notation

Many mathematical expression will be written using logical symbols so it is useful to see early on. They have the advantage that they abbreviate statements and (after you got used to them) make them easier to comprehend than statements that are written in words. Firstly, consider the following

- $\forall x$ means ' for all x'
- $\exists x$ means 'there exists an x'
- s.t means 'such that'

So for example the statement

'For all $\epsilon > 0$ there is an n_0 such that for all $n > n_0$ we have $|a_n - a| > \epsilon$

then reads

$$\forall \epsilon > 0 : \exists\, n_0\, s.t. \forall n > n_0 : |a_n - a| > \epsilon.$$

There is also the logical negation $\neg$ which takes a logical statement and negates it, so if statement x is true then $\neg x$ is false and vice versa. Some examples will clarify this notation:

- $\neg(x > 0) = x \leq 0$
- $\neg(\forall x : f(x) > 0) = \exists x : f(x) \leq 0$
- $\neg(\exists x : f(x) < 0) = \forall x : f(x) \geq 0$
- $\neg(\forall \epsilon > 0 : \exists n_0\, s.t. \forall n > n_0 : |a_n - a| > \epsilon) = \exists \epsilon > 0 s.t. \forall n_0 : \exists n > n_0 s.t |a_n - a| < \epsilon$

Further logical operations are

- $x \wedge y$ means 'x and y'
- $x \vee y$ means 'x or y'

1.4 Sequences, completeness and uncountable sets

In the previous section I have introduced a first idea of the concept of functions and I have given a specific example, namely that of a function from the natural number $\mathbb{N}$ into the rational numbers via the association $n \mapsto a_n$. Functions that map the natural numbers into another set of quantities, such as the rational numbers, are also called *sequences*. In fact, the concept of a sequence is older than that of functions and has a life and significance of its own. In the following I would like to study sequences for two reasons. Firstly, because they show up in physics all the time. In fact, usually your theory, be it classical mechanics, thermodynamics or quantum mechanics, will predict a value for a measurable quantity to, at least in principle, arbitrary precision. In your experiment however you will always have some measurement uncertainty, ie you will get some value together with some error bars around it. For example your experiment may tell you that the value is

1.45 ± 0.05. So you make the best guess that the true value is $m_1 = 1.45$. Then you improve your apparatus and you are able to measure more precisely and you find 1.415 ± 0.05. Now you write down $m_2 = 1.415$. This continues with every improvement of your apparatus and you get an arbitrarily long sequence of numbers $m_1, m_2, m_3, \ldots$ which ideally should converge to some specific value m. These intuitive ideas are made precise in this section. Apart from their practical motivation sequences also have a more abstract importance. In a natural way they force upon us the idea of real numbers, and can indeed be used for their definition.

Note that for the moment, when I am speaking about 'a number' I mean to say 'a rational number'! Remember that so far we do not know any other type of numbers. Nevertheless, I will formulate give the definitions, lemmas and propositions as far as possible without the mentioning the restriction to rational numbers (unless of course they only apply to rational numbers).

Definition 6 *A sequence, denoted by* $\{a_n\}_{n=1,\ldots}$*, associates with every natural number* n *another number* a_n*.*

Question: Would you gain anything new if you would pick the indices n in the sequence from the rational numbers rather than the natural numbers?

The definition implies automatically that a sequence contains infinitely many elements. Nevertheless also *finite sequences* of numbers fit into this definition when we assume that all the undefined elements actually take the value zero. Again, finite sequences are rather easy to deal with, but also rather boring. *Infinite sequences*, or simply *sequences* as I will call them from now on, however, have infinitely many non-zero elements and have non-trivial properties.

A particularly important place is taken by sequences whose values become more and more limited the larger the index n becomes. An example is

$$\{a_n = \frac{1}{n}\}_{n=1,2,\ldots} \tag{1.35}$$

The larger n becomes the smaller the value of a_n, ie a_n approaches 0 arbitrarily closely. Such a sequence we would call convergent to 0. Note

however that, although the elements a_n come closer and closer to zero they are never equal to zero in this case. So zero is not an element of the sequence but it is arbitrarily well approximated by the elements of the sequence. In fact, this property that one can use sequences of numbers to approximate some other numbers is one of the most important uses of sequences. Many problems in physics have solutions that cannot be expressed in neat and tidy analytical expressions, but one can only provide more and more precise approximations to them. These ideas are captured in the following formal definition

Definition 7 *A sequence of numbers, denoted by* $\{a_n\}_{n=1,...}$*, is called* convergent *when there is a number* a *such that for every* $\epsilon > 0$ *there is an* n_0 *such that for all* $n > n_0$ *we have* $|a_n - a| \leq \epsilon$*. The number* a *is also often written as* $a = \lim_{n\to\infty} a_n$*. A sequence that does not converge is called a divergent sequence.*

Examples (expand some of them):
(a) $a_n = \frac{1}{n}$ is convergent to 0
(b) $a_n = \frac{n}{n+1}$ is convergent to 1
(c) and loads more examples for them to practice
(d) What do you say to $a_n = n$? Is it convergent or not? What could be the limit of this sequence?

Lemma 8 *The convergence properties of a sequence remain unaffected when we change a finite number of its elements.*

Proof: Given the sequence $\{a_n\}_{n=1,...}$ and another sequence $\{a'_n\}_{n=1,...}$ that differs from it only in finitely many places. This implies that there is an $\tilde{n}$ such that for all $n > \tilde{n}$ we have that $a_n = a'_n$. Whenever we have a statement

'$\forall \epsilon > 0$ there is an n_0 such that for all $n > n_0$ we have'

where $n_0 < \tilde{n}$ we can replace it by

'$\forall \epsilon > 0$ there is an n_0 such that for all $n > \tilde{n}$'

This confirms that convergence is unaffected by changing finitely many elements. This concludes the proof.

Let us define the convergence of sequences in a slightly different way. This concept is called Cauchy-sequence.

Definition 9 *A sequence of numbers, denoted by* $\{a_n\}_{n=1,\ldots}$*, is called a* Cauchy sequence *when for every* $\epsilon > 0$ *there is an* n_0 *such that for all* $m, n > n_0$ *we have* $|a_n - a_m| < \epsilon$*.*

This looks very close to the concept of convergence, so now let us prove the following

Theorem 10 *A convergent sequence of numbers is a Cauchy sequence.*

Proof: As the sequence $\{a_n\}_{n=1,\ldots}$ is convergent, then there is an a such that for every $\frac{\epsilon}{2} > 0$ there is an n_0 such that for all $n > n_0$ we have $|a_n - a| < \frac{\epsilon}{2}$. Then we have for all $m, n > n_0$ that

$$|a_n - a_m| = |a_n - a + a - a_m| \leq |a_n - a| + |a - a_m| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \quad (1.36)$$

This is just the definition of a Cauchy sequence and therefore the proof is finished.

Now you would also expect, that any Cauchy-sequence is convergent. This is a very reasonable conjecture, but nevertheless I would like to investigate it a little bit because I will show you that it is almost correct, but not quite. Luckily it fails for an interesting reason. Indeed, the failure of this idea will lead us to the concept of real numbers and the idea of completeness. Let me consider the following sequence which is defined as: Let x_n be the largest rational number such that $10^n x_n$ is a natural number and $x_n^2 \leq 2$. The first few elements of the sequence are

$$x_0 = 1, x_1 = 1.4, x_2 = 1.41, x_3 = 1.414, x_4 = 1.4142, \ldots \quad (1.37)$$

Exercise: Check that these first for elements are correct. We could have defined the sequence slightly differently, namely: Be x_n the largest rational number with at most n digits after the decimal point such that $x_n^2 \leq 2$. Now its clear that the sequence increases monotonically, ie $x_{n+1} \geq x_n$ for all n. This sequence is also a Cauchy sequence because for all $n > m \geq n_0$ we have

$$|x_n - x_m| \leq 10^{-m} \quad (1.38)$$

Therefore, for every $\epsilon > 0$ we chose n_0 to be the smallest integer that is larger than $\log_{10}(1/\epsilon) = -\log_{10}\epsilon$. Then for all $m, n \geq n_0$ we have

$$|x_n - x_m| \leq 10^{-m} \leq \epsilon \tag{1.39}$$

so that the sequence x_n is indeed a Cauchy sequence. Clearly, by the construction of the sequence it approximates closer and closer the number $\sqrt{2}$. So, you would say that this sequence converges to $\sqrt{2}$. But stop! So far we only know rational number. So, we better ask ourselves whether $\sqrt{2}$ is actually a rational numbers. Lets us check by trying

$$\frac{k}{l} = \sqrt{2} \tag{1.40}$$

where we have assumed that k and l have no common factor. If they would have, then we could simply cancel this factor. Ok, now square both sides to find

$$\frac{k^2}{l^2} = 2 \tag{1.41}$$

Now multiply by l^2 and we have

$$k^2 = 2l^2 \tag{1.42}$$

So, clearly the right hand side can be divided by 2. So the left hand side must be divisible by 2, ie even, as well. This implies that k is even, simply because the product of two odd number is odd again. So, $k = 2q$ and we find

$$4q^2 = 2l^2 \Rightarrow 2q^2 = l^2 \tag{1.43}$$

and now we conclude that l must be even as well. But then both k and l are even and therefore they have a common factor in contradiction to the original assumption that they do not have a common factor. Therefore $\sqrt{2}$ is not rational!

As a consequence, while you would like to say that our sequence x_n converges to $\sqrt{2}$ as its limiting element, we now realize that $\sqrt{2}$ is not a (rational) number. The phenomenon that we have encountered here is that to be able to say that every Cauchy sequence converges we need another requirement, namely that the limiting element is an admissible number. We now realize that the rational numbers alone

are not sufficient for that! The fact that there are Cauchy sequences of rational numbers that do not converge against a rational number is called the incompleteness of the rational numbers. Indeed, this tells us that there is something else than simply rational numbers. We have to extend the set of rational numbers by the limiting elements of all Cauchy-sequences of rational numbers. The resulting set is what we call the real numbers. Does this coincide with our everyday idea of real numbers as having a potentially infinite sequence of decimal digits? Yes, indeed. Because every such number can be approximated arbitrarily well by a sequence where x_n is the number that coincides with our real number in the first n decimal digits. Then, the larger n becomes, the closer x_n comes to the real number x.

Definition 11 *We define the set of real numbers as the set of limiting points of Cauchy sequences of rational numbers.*

As a consequence we have the following

Lemma 12 *Every Cauchy sequence of real numbers converges to a real number.*

Real numbers that are not rational are called irrational and a simple example that I have shown to you is $\sqrt{2}$. The above definition shows that sequences of rational numbers are sufficient to approximate every real number a fact that can also be formulated in the form: The rational number lie dense in the real numbers.

Definition 13 *A set of numbers $\mathcal{A}$ lies dense in another set $\mathcal{B}$ when for every point $x \in \mathcal{B}$ and every $\epsilon > 0$ there is a $y \in \mathcal{A}$ such that $|x - y| \leq \epsilon$.*

Now we have seen the mathematical significance of the real numbers. They are there to make sure that every Cauchy sequence converges. But now you may ask whether we can confirm experimentally that some physical quantity takes the value of a real number. Well, the answer is that we simply cannot, because every experiment will have a finite precision. So, as arbitrarily close to an irrational number there is a rational number we cannot really decide whether the physical quantity is irrational or rational. The only thing that we can do when

we increase our experimental precision is to rule out more and more rational numbers.

Before we discuss more practically useful properties and rules for sequences, let me briefly come back to the concept of infinity. The simple question is: How many real numbers are there? Are the irrational numbers a rather rare occurrence or are they typical or even the majority. To answer questions of this type we have to count the irrational numbers as well as the real numbers. I want to show you that one cannot count real numbers, ie one cannot put them into one-to-one correspondence with natural numbers. The proof of this statement will be by contradiction. Let me assume that I can count the real numbers. That would imply that I can make a list which may look like this

$$\begin{aligned} 1 &\leftrightarrow x_1 = 0.123452524752572\ldots \\ 2 &\leftrightarrow x_2 = 0.245209542095424\ldots \\ 3 &\leftrightarrow x_3 = 0.765987652752756\ldots \\ 4 &\leftrightarrow x_4 = 0.298752845295874\ldots \\ \vdots &\leftrightarrow \vdots \end{aligned}$$

Now I construct a real number that is definitively not on the list. This number is $x = 0.8542\ldots$ and is constructed by taking the first decimal digit of x_1 and subtracting it from 9, the second digit of x_2 and subtracting it from 9 and so on. Clearly this number is different to any other number on the list as the n-th digit of x_n is different from the n-th digit of x. Therefore we have constructed a contradiction to the assumption that we can make a list of all the real numbers and the size of the set of real numbers, which we denote from now on by $\aleph_1$ is truly bigger than that of the natural numbers, ie we have $\aleph_1 > \aleph_0$.

Therefore the size of the set of real numbers is truly larger than that of the natural numbers. In fact, one can show that in some well-defined sense the size of the set of real numbers is $2^{\aleph_0}$ which equals $\aleph_1$. For more details please consult the book by PR Halmos. You could of course wonder whether there is a degree of infinity which lies in between the two number $\aleph_0$ and $\aleph_1 = 2^{\aleph_0}$. This question is called the continuum hypothesis. The answer to this question is extremely tricky and rather surprising. Cantor, who introduced set theory and

studied the concept of infinite tried to decide this question for many years but he was unsuccessful. It took about 80 years before the full solution to this question was found by results of Kurt Gödel and Paul Cohen. The proven answer is: From the standard axioms of set theory the question cannot be decided in principle. In other words: Both the assumption that there is no number between $\aleph_0$ and $\aleph_1 = 2^{\aleph_0}$ and the assumption that there is a degree of infinity between $\aleph_0$ and $\aleph_1 = 2^{\aleph_0}$ is fully compatible with the other axioms of set theory. This was not the answer that people had expected but it is not entirely unusual in mathematics. A similar status is held by the parallel axiom in Euclidean geometry. mathematicians tried to prove it in vain for nearly 2000 years from the other axioms of Euclidean geometry. It was only in the 18^{th} century that it was realized that this axiom is simply independent of the the others and therefore cannot be proven either. In fact, there are non-Euclidean geometries which take an important place in physics as they form the basis of Einsteins theory of general relativity. So they are real and not just mathematicians fun. However, some mathematicians are not happy with this situation and there are still thoughts spent on how to find another intuitive axiom for set theory that would allow to decide the question. After all, there does not seem to be any physical system on which we can simply make an experiment to see whether there is a different degree of infinity while this was the case for geometry. But what would happen if we find such an extra axiom that allows us to decide this question. Will it mean that every other theorem can be proven or disproven with this stronger set of axioms? That would be nice, but unfortunately it is known that as long as there are no contradictions in the set of axioms you will always be able to find further statements that are independent of those axioms, ie they can neither be proven or disproven within these axioms. Does this have any bearing on physics? Well, you know that there are always people that tell you that the final theory of all physics is around the corner (Hawking is an example). But now you have to start to worry, because if in any system of axioms for mathematics there are statements that can neither be proven or disproven, how can there be a final theory of physics? After all physics is based on mathematics and as a consequence for every theory that we formulate there should be statements that we can neither prove nor disprove mathematically.

Therefore we would have to test them experimentally to see which one is true and then extend our theory to cover for it. This appears to be a process which goes ad infinitum. So, it seems that we physicists will never be out of a job. I hope that these things make you think a little bit. Clearly if you want to make scientific statements about this question you will have to be much more rigorous than I was here. In any case, so far I haven't seen that any of the 'last theory of physics' guys has actually presented a satisfactory answer to this.

For the classwork: 1) Show that the Cantor dust is uncountable but not dense in the set of real numbers. (3-yadic expansion without 1's)

2) The set of all infinite sequences made up of digits 0,1, ..., 9 is uncountable as it is the same as the set of real numbers.

From now on, let us move on to practical aspects of sequences. Now that we officially know the real numbers, when I say number then I mean real number unless otherwise stated.

I want to give you a few basic properties of sequences that are useful to decide whether a sequence is converging or not. Obviously, there are many sequences that are very weird and wild. The digits of π for example show absolutely no pattern whatsoever, and in fact pass every random number test that has been used on them. But of course they are not random, as each one of them can be determined analytically. But the digits of π have one property. They have values $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ and therefore they are bounded. This gives rise to the idea of a bounded sequence

Definition 14 *A sequence of numbers, denoted by* $\{a_n\}_{n=1,...}$*, is called a bounded from above (below) if there is a number* r_{above} *(*r_{below}*) such that for all* n *we have* $a_n < r_{above}$ *(*$a_n > r_{below}$*). Sequence is called bounded if it is bounded both from below and above.*

Exercises: (a) If $\{a_n\}_{n=1,...}$ is bounded from below then $\{-a_n\}_{n=1,...}$ is bounded from above.
(b) If $\{a_n\}_{n=1,...}$ is bounded then $\{a_n^2\}_{n=1,...}$ is also bounded.
(c) Given a sequence $\{a_n\}_{n=1,...}$, when is the sequence $\{1/a_n\}_{n=1,...}$ bounded.

Why is the concept of boundedness useful? The reason is that it is quite closely connected to convergence. Firstly we have

Lemma 15 *Every convergent sequence is bounded.*

Proof: Convergent means that there is an a such that for example for $\epsilon = 1$ there is an n_0 such that for all $n > n_0$ we have $|a_n - a| \leq 1$. This means that for $n > n_0$ $a + 1 \geq a_n \geq a - 1$. Taking into account the remaining elements of the sequence a_k for $k = 1, \ldots, n_0$ we find that

$$\max\{a_1, \ldots, a_{n_0}, a + 1\} \geq a_n \geq \min\{a_1, \ldots, a_{n_0}, a - 1\}. \qquad (1.44)$$

Conversely however, not every bounded sequence is convergent. A simple example is the sequence with elements $a_n = (-1)^n$ which is clearly bounded, but does not converge. But from this sequence we see that if we pick every second element and form the new sequence $b_n = a_{2n}$ then we have a converging (in fact constant) sequence. Is this a general phenomenon that from a bounded sequence I can take a subsequence that is convergent? The answer is yes. Let us first define properly what a subsequence is

Definition 16 *Be $(a_n)_{n\in\mathbb{N}}$ a sequence and*

$$n_0 < n_1 < n_2 < \ldots$$

a strictly increasing sequence of natural numbers, then the sequence $(b_n)_{n\in\mathbb{N}}$ with $b_i = a_{n_i}$ is called a subsequence of $(a_n)_{n\in\mathbb{N}}$.

Examples: (i) Consider the sequence defined by $a_n = n$ of natural numbers. A subsequence would be $b_n = a_{2n} = 2n$, ie the sequence of even numbers.

Now we can formulate

Theorem 17 *A bounded sequence always possesses a convergent subsequence.*

Proof: As the sequence f_n is bounded it lies in an interval $[-a, a]$ for some value a. Now we will define two sequences $\{a_i\}$ and $\{b_i\}$. We start both with value $a_1 = -a$ and $b_1 = a$, ie the lower and upper boundary

of our interval. Now we split the interval $[a_1, b_1]$ in two equal halves. As the sequence has infinitely many elements, at least one of the two halves of the original interval must contain infinitely many elements. If this is the interval $[a_1, \frac{a_1+b_1}{2}]$ then we define $a_2 = a_1$ and $b_2 = \frac{a_1+b_1}{2}$. If it is the interval $[\frac{a_1+b_1}{2}, b_1]$ then we define $a_2 = \frac{a_1+b_1}{2}$ and $b_2 = b_1$. We then cut the new interval $[a_2, b_2]$ into two halves and so on. Continue this procedure to construct the sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$. Each of the intervals $[a_k, b_k]$ contains infinitely many points and we chose from each of them a single value c_k which equals an element of the sequence $(f_n)_{n\in\mathbb{N}}$. The resulting sequence $(c_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. This is so because for all $n > n_0$ we have that $c_n \in [a_{n_0}, b_{n_0}]$ and therefore for $n, m > n_0$ we have $|c_n - c_m| \leq b_{n_0} - a_{n_0} = 2a2^{-n_0}$. So clearly for any $\epsilon > 0$ I can find an n_0 such that one has for all $n, m > n_0$ that $|c_n - c_m| < \epsilon$. Therefore the sequence $(c_n)_{n\in\mathbb{N}}$ it is convergent. This concludes the proof.

The limit of a subsequence $(b_n)_{n\in\mathbb{N}}$ of $(a_n)_{n\in\mathbb{N}}$ is called an *accumulation point* of the sequence $(a_n)_{n\in\mathbb{N}}$.

Now something really useful if you want to avoid fiddling with $\epsilon's$ in the proof of the convergence of a sequence.

Lemma 18 *A monotonically increasing (decreasing) sequence $(a_n)_{n\in\mathbb{N}}$ that is bounded from above (below) converges.*

Proof: I just deal with the increasing sequence which is automatically bounded from below and by assumption also from above. Therefore the sequence is bounded and possesses a subsequence $(b_n)_{n\in\mathbb{N}}$ that converges to a value a. The elements of this subsequence are given by $b_k = a_{n_k}$ with an increasing sequence n_k. Now we show that also the sequence as a whole converges against a. Because the subsequence $(b_n)_{n\in\mathbb{N}}$ converges we have that for every ϵ there is a k_0 such that for all $k > k_0$ we have $|b_k - a| \leq \epsilon$. As the sequence $(a_n)_{n\in\mathbb{N}}$ is increasing we then have that for all $n > n_k$ we have $|a_n - a| \leq |b_k - a| \leq \epsilon$. This concludes the proof.

To see how much easier life has become with this theorem let us consider some of the previous examples.

Examples:

(a) $a_n = \frac{1}{n}$ is convergent because it is bounded from below and is monotonically decreasing

(b) $a_n = \frac{n}{n+1}$ is convergent because it is monotonically increasing and bounded from above by 1

(c) What do you say to $a_n = n$ does not converge. Although it is monotonically increasing it is not bounded from above and as a consequence it cannot converge. Remember: A converging sequence is bounded.

I am always going on about bounded sequences. Does this mean that unbounded sequences are useless? Well, not really. In fact, there are quite a few natural phenomena that lead to unbounded sequences.

Example: In the middle ages mathematics in Europe was not in a very good state. However, here and there were some smart people who nevertheless did work on mathematics. One of them was Leonardo da Pisa who is widely known as Fibonacci (which means blockhead and given that his head had a perfectly natural shape probably refers to his character). He traveled the far east and brought a lot of mathematical ideas back to Europe and he also studied mathematical problems himself. One particular problem he considered carries his name and pops up in surprisingly many aspects of mathematics, physics and even biology. In 1202 Fibonacci considered the following problem: Assume we have pairs of Rabbits. After they were born, they will take one season to grow up, and from then on, every season they will produce a new pair of rabbits. If we neglect that they are dying, how many rabbits will there be after n seasons? We start with one pair born in the first season, ie $F_1 = 1$. They have to grow up, so no children yet, ie. $F_2 = 1$. In the next season they have kids, so there is one more pair, ie $F_3 = 2$. The new pair won't have children yet but the old one continues to reproduce, ie. $F_4 = 3$. In fact, you can now see that the sequence is built in general as

$$F_{n+2} = F_{n+1} + F_n \tag{1.45}$$

ie the population of generation $n + 1$ increases by the population of generation n as they start to have children now. Quite clearly this

sequence is unbounded but nevertheless it is interesting to know how it develops. To get an idea, let us consider the slightly simpler case $G_{n+1} = G_n + G_n = 2G_n$. Clearly $G_n = 2^n G_0$ so the sequence grows exponentially. So let us make an Ansatz in the Fibonacci sequence, namely try $F_n = q^n$ and insert into the definition to find the quadratic equation

$$q^2 = q + 1 \tag{1.46}$$

which has solution

$$q_1 = \frac{1}{2}\left(1 + \sqrt{5}\right) \text{ and } q_2 = \frac{1}{2}\left(1 - \sqrt{5}\right). \tag{1.47}$$

So the general solution of the Fibonacci sequence is then a linear combination of the basic solutions, ie

$$F_n = \alpha_1 q_1^n + \alpha_2 q_2^n, \tag{1.48}$$

and we determine the constants α_1 and α_2 from the first few elements of the Fibonacci sequence such as $F_1 = 1$ and $F_2 = 1$ to find

$$\alpha_1 = -\alpha_2 = \frac{1}{\sqrt{5}}. \tag{1.49}$$

Inserting all this we find

$$F_n = \frac{q_1^n - (-q_1)^{-n}}{\sqrt{5}}. \tag{1.50}$$

Check this for a few terms to convince yourself that it is indeed correct.

It is remarkable how often Fibonacci numbers are encountered in nature. For example they govern the numbers of leaves, petals and seed grains of many plants.

Now an example for an important special generalization of sequences of Fibonacci numbers

Example: You all know fractions such as $\frac{2}{3}$ which are really a very simple concept. However, there is a slightly trickier concept that is closely related to both sequences a fractions. This is the idea of continued fractions which look like

$$b_0 + \frac{1}{b_1 + \frac{1}{b_2 + \frac{1}{b_3 + \ldots}}} \tag{1.51}$$

and are usually abbreviated by $[b_0; b_1, b_2, b_3, \ldots]$. Usually, such a fraction has infinitely many terms and you calculate it by building the sequence of partial fractions that break off after the n-th term, ie

$$\frac{A_k}{B_k} = [b_0; b_1, \ldots, b_k] \tag{1.52}$$

where the sequences of numerators obey

$$A_k = b_k A_{k-1} + A_{k-2} \tag{1.53}$$

with $A_0 = b_0, A_{-1} = 1$ and $A_{-2} = 0$ and the sequence of denominators obeys

$$B_k = b_k B_{k-1} + B_{k-2} \tag{1.54}$$

with $B_0 = 1$ and $B_{-1} = 0$. (Comment: That these are the sequences requires proof, and this will be done either in the lecture or the classworks.)

A first application of continued fractions is that they allow to find good approximations to given fractions. For example

$$\frac{964}{437} = [2; 4, 1, 5, 1, 12] \tag{1.55}$$

If I now want to approximate this fraction by one with a smaller denominator then I can for example break off the continued fraction after the second term. This will give $[2; 4, 1] = \frac{11}{5}$ which is correct to within 3 parts in 10^3. Not bad, isn't it?

Research project:[8] Try to find a method by by which one can obtain the continued fraction expansion of any rational number.

Another example is the following slight variation of continued fractions

$$\tan z = \left[\frac{z}{1-}\frac{z^2}{3-}\frac{z^2}{5-}\ldots\right] := \frac{z}{1 - \frac{z^2}{3-\frac{z^2}{5-\ldots}}}. \tag{1.56}$$

[8]I call this research project, because you really need to explore the problem or search the literature. And of course I will not give you the answer for quite a while. Thats just like real research where nobody knows the answer until you find it.

Now take the second order approximation, we find

$$\tan z \cong z\frac{15-z^2}{15-6z^2} \tag{1.57}$$

which is really surprisingly good. In fact, for $z = \frac{\pi}{4}$ this is about 0.9998 instead of the exact value 1. In stark contrast to that, the power series expansion with three terms $\tan z = z + \frac{z^3}{3} + \frac{z^5}{5}$ makes an error that is 32 times larger. The advantage of the continued fraction is that it uses power series both in numerator and denominator and can therefore reproduce functional behaviour much better.

As a curiosity let me tell you that the continued fraction $[1; 1, 1, \ldots]$, which is about the simplest continued fraction you can think of, has a particular importance in many areas including art. Its value is the "golden ratio" $g = \frac{1+\sqrt{5}}{2}$. In art it is recognized that for example rectangles where the long and the short side have lengths that have a ratio equal to g, are perceived as particularly nice and well balanced. Indeed, in paintings of great master you can sometimes realize that they have used this ratio (probably subconsciously). For an example see the delightful book "Number theory in science and communcation" by MR Schroeder published by Springer.

Classwork: Given an electrical network made up of resistances in the way shown in the figure.

Figure 1.4: This is a caption.

Now let us state some of the laws by which to manipulate convergent

sequences

Theorem 19 *Given the convergent sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ with limits a and b respectively as well as real numbers λ and μ.*

1. $\lim_{n\to\infty}(a_n + b_n) = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n$
2. $\lim_{n\to\infty}(a_n b_n) = \lim_{n\to\infty} a_n \lim_{n\to\infty} b_n$
3. $\lim_{n\to\infty}(a_n/b_n) = \lim_{n\to\infty} a_n / \lim_{n\to\infty} b_n$
4. $\lim_{n\to\infty} \lambda a_n = \lambda a$
5. *If $a_n \leq b_n$ for all n then $\lim_{n\to\infty} a_n \leq \lim_{n\to\infty} b_n$ but note that $a_n < b_n$ for all n does* **not** *imply $\lim_{n\to\infty} a_n < \lim_{n\to\infty} b_n$*

Proofs: Exercise!

These rules apply for converging sequences, but we can also make a few rules for sequences $(a_n)_{n\in\mathbb{N}}$ that diverge to infinity. For example $\lim_{n\to\infty} a_n = \infty \Rightarrow \lim_{n\to\infty} \lambda a_n = \infty$ for $\lambda \neq 0$ which represents the rule $\infty = \lambda\infty$. A tricky point is the case when $\lambda = 0$ ie what happens to expressions such as $0 \cdot \infty$? Can one make sense of them? Well, yes one can. Given another sequence $(b_n)_{n\in\mathbb{N}}$ which converges to 0 then we can consider what happens to $\lim_{n\to\infty} a_n b_n$. The answer is however not unique. It very much depends on the individual sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$. This can be seen with some simple examples:

- $a_n = n$ and $b_n = \frac{1}{n}$ then we have
$$\lim_{n\to\infty} a_n b_n = 1$$
- $a_n = n$ and $b_n = \frac{1}{n^2}$ then we have
$$\lim_{n\to\infty} a_n b_n = 0$$
- $a_n = n$ and $b_n = \frac{1}{\sqrt{n}}$ then we have
$$\lim_{n\to\infty} a_n b_n = \infty$$

So you can see that anything can happen and that great care is needed when one is dealing with sequences that do not converge. Soon we will see even more examples for the counterintuitive features of infinite sequences.

As I have said already, sequences are useful for approximating numbers such as square roots by rationals. Of course there are methods that converge rapidly and require a lot of effort to compute the successive elements of the sequence and others that work very efficiently. Previously I had constructed a sequence for you that converges against $\sqrt{2}$. But the construction was rather clumsy from a numerical point of view (it was however very useful as one could immediately see that it is a Cauchy sequence). Now I will show you a very efficient way of computing roots of numbers, ie we compute $a^{1/k}$ for any $k \in \mathbb{N}$ efficiently.

First I will show you how to compute $\sqrt{a}$ and then, as an exercise, you will show how to do this for $a^{1/k}$ with any $k \in \mathbb{N}$. Let us define the following sequence

$$x_0 > 0 \tag{1.58}$$

$$x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n}). \tag{1.59}$$

For $a = 2$ and $x_0 = 1$ the first few elements of this sequence have the form

$$x_0 = 1, x_1 = \frac{3}{2}, x_2 = \frac{17}{12}, x_3 = \frac{577}{408}, x_4 = \frac{665857}{470832}, \ldots \tag{1.60}$$

Just by checking the values you can see that they come very close to $\sqrt{2}$. Indeed,

$$\frac{x_0}{\sqrt{2}} \approx 0.70710678118654752440084436210485 \tag{1.61}$$

$$\frac{x_1}{\sqrt{2}} \approx 1.06066017177982128660126654315 73 \tag{1.62}$$

$$\frac{x_2}{\sqrt{2}} \approx 1.00173460668094232623452951298 19 \tag{1.63}$$

$$\frac{x_3}{\sqrt{2}} \approx 1.00000150182509294504727254150 61 \tag{1.64}$$

$$\frac{x_4}{\sqrt{2}} \approx 1.00000000000112773761123505712 88\,. \tag{1.65}$$

As you can see after only 4 steps the relative difference between x_4 and $\sqrt{2}$ is roughly 10^{-12} which is rather good. Even better, it seems that in every step we roughly double the number of correct digits. This looks like a rather useful scheme! So, I would like to prove to you that indeed the method does converge to $\sqrt{2}$ analytically and I will also show you that the number of valid digits is roughly doubled in every step of the iteration.

Proof: First we prove convergence. The proof proceeds in a number of small steps

1. For all k we have $x_k > 0$ which one can see by induction as $x_k > 0$ implies $x_{k+1} = \frac{1}{2}(x_k + \frac{a}{x_k})$. As $x_0 > 0$ all subsequent $x_k > 0$.

2. For all $k \geq 1$ we have $x_k^2 > a$. To see this consider

$$\begin{aligned} x_k^2 - a &= \frac{1}{4}\left(x_{k-1} + \frac{a}{x_{k-1}}\right)^2 - a \\ &= \frac{1}{4}\left(x_{k-1}^2 + 2a + \frac{a^2}{x_{k-1}^2}\right) - a \\ &= \frac{1}{4}\left(x_{k-1}^2 - 2a + \frac{a^2}{x_{k-1}^2}\right) \\ &= \frac{1}{4}\left(x_{k-1} - \frac{a}{x_{k-1}}\right)^2 \\ &> 0 \end{aligned}$$

3. For all $k \geq 1$ the sequence is monotonically decreasing, ie $x_{k+1} \leq x_k$.

$$x_k - x_{k+1} = x_k - \frac{1}{2}(x_k + \frac{a}{x_k}) = \frac{1}{2}(x_k - \frac{a}{x_k}) = \frac{x_k^2 - a}{2x_k} \geq 0.$$

4. Apart possibly of the first element, the sequence is monotonically decreasing and bounded from below. Therefore is converges to a value x.

5. Now we find an equation that x must satisfy.

$$\begin{aligned} x &= \lim_{n\to\infty} x_n \\ &= \lim_{n\to\infty} \frac{1}{2}\left(x_{n-1} + \frac{a}{x_{n-1}}\right) \\ &= \frac{1}{2}\lim_{n\to\infty} x_{n-1} + \frac{a}{2}\lim_{n\to\infty}\frac{1}{x_{n-1}} \\ &= \frac{1}{2}\left(x + \frac{a}{x}\right). \end{aligned}$$

The unique positive solution of this equation for x is indeed $x = \sqrt{a}$.

This concludes the proof.

A hard exercise for the daring: Prove that for any $k > 1$ the sequence

$$x_0 > 0 \tag{1.66}$$

$$x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n^{k-1}}) \tag{1.67}$$

converges to

$$\lim_{n\to\infty} x_n = a^{\frac{1}{k}} \tag{1.68}$$

Hints: The tricky point here is that the sequence is alternating and not simply monotonically increasing. You need to show that if $x_1 < a^{1/k}$ the subsequences x_{2k} is monotonically decreasing while the subsequence x_{2k+1} is monotonically increasing. Then you show that they are bounded so that each of the converges, but as also the whole sequence converges you have total convergence. For $k = 3$ it helps to realize that for $x < a^{1/3}$ one has $(x - a^{1/3})^3 + a^{1/3}(x - a^{1/3})(x - 2a^{1/3}) < 0$

Classwork: Investigate the behaviour of the sequence

$$x_0 < 0 \tag{1.69}$$

$$x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n}) \tag{1.70}$$

for $a > 0$ and show that it converges to $\lim_{n\to\infty} x_n = -\sqrt{a}$.

Exercise: Investigate the behaviour of the sequence

$$x_0 > 0 \tag{1.71}$$

$$x_{n+1} = \frac{1}{2}(x_n - \frac{a}{x_n}) \tag{1.72}$$

for $a > 0$ What do you observe? We will come back to this sequence when we study complex numbers again.

1.5 Series

An important special case of sequences is that of infinite sums which are also called series. Often in physics you will be confronted with systems which we approximate as infinite. Examples are condensed matter physics, thermodynamics and others. If one then wishes to determine quantities such as the energy of such a system one will then have to add infinitely many terms. As usual one will do this by adding finitely many and then taking the limit towards larger and larger number of terms in the sum. In that way one arrives at a limit which is then the value of the infinite sum. Let us make this definition precise

Definition 20 *Given a sequence* $(a_n)_{n=1,\ldots}$. *The sequence of partial sums*

$$s_n = \sum_{k=1}^{n} a_k \tag{1.73}$$

is called series *and denoted by* $\sum_{k=1}^{\infty} a_k$. *In the case of convergence the limit will also be denoted by* $s = \sum_{k=1}^{\infty} a_k$.

As a series is a special case of a sequence the definitions for convergence, Cauchy sequence and boundedness carry over in a trivial fashion. For example

Definition 21 *A series* $\sum_{k=1}^{\infty} a_k$ *is a Cauchy series when* $\forall \epsilon > 0$: $\exists n_0\, s.t. \forall n > m$ *with* $m, n > n_0$ *we have*

$$|\sum_{i=m}^{n} a_i| \leq \epsilon. \tag{1.74}$$

Definition 22 *A series* $\sum_{i=1}^{\infty} a_i$ *is called convergent if* $\forall \epsilon > 0 : \exists n_0 \, s.t. \, \forall n > n_0$

$$|\sum_{i=n}^{\infty} a_i| \leq \epsilon. \tag{1.75}$$

Examples: Consider the sum

$$\sum_{i=1}^{\infty} \frac{1}{i}.$$

This sum diverges as can be seen by grouping the terms conveniently.

$$\begin{aligned} \sum_{i=1}^{\infty} \frac{1}{i} &= 1 + \left(\frac{1}{2}\right) + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) + \dots \\ &= \sum_{n=0}^{\infty} \left(\sum_{k=2^n+1}^{2^{n+1}} \frac{1}{k} \right). \end{aligned}$$

We see that every bracket has a value greater or equal to $1/2$ because

$$\sum_{k=2^n+1}^{2^{k+1}} \frac{1}{k} \leq \sum_{k=2^n+1}^{2^{k+1}} \frac{1}{2^{n+1}} = \frac{2^{n+1} - 2^n}{2^{n+1}} = \frac{1}{2} \tag{1.76}$$

and there are infinitely many of them so that the series diverges to infinity.

Now consider the series

$$\sum_{i=1}^{\infty} \frac{1}{n(n+1)} \, .$$

To determine the value of the sum we prove first by induction that

$$\sum_{i=1}^{k} \frac{1}{n(n+1)} = 1 - \frac{1}{k+1} \, .$$

This is certainly true for $k = 1$ as one easily check. Now assume its true for k. Then for $k + 1$ we find

$$\sum_{i=1}^{k+1} \frac{1}{n(n+1)} = \sum_{i=1}^{k} \frac{1}{n(n+1)} + \frac{1}{(k+1)(k+2)} = 1 - \frac{1}{k+2}$$

so that the statement also holds for $k + 1$. As a consequence its true for all values of k. Now we see that

$$\lim_{k\to\infty} \sum_{i=1}^{k} \frac{1}{n(n+1)} = \lim_{k\to\infty} (1 - \frac{1}{k+1}) = 1.$$

Finally let us briefly consider the geometric series for $q < 1$

$$\sum_{n=0}^{\infty} q^n = \frac{1}{1-q} \tag{1.77}$$

This can be seen by considering the partial sums

$$(1-q)\sum_{n=0}^{k} q^n = \sum_{n=0}^{k} q^n - \sum_{n=0}^{k} q^{n+1} = \sum_{n=0}^{k} q^n - \sum_{n=1}^{k+1} q^n = 1 - q^{k+1} \tag{1.78}$$

so that

$$\sum_{n=0}^{k} q^n = \frac{1-q^{k+1}}{1-q}. \tag{1.79}$$

Then clearly we have

$$\sum_{n=0}^{\infty} q^n = \lim_{k\to\infty} \sum_{n=0}^{k} q^n = \lim_{k\to\infty} \frac{1-q^{k+1}}{1-q} = \frac{1}{1-q}. \tag{1.80}$$

After these examples we would now like to have a look whether it is possible to find simple criteria to decide whether a series converges or not. Of course we know convergence criteria already, but they are rather hard to use, with their $\epsilon's$ and $\forall's$ and $\exists's$. So, let us formulate some useful lemma's.

Lemma 23 *If the series $\sum_{i=1}^{\infty} a_i$ converges, then the sequence $(a_n)_{n=1,\ldots}$ converges to zero.*

Proof: As the series converges it is a Cauchy sequence. Then for every $\epsilon > 0$ there is an n_0 such that for $n > m > n_0$ we have

$$|\sum_{i=m}^{n} a_i| \leq \epsilon. \tag{1.81}$$

In particular we then have for $m = n > n_0$ that

$$|a_n| = |\sum_{i=n}^{n} a_i| \leq \epsilon \tag{1.82}$$

which implies that the sequence $(a_n)_{n=1,...}$ converges to zero. This finishes the proof.

This lemma can be used in the following way. Check whether the sequence of $(a_n)_{n=1,...}$ converges. If it does not converge, then this implies that the corresponding series $\sum_{i=1}^{\infty} a_i$ does not converge either.

Lemma 24 *The convergence of the series $\sum_{i=1}^{\infty} |a_i|$ implies the convergence of the sequence $\sum_{i=1}^{\infty} a_i$.*

Proof: By the triangular inequality we have

$$\epsilon \geq \sum_{i=m}^{n} |a_i| \geq |\sum_{i=m}^{n} a_i|$$

so if $\sum_{i=1}^{\infty} |a_i|$ is a Cauchy series, so is $\sum_{i=1}^{\infty} a_i$ and both converge. This finishes the proof.

More generally we can say

Lemma 25 *(Comparison criterion) Let $\sum_{n=1}^{\infty} c_n$ be a convergent series where all the c_n are positive. Any other series $\sum_{n=1}^{\infty} a_n$ for which we have $|a_n| \leq c_n$ converges absolutely.*

Proof: Because $\sum_{n=1}^{\infty} c_n$ converges, it is a Cauchy sequence. Then for any $\epsilon > 0$ there is an n_0 such that for all $m, n > n_0$ we have

$$\sum_{k=m}^{n} |a_k| \leq \sum_{k=m}^{n} c_k = |\sum_{k=m}^{n} c_k| \leq \epsilon \tag{1.83}$$

Therefore also $\sum_{k=1}^{\infty} |a_k|$ is a Cauchy sequence. This completes the proof.

Thus we have a reasonably efficient way to identify some series that diverge. But we would really like to have some positive statements as well. Here is one of these

Theorem 26 *(Ratio test of D'Alembert) For a series $\sum_{i=1}^{\infty} a_i$ we form the sequence $q_i = |a_{i+1}/a_i|$ and determine its limit*

$$\lim_{n\to\infty} q_n = q.$$

The series $\sum_{i=1}^{\infty} a_i$ converges if $q < 1$, it diverges if $q > 1$ and for $q = 1$ it may either converge or diverge.

Proof: Study first the case $q < 1$. Then there is an $\epsilon > 0$ and an n_0 such that for all $n > n_0$ we have $q < 1-2\epsilon$ and $|q_n - q| < \epsilon$. Then by the triangular inequality we have $|q_n| = |q_n - q + q| \leq |q_n - q| + |q| \leq 1 - \epsilon$. As the convergence of the series is not influenced by changing a finite number of terms (though its value is changed) we can now assume that $|q_n| \leq 1-\epsilon$ for all n. This implies that $|a_{n+1}| \leq (1-\epsilon)|a_n| \leq (1-\epsilon)^n |a_0|$ and we have

$$\sum_{i=m}^{n} |a_i| \leq |a_0| \sum_{i=m}^{n} (1-\epsilon)^i = |a_0| \frac{(1-\epsilon)^m - (1-\epsilon)^n}{\epsilon}.$$

For sufficiently large m, n this can be made arbitrarily small and so we have a Cauchy series and by the previous lemma we have completed the proof of this case.

For the case $q > 1$ we have an $\epsilon > 0$ and an n_0 such that for all $n > n_0$ we have $q > 1 + 2\epsilon$ and $|q_n - q| < \epsilon$. As a consequence, again of the triangular inequality, we have $|q_n| = |q - (q - q_n)| \geq |q| - |q - q_n| > q - \epsilon > 1 + \epsilon$. Therefore we conclude that $|a_n| \geq (1+\epsilon)^n |a_0|$ so that the sequence of $|a_n|$ does not converge. Then, as a consequence by the above lemma also $\sum_{n=1}^{\infty} a_n$ cannot converge. This completes the proof.

For the case $q = 1$ I need to present one series that diverges and one the converges to convince you that in this case the quotient criterion is useless. Consider the series

$$\sum_{n=1}^{\infty} \frac{1}{n}$$

which diverges but has $q_i = \frac{i+1}{i}$ which converges to 1, and

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = 1$$

which converges but also has $q_i = \frac{(n+1)^2}{n^2}$ which converges to 1as well. This concludes the proof.

A further criterion is

Theorem 27 *(Root test of Cauchy) For a series $\sum_{i=1}^{\infty} a_i$ we form the sequence $q_i = |a_i|^{1/i}$ and determine its limit*

$$\lim_{n\to\infty} q_i = q$$

The series $\sum_{i=1}^{\infty} a_i$ converges if $q < 1$, it diverges if $q > 1$ and for $q = 1$ it may either converge or diverge.

Proof: Works analogously to the proof of the the previous theorem and is left as an exercise.

1.5.1 Absolute convergence

Up until now we have generally investigated the sequences of the form $\sum_{i=1}^{\infty} |a_i|$ and from their convergence we concluded that $\sum_{i=1}^{\infty} a_i$ must also converge. This is fine, but there are sequences that cannot be attacked like that. An example is the following sequence

$$\sum_{n=1}^{\infty} \frac{(-1)^{(n+1)}}{n} = \ln 2$$

which converges but for which the sequence $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges to infinity.

Exercise: Prove convergence (but not the value) of the above series by collecting neighboring terms in pairs to obtain a sum for which you can show convergence more easily. This is a trick to accelerate the convergence of the sum and later you will learn more of these.

So, we see that the series converges. Surely as we just grouped the terms in pairs, we may also reorder the terms as we please without affecting the convergence? This may seem natural, but it is definitively wrong. Before I will prove this to you I will first define what I mean by a reordering and then I will show that for every absolutely convergent series the order of the summation does not play any role.

Definition 28 *Given a sequence $(a_n)_{n=1,\dots}$ and a bijective map $\tau : \mathbb{N} \to \mathbb{N}$ then we call the sequence $(b_n)_{n=1,\dots}$ with $b_n = a_{\tau(n)}$ a reordering of the sequence $(a_n)_{n=1,\dots}$.*

Now we can state

Lemma 29 *Let $\sum_{i=1}^{\infty} a_i$ be an absolutely convergent series with the limit a, then every reordering is a convergent series with limit a.*

Proof: We need to show that for any reordering

$$\lim_{n\to\infty} \sum_{i=1}^{n} a_{\tau(i)} = a.$$

Due to the absolute convergence of the series, for every ϵ there is an n_0 such that for all $n > n_0$ we have

$$\sum_{i=n_0}^{\infty} |a_i| \leq \frac{\epsilon}{2}\,.$$

Now chose $N(n_0)$ so large that $\{1, 2, \dots, n_0\} \subset \{\tau(1), \tau(2), \dots, \tau(N)\}$. Then we find for all $n > N$

$$\begin{aligned} |\sum_{i=1}^{n} a_{\tau(i)} - a| &\leq |\sum_{i=1}^{n} a_{\tau(i)} - \sum_{i=1}^{n_0-1} a_i| + |\sum_{i=1}^{n_0-1} a_i - a| \\ &\leq \sum_{i=n_0}^{\infty} |a_i| + \frac{\epsilon}{2} \\ &\leq \epsilon. \end{aligned}$$

This finishes the proof.

Now we come to the proof that the series $\sum_{n=1}^{\infty} \frac{(-1)^{(n+1)}}{n}$, which is not absolutely convergent, has a reordering that diverges. In fact one can prove that in general for any convergent series that is not absolutely convergent and any given c there is a reordering such that the reordered series has the limiting value c. I will not show you the proof for this general statement but I will show you that there is a reordering of the sequence $\sum_{n=1}^{\infty} \frac{(-1)^{(n+1)}}{n}$ that diverges to infinity. So, remember that the elements of the series are given by $a_n = \frac{(-1)^n}{n}$. Let us have a look at the elements with odd index ranging from $\frac{1}{2^n+1}$ to $\frac{1}{2^{n+1}+1}$. Then for every $n \geq 1$ we have

$$\frac{1}{2^n+1} + \ldots + \frac{1}{2^{n+1}+1} \geq \frac{2^{n-1}}{2^{n+1}} = \frac{1}{4}.$$

Therefore we reorder the series in the following way

$$\begin{array}{rcl} 1 & - & \frac{1}{2} \\ & + & \frac{1}{3} - \frac{1}{4} \\ & + & \left(\frac{1}{5} + \frac{1}{7}\right) - \frac{1}{6} \\ & + & \left(\frac{1}{9} + \frac{1}{11} + \frac{1}{13} + \frac{1}{15}\right) - \frac{1}{8} \\ & \vdots & \vdots \end{array}$$

which clearly diverges as every row gives a contribution that is definitively larger than $\frac{1}{12}$. This completes the proof.

Research project: The brave amongst you may try to prove the following statement. Given a series $\sum_{i=1}^{\infty} a_i$ that is **not** absolutely convergent. Show that for any value c there is a reordering of the series such that $\sum_{i=1}^{\infty} a_{\tau(i)} = c$. (Note that this is a tough one. I had to do this as a student in my first year myself and I remember that it took me a while.)

As a consequence we have to be very careful when we are dealing with sums that are not absolutely convergent. You may think that this

has little relevance for physical problems, but let us think about the following problem.

Imagine we have a crystal made up of a periodic array of positive and negative charges. To make life simple, we consider this crystal to be one-dimensional and the charges arranged equally spaced with a distance 1 between neighboring charges (see figure). Generally in solid

Figure 1.5: This is a caption

state physics one assumes that the crystal has infinite extension. Now we would like to compute the potential energy of a single charge, which is just the work that is needed to remove it from the charge from the crystal. The potential energy is inversely proportional to the distance between two charges and is proportional to the product of the two charges. So, if we wish to compute the potential energy of the positive charge sitting at $n = 0$ we find

$$\sum_{n=-\infty}^{\infty}{}' \frac{(-1)^{(n+1)}}{n} = 2\ln 2 \tag{1.84}$$

where the term with $n = 0$ is excluded and if we sum in the order given by the summation index. Now it seems odd, that in a physical situation it should matter how we order the terms in the sum. one can try various ways out. One attempt could be to say that the crystal is finite in reality and for a finite sum it does not matter in which order one sums up the terms. But then one would still have the problem that one would like to make the crystal larger and larger (a real crystal contains

easily 10^{23} charges which is pretty much infinite for most purposes) and given that one can make the series diverge to infinity one then expects that one can make the energy per particle arbitrarily large. That's not very encouraging. The better answer is that indeed the fact that one can change the convergence of the series by changing the order of the terms reflects something physical. The order in which the terms occur reflects the way in which the crystal is formed. Let us consider two possible ways to build the crystal

1) We start with a particle in position 0. Then we add two particles with the opposite charge to the first one in the positions $+1$ and -1. Then we add the particles in the positions $+2$ and -2 and so on. This way yields exactly the summation order that we have used to obtain $\ln 2$ as the value of the sum. However, other ways can be considered.

2) An extreme case would be that one first adds all the particles that have the same charge as the one at the position 0. First we put the ones at position ± 2, then those at position ± 4 and so on. Of course this requires some work as we have to overcome the electrostatic repulsion. In fact, now we find that completing first the positioning of all the particles with the same charges requires infinitely much work or, in other words, its not possible.

But of course, as I have told you already that by reordering the sum I can achieve any limit I like we realize that depending on the way in which I build the crystal, different amounts of work will be required in the limit of very large crystals. So what looked like some mathematical curiosity has actually acquired some real physical meaning.

In any case, the main point to remember is that in a convergent series that is not absolutely convergent the order of the summation is absolutely crucial and must not be changed!

1.5.2 Methods to enhance the speed of convergence of series

In the previous sections I have introduced convergence criteria for series. Often it is the case that we know that the series converges to some finite value from our convergence criteria but we do not know the limit. Indeed, it is the standard situation that you cannot give a simple value

to the series, such as 2 or $\ln 2$ which you would accept as nice answers. But to be precise even $\ln 2$ is some real number for which we have given a nice name, but if you want to have the digits you will have to resort to actually working out the series it represents term by term until you have the required precision. Of course you would not like to compute too many terms in the sequence to obtain the desired precision. The number of terms that you need to compute for this depends both on the precision that you want to achieve but also on the nature of the sequence. Some sequences will converge rather fast so that you do not need to compute many terms others will converge very slowly. An example for a rapidly converging sequence is

$$\sum_{n=0}^{\infty} \frac{1}{n!} = e \cong 2.7182818265\ldots \tag{1.85}$$

where it is enough to compute the first 10 terms to get the final value to within $3 \cdot 10^{-8}$. An example for a rather slowly converging sequence is

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} = \ln 2 \cong 0.693\ldots \tag{1.86}$$

where you need to compute 10^7 terms to obtain the final value to within $5 \cdot 10^8$. So that would be really hard work. So, obviously you like to find ways to accelerate the convergence of the sequence. One obvious way to do this is to use the observation that the terms are alternatingly positive and negative in the sum. So why not grouping them in pairs so that you only have positive terms. This give the summation

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} = \sum_{n=1}^{\infty} \frac{1}{2n(2n-1)} = \ln 2 \cong 0.693\ldots \tag{1.87}$$

Now we have a sum that requires 10^7 terms to get to $\ln 2$ to within $2.5 \cdot 10^{-8}$. So we gained a little bit, but not really very much. In fact, you still have to compute roughly the same, very large, number of terms. We are really looking for something that requires instead of N terms maybe $\sqrt{N}$ terms or some even stronger improvement. In the following I will show you a neat trick to do that. It will not work for all series, but for quite a large number of them. Also, whenever you will

come into a situation where you may need to accelerate the convergence of a sum you may remember that there are methods to do so and then you can have a look for them in books. Ok, now lets start with what is called the Shanks Transformation. [9]

What is the main idea behind the Shanks transformation and all the other convergence enhancing transformations? When you are given a series then the terms in that series are not random, but they possess some regularities which may be hidden under some transient behaviour. In a sense the eventual value of the series will be hidden by this transient behaviour. But sometimes we can make guesses of the form of the transient behaviour and reduce their impact. In that way we will 'see' the final value more clearly. Let us study these ideas with an example.

Imagine first a particularly simple situation. Consider a series $\sum_{n=0}^{\infty} a_n$ where the n-th partial sum is given by

$$\sum_{k=0}^{n} a_k = s_n = a + \alpha q^n. \tag{1.88}$$

Clearly the final value of the sum will depend on the three free parameters a, α and q. We can determine these three values from the three equations

$$\begin{aligned} s_n &= a + \alpha q^n & (1.89)\\ s_{n+1} &= a + \alpha q^{n+1} & (1.90)\\ s_{n+2} &= a + \alpha q^{n+2}. & (1.91) \end{aligned}$$

Indeed we find that

$$a = \frac{s_{n+2}s_n - s_{n+1}^2}{s_{n+2} + s_n - 2s_{n+1}}. \tag{1.92}$$

So, if we would have indeed a sum where the partial sums have such a simple behaviour we have just shown in eq. (1.88), we could determine the final value of the sum from only three terms. The idea is now to

[9] Shanks was a mathematician who used early computers to determine many thousands of digits of the number π which he did via series expansions. To get good precision, he had to find methods for speeding up the convergence of those series.

apply this transformation to sums that may have a somewhat more complicated behaviour and hope that, while we will not get the exact value of the sum, we may accelerate the convergence. This means that instead of the very simple behaviour for the partial sums assumed above we will have

$$s_n = a(n) + \alpha q^n + \beta r^n + \ldots, \tag{1.93}$$

where $a(n)$ is hopefully some weak n-dependence. Then by the Shanks transformation we get a new sum

$$S(s_n) = \frac{s_{n+2}s_n - s_{n+1}^2}{s_{n+2} + s_n - 2s_{n+1}}, \tag{1.94}$$

for which we hope that it has a faster convergence. It is in fact not easy to prove when this method works and when it doesn't (also I simply do not know these proofs) so what most people do is to just try the transformation and then look whether the convergence improves. If it doesn't then one can try various other convergence enhancing transformations.

Let us now study a few examples to see that sometimes indeed, the Shanks transformation leads to excellent improvements in the rate of convergence.

Example 1: Let us consider the simple geometric series

$$s_n = \sum_{k=0}^{n} (-x)^k = \frac{1 - (-x)^{n+1}}{1 - (-x)} \tag{1.95}$$

When we apply the Shanks transformation we obtain

$$S(s_n) = \frac{1}{1 + x} \tag{1.96}$$

which is the exact sum. So here convergence is immediate.

Example 2: Let us consider the slightly more complicated series

$$\sum_{i=0}^{\infty} \left(1 - \frac{1}{2^{i+1}}\right) (-x)^i = \frac{1}{(x+1)(x+2)} \tag{1.97}$$

If we wish to evaluate the original sum for $x = 0.99$ we need about 1500 terms to get the first 6 decimal digits of the sum. That's a lot of work. Why could the Shanks transformation work in this case? To see this let us write out the partial sums which are

$$s_n = \sum_{i=0}^{n} \left(1 - \frac{1}{2^{i+1}}\right)(-x)^i = \frac{1}{(x+1)(x+2)} - \frac{(-x)^{n+1}}{x+1} + \frac{(-x/2)^{n+1}}{x+2}. \tag{1.98}$$

You can prove this by induction (Exercise). So you can see that the partial sums indeed have a very close similarity to the behaviour assumed in eq. (1.88). Indeed, for a general behaviour of the form

$$s_n = a + \alpha_1 q_1^n + \alpha_2 q_2^n \tag{1.99}$$

If we consider $1 \gg q_1 \gg q_2$ then we find for the Shanks transformation

$$\begin{aligned} S(s_n) &= a + \frac{\alpha_1\alpha_2(q_1-q_2)^2}{\alpha q_2(q_1-1)^2 q_1^n + \alpha_2 q_1 (q_2-1)^2 q_2^n} q_1^n q_2^n \\ &\cong a + \frac{\alpha_1\alpha_2 q_1^2}{\alpha_1 q_2 q_1^n + \alpha_2 q_1 q_2^{n+2}} q_1^n q_2^n \qquad (1.100) \\ &\cong a + \frac{\alpha_1\alpha_2 q_1^2}{\alpha_1 q_2 q_1^n} q_1^n q_2^n \qquad (1.101) \\ &\cong a + \alpha_2 \frac{q_1^2}{q_2} q_2^n \qquad (1.102) \end{aligned}$$

Before the Shanks transformation the dominant transient behaviour comes from q_1. But after the Shanks transformation we have an approximate transient behaviour which is proportional to q_2^n which is much weaker. In general one can show that the transient behaviour of the partial sums after the Shanks transformation is much smaller than for the original series. This suggests that the Shanks transformation is improved by successive iteration.

Of course you can iterate the Shanks transformation, ie once you have used the Shanks transformation to generate a new series, you can apply the Shanks transformation to that series and so on.

In the following table I show you the remarkable success of the Shanks transformation for this sum for $x = 0.99$ for which the exact result is $0.1680644\ldots$.

n	s_n	$S(s_n)$	$S(S(s_n))$	$S(S(S(s_n)))$
0	+0.5000000	–	–	–
1	−0.2425000	0.1554524	–	–
2	−0.6150875	0.1736603	0.1679926	–
3	−0.2945678	0.1654309	0.1680796	0.1680642
4	+0.6360096	0.1693366	0.1680609	0.1680644
5	−0.3001213	0.1674421	0.1680652	0.1680644
6	+0.6340036	0.1683706	0.1680642	0.1680644
7	−0.2944209	0.1679133	0.1680645	0.1680644
10	+0.6178369	0.1680827	0.1680644	0.1680644
15	−0.2597995	0.1680639	0.1680644	0.1680644
20	+0.5749627	0.1680644	0.1680644	0.1680644

You can clearly see the impressive enhancement in the rate of convergence.

Example 3: For our last example let us come back to the series for the $\ln 2$, ie

$$\sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} = \ln 2 \tag{1.103}$$

The numerical value is $\ln 2 \cong 0.6931472\ldots$ and we find

n	s_n	$S(s_n)$	$S(S(s_n))$	$S(S(S(s_n)))$
1	1.0000000	–	–	–
2	0.5000000	0.7000000	–	–
3	0.8333333	0.6904762	0.6932773	–
4	0.5833333	0.6944444	0.6931058	0.6931489
5	0.7833333	0.6924242	0.6931633	0.6931467
6	0.6166667	0.6935897	0.6931399	0.6931474
7	0.7595238	0.6928571	0.6931508	0.6931471
8	0.6345238	0.6933473	0.6931452	0.6931472
15	0.7253719	0.6931138	0.6931473	0.6931472
25	0.7127475	0.6931397	0.6931472	0.6931472
35	0.7072289	0.6931444	0.6931472	0.6931472

Again the improvement in the rate of convergence is really impressive.

If you want to learn more about other possible methods for enhancing the rate of convergence, then you should have a look at the book by CM Bender and SA Orszag "Advanced Mathematical Methods for Scientists and Engineers" published in Springer Verlag.

1.6 Complex numbers

So far we have got to know the natural numbers, the integers, the rational numbers and the real numbers. But even the set of real numbers is not really sufficient for all purposes. The real numbers are not too difficult to imagine as they are the limiting case of rational numbers. So, in some sense they are not so much different from the rational numbers. The completeness of the real numbers also tells us that all real Cauchy sequences converge to a real number. However, we have also seen some strange behaviour. For example, I showed you an iterative method for computing the square root of a numbers by employing the rule $x_{n+1} = \frac{1}{2}(x_n + a/x_n)$ to create a sequence that converges to $\sqrt{a}$ for $a > 0$. Then I suggested to you to try and see what happens when you replace a by $-a$. Then we have the sequence generated by $x_{n+1} = \frac{1}{2}(x_n + (-a)/x_n)$. It turns out that this sequence does not converge and in fact it is jumping around quite wildly. What is the reason for that? If the sequence would converge, then we would have for the limiting value x the equation $x = \frac{1}{2}(x - a/x)$ which has the 'solution' $x = \sqrt{-a}$ for $a > 0$. But does that make sense? Is there a way in which the square root of a negative number makes sense? Do we have to make sense of it? Indeed, these were questions that mathematicians asked themselves about 500 years ago and it took about 250 years from the realization that there is a problem to arriving at a satisfactory understanding. Indeed, for many years complex numbers were considered an amusing curiosity with no real significance. Let us see how the story went[10].

Commonly the birth of the complex numbers is dated as 1545 with the publication of Cardano's book *Ars Magna*. This, however, is not

[10]Much more than what I am going to tell you now can be found in the excellent book by T. Needham "Visual Complex Analysis" published in Clarendon Press which I thoroughly recommend as an introduction to complex analysis.

quite true as Girolamo Cardano introduced those numbers but then also dismissed them as useless. Why did he think that? He was certainly no fool so he must have had a reason. Amongst other things Cardano studied quadratic equations, ie equations of the form

$$x^2 = mx + b. \tag{1.104}$$

We can very easily write down the solutions for this equation. They are

$$x_{1/2} = \frac{1}{2}\left[m \pm \sqrt{m^2 + 4b}\right]. \tag{1.105}$$

That's ok, and all of you know this formula. But what do we do when $m^2 + 4b$ is negative. Then we have to take the square root of a negative number and the formula tells us that the solution to the quadratic equation is one of these 'imaginary' numbers. Well, you could accept this and say, ok these numbers occur so they must make sense. But Cardano, like most mathematicians of the time, thought more in terms of geometry and this way of thinking led him to conclude that the complex numbers do not make much sense. For him the quadratic equation simply means that we look for the intersection of a straight line $y = mx + b$ with a parabola $y = x^2$. Sometimes this problem

Figure 1.6: This is a caption

has a solution and sometimes it doesn't. Well, indeed the existence of solutions is indicated by a real solution x while the non-existence of a solution is indicated by the occurrence of roots of negative numbers

in the expression for x. So, for Cardano the occurrence of these weird imaginary numbers simply indicated that there is no solution and so he decided that these numbers had no particular use or meaning. Well, this was not as unreasonable as it may seem today. So, imaginary numbers lay dormant for quite a bit longer. Cardano also considered solutions of more complicated polynomials, namely those of cubic polynomials. In general they have the form

$$ay^3 + by^2 + cy + d = 0 \tag{1.106}$$

but dividing by a and using the substitution $y = x - \frac{b}{3a}$ this can be reduced to the standard form

$$x^3 = 3px + 2q \tag{1.107}$$

ie again we can easily interpret this equation as the quest for finding an intersection between a cubic curve $y = x^3$ and a straight line $y = mx+b$. You can see straight away (have a look at figure) by drawing this problem that this always has a solution. Ok, Cardano also derived a

Figure 1.7: This is a caption

formula to solve this problem.

$$x = \sqrt[3]{q + \sqrt{q^2 - p^3}} + \sqrt[3]{q - \sqrt{q^2 - p^3}} \tag{1.108}$$

which is not at all easy to find. Either Cardano didn't realize or he didn't care anymore after he decided that the imaginary numbers are

not very significant, but here a remarkable phenomenon happens. It was realized by Bombelli more than 30 years later. Clearly, when $p^3 > q^2$ then there are again roots of negative numbers. And this time their occurrence is more serious because they cannot be dismissed with the remark that they correspond to the lack of actual solution for the intersection of two curves because we know that there always is such an intersection. Bombelli considered

$$x^3 = 15x + 4 \tag{1.109}$$

which clearly has the solution $x = 4$ but for which Cardanos formula yields

$$x = \sqrt[3]{2 + 11i} + \sqrt[3]{2 - 11i} \tag{1.110}$$

where I have abbreviated $\sqrt{-1}$ with i. Now Bombelli was trying to make sense of the situation and was fighting very hard with this problem. Finally he had a very good idea indeed (he called it a "wild thought" which it certainly was at that time). Perhaps, he thought, the solution $x = 4$ is indeed equal to $x = \sqrt[3]{2 + 11i} + \sqrt[3]{2 - 11i}$. So he made the Ansatz

$$\sqrt[3]{2 + 11i} = 2 + in \quad \sqrt[3]{2 - 11i} = 2 - in. \tag{1.111}$$

Of course he already made some assumption here, namely that 'real' part of the numbers are equal, but that seemed quite a decent guess. Furthermore he had to make the very reasonable assumption that one can add imaginary numbers $z_1 = a_1 + ib_1$ and $z_2 = a_2 + ib_2$ according to the law

$$z_1 + z_2 = (a_1 + ib_1) + (a_2 + ib_2) = (a_1 + a_2) + i(b_1 + b_2). \tag{1.112}$$

Of course, he also needed to compute $(2+in)^3$, ie he needed some rules for multiplication. He chose quite naturally

$$(a_1 + ib_1)(a_2 + ib_2) = a_1b_1 + i(a_1b_2 + b_1a_2) + i^2b_1b_2 \tag{1.113}$$

and he made the reasonable assumption that $i^2 = -1$ given that we defined i as $\sqrt{-1}$ and he arrived at

$$(a_1 + ib_1)(a_2 + ib_2) = a_1b_1 - b_1b_2 + i(a_1b_2 + b_1a_2). \tag{1.114}$$

With these rules he was now able to verify $(2 \pm i)^3 = 2 \pm 11i$, ie $\sqrt[3]{2 \pm 11i} = 2 \pm i$. Therefore he was able to make sense of imaginary numbers at least in this situation and also motivated the multiplication law for these numbers.

For the next 250 years there was not too much progress in the study of complex numbers. One of the reasons was the problem that no-one was really able to visualize these numbers, quite unlike the real numbers for example which are points on a line. The real progress started when Wessel and Argand independently realized that one can represent imaginary numbers in a plane, the complex plane. The real numbers would form one axis and the other axis is formed by the real multiples of i. This in itself would not have been a big progress if it weren't for the fact that now the laws for addition and multiplication gained a very simple geometrical intuition. Indeed, the addition of two complex numbers was just the addition of the two corresponding vectors in the complex plane. That's fairly straightforward. Less straightforward is the realization that the multiplication of two complex numbers (following Bombellis rule) is equivalent to obtaining the resulting vector by multiplying the length of the two vectors and adding the angles that they make with the real axis. (Prove this as an exercise). This is very nice indeed. The fact that multiplication is a combined stretching and rotation makes it a bit more plausible how Euler could get the intuition for his famous Euler formula

$$z = re^{i\phi} = r(\cos\phi + i\sin\phi) \tag{1.115}$$

where r is a real number giving the length of the corresponding vector and ϕ gives the angle it makes with the x-axis.

This formula is really useful in many aspects of mathematics and especially physics.

Example: Verify the law

$$e^{i\theta} + e^{i\phi} = 2\cos\frac{\theta - \phi}{2} e^{\frac{i(\theta+\phi)}{2}} \tag{1.116}$$

both by calculation and with a picture.

Examples: Derive $\cos(3\phi) = 4(\cos(\phi))^3 - 3\cos\phi$ using complex numbers.

Example: Let $S_n = \sum_{k=1}^{n} \cos(2k-1)\theta$. Show that

$$S_n = \frac{\sin 2n\theta}{2\sin\theta} \tag{1.117}$$

Example: Finally let us study again the map

$$z_{n+1} = \frac{1}{2}(z_n + \frac{a}{z_n}) \tag{1.118}$$

which allows us to compute $\sqrt{a}$ numerically. Remember that for a real initial value z_0 the sequence converges to the positive square root $\sqrt{a}$ if $z_0 > 0$ and to the negative square root $\sqrt{a}$ if $z_0 < 0$. Now let us assume however, that we start with a purely imaginary value $z_0 = ir_0$. In that case we get a recursion relation for $z_n = ir_n$ which is given by

$$r_{n+1} = \frac{1}{2}\left(r_n - \frac{1}{r_n}\right)$$

Now if we start with some value of r_0 the sequence of r_n does not converge at all. Indeed, it is jumping back and forth quite wildly. For example, $g = \frac{1+\sqrt{5}}{2} = 1.618....$ maps to $0.5, -0.5, 0.291666..., -1.56845, ...$ etc. But some r_0 behave differently. An example is $r_0 = 1+\sqrt{2}$ which maps to $1, 0, \infty$. This seems to be some rather wild behaviour altogether and we have to try to get some handle on the problem to be able to understand what can and cannot happen. To achieve this, we use a rather nice little trick, namely we make the substitution

$$r = -\text{cotan}(\alpha\pi) \equiv -\frac{\cos(\alpha\pi)}{\sin(\alpha\pi)}$$

Now we can insert this into the recursion relation for r_n to find

$$\begin{aligned}
\frac{1}{2}\left(r_n - \frac{1}{r_n}\right) &= \frac{1}{2}\left(-\frac{\cos(\alpha_n\pi)}{\sin(\alpha_n\pi)} - \frac{\sin(\alpha_n\pi)}{\cos(\alpha_n\pi)}\right) \\
&= -\frac{\cos^2(\alpha_n\pi) - \sin^2(\alpha_n\pi)}{2\cos(\alpha_n\pi)\sin(\alpha_n\pi)} \\
&= -\text{cotan}(2\alpha_n\pi) = r_{n+1}
\end{aligned}$$

Taking into account that $cotan(\alpha\pi) = cotan((\alpha+1)\pi)$, we find as a consequence a recursion relation for the α_n which is

$$\alpha_{n+1} = 2\alpha mod 1$$

where the symbol $amodx$ means that we take the remainder of a upon division by x. Here x has to be a natural number. So, in the above recursion relation we only keep the digits of $2\alpha_n$ after the decimal point. Now the behaviour of the recursion relation becomes far more transparent. If we consider α_n written in binary notation, we see that one step of the recursion relation amounts to shifting the digits in the binary representation 'to the left' by one place. Therefore, we can see that there are three different types of behaviour.

1. If the binary representation of α_0 terminates after n digits, then the recursion relation will lead to $\alpha_n = 0$ and as a consequence $r_n = \infty$, ie the recursion diverges.

2. If the binary representation of α_0 is periodic and non-terminating (for example $2/3 = 0.10101010...$) with period n, then the recursion relation will lead to period n.

3. If the binary representation of α_0 is aperiodic and non-terminating, then the sequence of the α_n is also aperiodic and non-terminating.

So, what we observe here, is that a tiny difference in the initial conditions can lead to a hugely different behaviour of the sequence generated by the recursion relation. This phenomenon, that can be observed for many recursion relations, is a manifestation of 'chaos'. Indeed, initially physicists were not aware of this behaviour, but nowadays it is well-known that chaos reigns in many areas of physics. The dynamics of the solar system is one example.

So, two almost identical initial conditions diverge which is a signature for chaos. Then to quantify the degree of chaos one could quantify the rate at which the two closely spaced points begin to diverge. This is usually done via the so-called Lyapunov exponent, that is defined by

$$\lambda = \ln(\frac{\alpha_{n+1}}{\alpha_n}) \tag{1.119}$$

This tells you by how much the value of α grows in successive steps of the iteration and it therefore also tells you the rate by which any difference between the initial conditions grow. In the above recursion you find $\lambda = \ln 2 = 0.693.....$

An even more wild a weird behaviour can be found from the recursion relation

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

Chapter 2

Functions of real variables

In the last section I have introduced some of the various concepts of numbers that are in use in physics and mathematics. You have also learnt about sequences and series and their properties. We viewed a sequence just as an ordered list of numbers. But one can also consider a sequence as a function that maps the natural numbers into the real (or complex) numbers, namely with each natural number n we associate exactly one other real (or complex) number, namely the n-th element of the sequence. So, here a function is a way of enumerating (using the natural numbers) and ordering a set of real (complex) numbers. See figure ... for a way to depict such a sequence graphically. The horizontal

Figure 2.1: This is a caption

axis is used to enumerate the natural numbers and the vertical direction

is used to plot the value of the corresponding element of the sequence. As you can see, there are plenty of gaps, simply because we enumerate the elements of our sequence by natural numbers. Of course, we could have also made a more fancy sequence, namely we could have associated with every rational number another real (or complex) number. This is generally not what is done, but it is possible, because there is a one-to-one mapping between the natural and the rational numbers. Such sequences could also be plotted and they would look as if there were no gaps, but of course you know that there are, simply because not a single irrational number is used to enumerate the sequence. Therefore, if you would like to draw the graph of this sequence, analogously to that of the sequence enumerated by natural numbers, you could not let the pen on the paper all the time, but you would have to lift it up and down to draw single dots for each of the elements of the sequence. The reason for that is of course the fact that the rational numbers have lots and lots of gaps between them, namely the irrational numbers such as $\sqrt{2}$ etc. Ideally we would like to fill these gaps as well. So, we need sequences that are even more densely enumerated that by rational numbers. This idea then leads to the concept of functions from real numbers to real (or complex) numbers. This little heuristic argument should indicate to you that those function are in a sense continuous versions of sequences. This innocious step however, allows us to introduce new concepts and generalizations of old concepts which make the idea of functions so much more powerful than that of sequences and series. As a bonus we will obtain new tricks to evaluate discrete sums.

2.1 More about sets

In the first chapter we have encountered the basic concept of sets as well as some of the operations that one can do with them. These operations included the intersection and the union of sets and some more. We have also considered the concept of limits and we have encountered the remarkable phenomenon that a sequence of rational numbers may not converge to any rational number itself. However, any sequence of natural numbers that converges, actually converges against a natural number. Therefore convergence and the properties of the limiting value

of a sequence within a set can tell you something about the basic structure of a set. We will now distinguish two main types of sets, namely open and closed sets.

Definition 30 *A subset $\mathcal{C}$ of Ω is called closed, when it contains the limit points for* all *convergent sequences that can be formed with its elements.*
A subset $\mathcal{O}$ of Ω is called open when it is the complement of a closed set (the complement of $\mathcal{O}$ is defined as $\bar{\mathcal{O}} = \{x \in \Omega | x \notin \mathcal{O}\}$).

Let us consider some examples from the real numbers to make these definitions clear.

Examples: An interval is a subset of the real numbers which can be open, closed and half-open. These three types are defined by

- A closed interval from x_0 to x_1 denoted by $\mathcal{I}_1 = [x_0, x_1]$ is an interval that contains all the numbers x which satisfy $x_0 \leq x \leq x_1$. Clearly, any sequence $(a_n)_{n=1,...}$ from points in this interval satisfies $x_0 < a_n < x_1$. Therefore, the limit $a = \lim_{n\to\infty} a_n$ can neither be large than x_1 nor smaller than x_0, therefore $x_0 \leq a \leq x_1$. As a consequence all the limiting points of sequences of elements from the interval $\mathcal{I}_1$ lie in $\mathcal{I}_1$.

- An open interval from x_0 to x_1 denoted by $\mathcal{I}_2 =]x_0, x_1[$ is an interval that contains all the numbers x which satisfy $x_0 < x < x_1$ but **not** the points x_0 and x_1. Clearly, this interval is the complement of the set of points $]-\infty, x_0] \cup [x_1, \infty[$ which is a set which contains all the limiting points of sequences made up of elements of that set.

- A half-open interval from x_0 to x_1 is denoted by $\mathcal{I}_3 =]x_0, x_1]$ when it contains all the numbers x which satisfy $x_0 < x \leq x_1$ and it is denoted by $[x_0, x_1[$ when it contains all the points $x_0 \leq x < x_1$. This interval is neither open nor closed.

- The points x_0 and x_1 are called boundary points of the interval.

- The set $]x_0, \infty[= \{x | x > x_0\}$ is open because both boundary points x_0 and ∞ are not part of the set.

- The set $[x_0, \infty[= \{x | x \geq x_0\}$ is closed because its complement is open.
- If two sets $\mathcal{A}$ and $\mathcal{B}$ are open, so is their intersection $\mathcal{A} \cap \mathcal{B}$
- Any *finite* intersection of open intervals $\mathcal{A}_i$, denoted by $\cap_{i=1}^{n} \mathcal{A}_i$ is open.
- An infinite intersection of open intervals $\mathcal{A}_i$, denoted by $\cap_{i=1}^{\infty} \mathcal{A}_i = \lim_{n\to\infty} \cap_{i=1}^{n} \mathcal{A}_i$ can be open or closed depending on the sequence. For example $\cap_{i=1}^{\infty}]-\frac{1}{i}, \frac{1}{i}[= \{0\}$ is closed while $\cap_{i=1}^{\infty}]0, i[=]0, \infty[$ is open.
- The empty set $\emptyset$ is open because there is no point that does not have a ball around it that lies in the set.
- The set of real numbers $\mathbb{R}$ is open since for every point there is a ball that contains that point and lies entirely in $\mathbb{R}$.
- Now, maybe surprisingly, the sets $\emptyset$ and $\mathbb{R}$ are both closed as their complements are open.

Exercises: Prove the following statements

- Any finite intersection of open sets is open.
- The intersection of infinitely many open sets may be open or closed.
- The intersection of any number of closed sets is a closed set.

Open and closed sets play some role in the definition of functions and their properties, which is why I mentioned them here.

2.2 The basic definition of a function

In the general introduction to this section I have outlined the idea of functions. I have said that sequences are a mapping of the natural numbers to the real numbers etc. In the following I would like to

introduce names for many of the basic concepts of functions to allow us to speak more easily about them.

Definition 31 *A real function $f : \mathcal{D} \mapsto \mathcal{I} \subset \mathbb{R}$ associates with every $x \in \mathcal{D} \subset \mathbb{R}$ a unique real number $y = f(x) \in \mathcal{I}$.*

- *The set $\mathcal{D} \subset \mathbb{R}$ is called the* domain *of the function. We will also write $\mathcal{D} = dom(f)$.*
- *The set $\mathcal{I} \subset \mathbb{R}$ of all possible values that the function can assume is defined by $\mathcal{I} = \{y | \exists x \in \mathcal{D} such\ that y = f(x)\}$ is called the* image *of the function. We will also write $\mathcal{I} = im(f)$.*
- *The set $\mathbb{R}$, which includes $\mathcal{I}$ is also called the* range *or* codomain *of the function.*

Examples: 1) The function $f : \mathbb{R} \mapsto \mathbb{R}$ which is given by $f(x) = x$ has the domain $\mathcal{D} = \mathbb{R}$, ie it is defined on all real numbers. Its image is also the set of all real numbers.

2) The function $f : \mathbb{R}_+ \mapsto \mathbb{R}$ which is given by $f(x) = \sqrt{x}$ has the domain $\mathcal{D} = \mathbb{R}_+$, ie it is defined on all positive real numbers. Its image is also the set of all positive real numbers.

3) The function $f : \mathbb{R} \mapsto \mathbb{R}$ which is given by $f(x) = x^2$ has the domain $\mathcal{D} = \mathbb{R}$, ie it is defined on all real numbers. Its image is also the set of all positive real numbers.

The domain of a function can have many forms. It can be discrete such as the natural numbers or the rational numbers in which case we are taken back to the case of sequences. It can of course also be continuous.

We have now defined functions and so we should learn some basic operations on functions. Firstly, a function $f : dom(f) \mapsto im(f)$ maps the domain of f to the image of f. Of course, I could then try to apply another function $g : dom(g) \mapsto im(g)$ to the image of f again, see the figure. This is only well-defined when $im(f) \subset dom(g)$. Then we get a new function denoted by $h = g \circ f$. We have that $dom(h) \subset dom(g)$.

Figure 2.2: This is a caption

Of course $f = g$ is a possibility and we may iterate the function many times. Then we denote

$$\begin{aligned} f^2(x) &= f \circ f(x) = f(f(x)) && (2.1)\\ f^n(x) &= \underbrace{f \circ ... \circ f}_{n\,times}(x) = f(f(\dots f(x)\dots)) && (2.2) \end{aligned}$$

Note the difference between the notation $f^n(x)$ which is the n-fold iterated application of the function f and the notation $(f(x))^n$ which is the n-th power of the value $f(x)$. To make the difference clearer, consider $f(x) = x^2$ and $n = 4$. Then we have $(f(x))^n = x^8$ and $f^n(x) = x^{16}$. For general n we have $(f(x))^n = x^{2n}$ and $f^n(x) = x^{2^n}$ which is a clear difference.

Examples: Already in the first chapter you have seen a specific example for an iterated map, namely the one for the computation of the square root which I now write in functional form

$$x_{n+1} = f(x_n) = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) \tag{2.3}$$

so that we have

$$x_n = f^n(x_0). \tag{2.4}$$

A little later-on we will see how, using the concept of functions, we can derive a whole lot of useful iteration procedures to solve equations other than $x^2 = a$.

Another interesting map is the so-called logistic map which is defined as

$$f(x) = ax(1 - x) \tag{2.5}$$

which exhibits chaos as well.

Given a function f we would sometimes invert it, ie for a given value from the image of the function f we would like to find out which element from the domain of f led to it, for an $y \in im(f)$ we would like to find the $x \in dom(f)$ such that $y = f(x)$. The formal definition is

Definition 32 *For a one-to-one function $f : dom(f) \mapsto im(f)$ we define the inverse function, denoted by $f^{-1} : im(f) \mapsto dom(f)$ by*

$$f^{-1}(y) = x \tag{2.6}$$

such that x is the unique value that satisfies $f(x) = y$.

A word of caution is necessary here as not all functions can be inverted. Indeed, the condition that f is a one-to-one function from $dom(f)$ to $im(f)$ is very important. If the function would associate with two different values x_0 and x_1 the same result $y = f(x_0) = f(x_1)$ then we could not invert it because given y we would be unable to decide on a unique x that ensures $f(x) = y$.

But iterating functions is not the reason why they are important. In fact, if it were just for the iteration of functions there would have been no real need to introduce the concept of functions but we could have stuck with sequences just as well. The real use of functions comes from the ideas of continuity, differentiation, integration etc. This will be introduced in the next section.

2.3 Continuity

Functions may be very regular but they may also be very irregular. A first property of regularity is that of continuity. In a hand-waving

way this implies, that you can draw the function without ever lifting the pencil or in other words, there are no jumps in the function. This definition is useful for intuition, but not so good when you actually want to prove something. So, here is a rigorous definition.

Definition 33 *A real function $f : \mathcal{D} \mapsto \mathbb{R}$ is called continuous in the point x_0 if for all $\epsilon > 0$ there is a $\delta > 0$ such that for all x with $|x - x_0| \leq \delta$ we have that $|f(x) - f(x_0)| \leq \epsilon$.*

The function f is called continuous in an interval $\mathcal{I}$ if it is continuous for all $x \in \mathcal{I}$.

This definition has quite some similarities with the limit of sequences. To see this more clearly, let us see how we could define the limit of a function, ie we would like to give sense to the expression $\lim_{x \to x_0} f(x) = b$. Clearly for sequences, ie functions $f : \mathbb{N} \mapsto \mathbb{R}$ we defined the limit as $\lim_{n \to \infty} f(n) = b$ by the requirement that for all $\epsilon > 0$ there has to be an n_0 such that for all $n \in [n_0, \infty[$ we have $|f(n) - b| \leq \epsilon$. That means the closer we come into the neighborhood of ∞, the closer we come to the value b. Therefore a natural definition for convergence for functions is

Definition 34 *A real function $f : \mathcal{D} \mapsto \mathbb{R}$ converges to the point b in the limit of x approaching x_0 if for all $\epsilon > 0$ there is a $\delta > 0$ such that for all $x \in [x_0 - \delta, x_0 + \delta]$ we have that $|f(x) - b| \leq \epsilon$.*

Now we realize that continuity of the function f in the point x_0 equivalent to the statement $\lim_{x \to x_0} f(x) = f(x_0)$. We can visualize these statements nicely in figure such as those in figure ....

Continuous functions are also nicely behaved in the sense that they do not grow too fast. This is captured in

Lemma 35 *A function f that is continuous on $\mathcal{D}$ is bounded on every closed interval $[x_0, x_1] \subset \mathcal{D}$.*

Proof: Assume that f is not bounded on the interval $[x_0, x_1]$ then I can split the interval in two equal halves. In at least one half the function will be unbounded. Chose that interval, call it middle point y_1. Cut the interval in half again and identify the interval in which the function is unbounded. Name the middle point y_2. Continue in this way ad

Figure 2.3: This is a caption

infinitum so that you will define a sequence $\{y_i\}$ that converges to y such that $\lim_{y_i \to y} f(y_i) = \infty$. So that the sequence does not converge. This is a contradiction so that the function has to be bounded. This completes the proof.

Again, continuity is a mathematical concept which is highly convenient, but it is also a property that you cannot verify experimentally in a lab. You can just check it to higher and higher precision. If the function that you are measuring has a jump that is smaller than your measurement precision, then you cannot see it. Of course, if the function does indeed have a discontinuity in some point, then you will eventually be able to verify that there is a jump, once your experimental precision is sufficiently high. However, the continuity in a certain point cannot be proven by measurement. Nevertheless, continuity is a very useful concept both in theoretical physics and in mathematics.

Examples:

1. The function $f(x) = x$ is continuous in every point x_0. This can be proven easily by starting with any $\epsilon > 0$ and choosing $\delta = \epsilon$ which is sufficient to check that for all x with $|x - x_0| \leq \delta$ we have $|f(x) - f(x_0)| = |x - x_0| \leq \epsilon$.

2. The function $f(x) = x^2$.

Proof: Exercise

3. Given two functions $f : \mathcal{D} \mapsto \mathbb{R}$ and $g : \mathcal{D} \mapsto \mathbb{R}$ that are both continuous on the whole interval $\mathcal{D}$ and let α and β be real numbers, then the functions

   - $\alpha f + \beta g : \mathcal{D} \mapsto \mathbb{R}$ is also continuous on the interval $\mathcal{D}$.
   - $f \cdot g : \mathcal{D} \mapsto \mathbb{R}$ is continuous on the interval $\mathcal{D}$.
   - If g has no zeros in $\mathcal{D}$ then also $\frac{f}{g}$ are continuous in $\mathcal{D}$.

   The proofs for this are an exercise.

4. If f is continuous, then so is the function $g(x) = (f(x))^2$.
   Proof: Exercise.

Theorem 36 *Given a continuous function $f : [a, b] \to \mathbb{R}$. Then for every $y \in [f(a), f(b)]$ there is an $x \in [a, b]$ such that $y = f(x)$*

Proof: For $y = f(a)$ and $y = f(b)$ the statement is evidently true. Furthermore, let us assume $f(a) < f(b)$. An analogous proof works for $f(a) > f(b)$. For $f(a) = f(b)$ the statement is trivial. Therefore, let us assume that $f(a) < y < f(b)$. Let define a sequence of intervals according to the following rule. We start with $\mathcal{I}_0 = [a, b] \equiv [a_0, b_0]$. Now, given an interval $\mathcal{I}_n = [a_n, b_n]$ with $y \in [f(a_n), f(b_n)]$ we examine the two intervals $[a_n, \frac{a_n+b_n}{2}]$ and $[\frac{a_n+b_n}{2}, b_n]$. If $y \in [f(a_n), f(\frac{a_n+b_n}{2})]$ then $\mathcal{I}_{n+1} = [a_n, \frac{a_n+b_n}{2}] \equiv [a_{n+1}, b_{n+1}]$. If $y \in [f(\frac{a_n+b_n}{2}), f(b_n)]$ then $\mathcal{I}_{n+1} = [\frac{a_n+b_n}{2}, b_n] \equiv [a_{n+1}, b_{n+1}]$. Therefore we get a monotonously increasing sequence $(a_n)_{n=1,\ldots}$ and a monotonously decreasing sequence $(b_n)_{n=1,\ldots}$. The sequences are evidently bounded and therefore they converge. As we also have $|b_n - a_n| = 2^{-n}|b_0 - a_0|$ the limites coincide, ie $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n$. As the function is continuous, we have that $\lim_{n\to\infty} f(a_n) = \lim_{n\to\infty} f(b_n) = f(a_\infty)$ where $a_\infty \in [a, b]$. This finishes the proof.

2.3.1 Functions of many variables

If have shown you that continuity and taking the limit of a function are closely related. This is all pretty harmless when we are talking about functions of a single variable. But in physics often functions can have more than a single variable. For example if you have a particle that is moving in a plane, then the absolute value of its velocity will depend on two variables that specify its position, ie $v = v(x, y)$. In such functions there are some important subtleties involved in carrying out limites. To illustrate this, let me consider the following function

$$f(x, y) = \frac{x^2}{x^2 + y^2} \tag{2.7}$$

which is defined everywhere except for the point $(x, y) = (0, 0)$. Now we could try to find the limit of this function

$$\lim_{(x,y)\to(0,0)} f(x, y) =? \tag{2.8}$$

What do we mean by this limit? A real function of two variables $f : \mathcal{D} \mapsto \mathbb{R}$ converges to b in the limit of (x, y) approaching (x_0, y_0) if for all $\epsilon > 0$ there is a $\delta > 0$ such that for all $|(x, y) - (x_0, y_0)| \leq \delta$ we have that $|f(x, y) - b| \leq \epsilon$. This is just the definition that we brought above already for functions of a single variable. The only thing that I should specify is what $|(x, y) - (x_0, y_0)|$ actually means. Well, it is the distance between the two vectors (x, y) and (x_0, y_0), ie we have $|(x, y) - (x_0, y_0)| = \sqrt{(x - x_0)^2 + (y - y_0)^2}$.

So, now let us see whether we can possibly have convergence for the function $f(x, y) = \frac{x^2}{x^2+y^2}$. If we chose $y = x$ then we would have $f(x, x) = \frac{1}{2}$ for any value of x as long as it is unequal to 0. Because $|(x, x) - (0, 0)| = \sqrt{2}x$ we have in every proximity to $(0, 0)$ points (x, y) where $f(x, y) = \frac{1}{2}$. Likewise, for $y = 2x$ we have $f(x, 2x) = \frac{1}{5}$ for any value of x as long as it is unequal to 0. Because $|(x, 2x) - (0, 0)| = \sqrt{5}x$ we have in every proximity to $(0, 0)$ points (x, y) where $f(x, y) = \frac{1}{5}$. Therefore the function does not converge towards any fixed value. That in itself would not be so surprising. There are clearly functions that converge and others that don't. However, let us now take first the limit

in y and then in x. Then we find

$$\lim_{x\to 0}(\lim_{y\to 0}\frac{x^2}{x^2+y^2}) = \lim_{x\to 0} 1 = 1 \tag{2.9}$$

So, here the limit exists. If we take the limit the other way around the we find

$$\lim_{y\to 0}(\lim_{x\to 0}\frac{x^2}{x^2+y^2}) = \lim_{y\to 0} 0 = 0 \tag{2.10}$$

So, we make a very important observation here. Limites cannot necessarily be interchanged. Of course there may be compelling physical reasons for choosing a particular ordering for the limits but in the absence of such reasons, one has to take great care with multiple limits.

The problem in our example was that the limit $\lim_{(x,y)\to(0,0)} f(x,y)$ does not exist and as a consequence taking the limits one by one, first in x and then in y and vice versa gives different results. However, when we know that the limit $\lim_{(x,y)\to(0,0)} f(x,y)$ exists, then we can also do the iterated limites, or more precisely.

Theorem 37 *If* $\lim_{(x,y)\to(a,b)} f(x,y) = c$ *and if* $\lim_{y\to b} f(x,y)$ *exists then so does* $\lim_{x\to a}(\lim_{y\to b} f(x,y))$ *and it equals* c.

Proof: As the total limes exists, then for all $\epsilon > 0$ there is a $\delta > 0$ such that $|f(x,y)-c| \leq \epsilon$ for all $|x-a| \leq \delta$ and $|y-b| \leq \delta$. From the existence of the limit $\lim_{y\to b} f(x,y)$ it follows that $|\lim_{y\to b} f(x,y) - c| \leq \epsilon$ for all $|x-a| \leq \delta$ and this finishes the proof.

2.4 Convexity I

Now let us study some other most useful property of function that is quite closely related to continuity. This is the concept of convex and concave function which will later turn out to be very useful for deriving in a simple way a number of often used inequalities.

Definition 38 *A function* f *is called convex on the interval* $[a,b]$ *when for all intervals* $[x,y] \subset [a,b]$ *and all* $\lambda \in [0,1]$ *we have that*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \tag{2.11}$$

A function f *is called concave, if the function* $-f$ *is convex.*

Geometrically this means that for a convex function for any pair of points the function lies below the straight line joining the points $(x, f(x))$ with $(y, f(y))$. This is shown in figure ... where I have plotted two functions. The left function is convex while the function on the right is concave.

Figure 2.4: This is a caption

Examples:

1. The function $f(x) = x$ is evidently both convex and concave.
2. The function $f(x) = x^2$ is convex. This can be seen by

$$\begin{aligned} \lambda f(x) + (1-\lambda)f(y) - f(\lambda x + (1-\lambda)y) &= \lambda x^2 + (1-\lambda)y^2 - (\lambda x + (1-\lambda)y)^2 \\ &= \lambda(1-\lambda)(x-y)^2 \\ &\geq 0 \end{aligned}$$

   This completes the proof.

3. It is more difficult to prove that the function $f(x) = x^{2^n}$ is concave for all $n \in \mathbb{N}$. To see this directly one would have to verify in particular for $\lambda = 1/2$ the equation

$$(\lambda x + (1-\lambda)y)^{2n} \leq \lambda x^{2n} + (1-\lambda)y^{2n} \tag{2.12}$$

   This is indeed a useful inequality to know, but maybe not so easy to prove. In the following I will use a useful lemma to solve this

problem. In the next subsection however we will find a more direct and much easier way of approaching this questions.

Lemma 39 *Given a convex function $f : \mathcal{D} \to \mathcal{I}$ and another function g which is both convex and monotonuously growing on $\mathcal{I}$. Then the composition $g \circ f$ is also a convex function on D.*

Proof: Convexity of f on $\mathcal{D}$ means that for all $x_0, x_1 \in \mathcal{D}$ and $\lambda \in [0,1]$ we have $f(\lambda x_0 + (1-\lambda)x_1) \leq \lambda f(x_0) + (1-\lambda)f(x_1)$. Then we have from the monotonicity of g on the interval $\mathcal{D}$ as well because

$$\begin{aligned} g \circ f(\lambda x_0 + (1-\lambda)x_1) &= g(f(\lambda x_0 + (1-\lambda)x_1)) \\ &\leq g(\lambda f(x_0) + (1-\lambda)f(x_1)) \\ &\leq \lambda g \circ f(x_0) + (1-\lambda) g \circ f(x_1). \end{aligned}$$

This proves the convexity of $g \circ f$.

Now we realize that the composition of $f(x) = x^2$ with itself is simply $f \circ f(x) = f(f(x)) = f(x^2) = x^4$. As the image of f are the positive real numbers on which f is convex and monotonous we conclude from the lemma, that $g_2(x) = x^4$ is convex as well. By induction we can then conclude that $g_n(x) = x^{2^n}$ is convex.

4. The function $f(x) = \ln x$ is concave. Again the proof of convexity is not that straightforward and is equivalent to a useful inequality. Concavity requires that

$$\ln(\lambda x + (1-\lambda)y) \geq \lambda \ln(x) + (1-\lambda)\ln(y) \tag{2.13}$$

for all $\lambda \in [0,1]$. This is equivalent to the following statement. For all positive p and q with $\frac{1}{p} + \frac{1}{q} = 1$ and all positive x and y we have

$$x^{1/p} y^{1/q} \leq \frac{x}{p} + \frac{y}{q} \tag{2.14}$$

The proof of this inequality is not so straightforward and again I leave the proof for later.

5. The function $f(x) = e^x$ is convex on all real numbers. One could try to prove this directly by verifying that

$$e^(\lambda x_0 + (1-\lambda)x_1) \leq \lambda e^{x_0} + (1-\lambda)e^{x_1} \tag{2.15}$$

but again a useful lemma comes in very handy indeed because it will allow us to conclude the convexity of e^x from the concavity of $\ln x$. We have

Lemma 40 *If a monotonously growing and invertible function $f : \mathbb{R} \to \mathbb{R}$ is convex exactly if the inverse function $f^{-1} : \mathbb{R} \to \mathbb{R}$ is concave.*

Proof: First note that the inverse of a monotonically growing function is also monotonically growing. This can be seen as follows. Monotonicity is equivalent to

$$x_1 \geq x_2 \Leftrightarrow f(x_1) \geq f(x_2) \tag{2.16}$$

Now using $y_i = f(x_i)$ we have the inverse function $x_i = f^{-1}(y_i)$ and as a consequence

$$f^{-1}(y_1) \geq f^{-1}(y_2) \Leftrightarrow y_1 \geq y_2 \tag{2.17}$$

which confirms the monotonicity of the inverse function f^{-1}.

As f is convex, we have

$$\lambda f(x_0) + (1-\lambda)f(x_1) \geq f(\lambda x_0 + (1-\lambda)x_1)$$

Denoting again $y_i = f(x_i)$ and using $f \circ f^{-1} = \mathbb{1}$ we find

$$f^{-1}(\lambda y_0 + (1-\lambda)y_1) \geq f^{-1}(f(\lambda x_0 + (1-\lambda)x_1)) = \lambda f^{-1}(y_0) + (1-\lambda)f^{-1}(y_1)$$

which is the condition for the concavity of f^{-1}. The reverse direction of the proof works analogously. This finishes the proof.

As the $f(x) = e^x$ is the inverse function of $g(x) = lnx$ we can use the concavity of $g(x) = \ln x$ to conclude the convexity of $f(x) = e^x$.

Before we go on to the next topic which will yield as a byproduct a simple way to prove convexity I would like to show you that convexity of a function is closely related to its continuity, in fact

Lemma 41 *A bounded function $f : \mathcal{D} \to \mathbb{R}$ that is convex on an open interval is continuous in the same interval.*

Proof: We will proof this by assuming that the function is discontinuous and conclude that it cannot be convex then. To get an idea for the proof have a look at the figure The function $f(x)$ has a jump at

Figure 2.5: This is a caption

the position x_0. This is the only way that a bounded function can be discontinuous. Now we can see quite clearly, that if we draw a line from the point $(x_0 - \epsilon, f(x_0 - \epsilon))$ to the point $(x_0 + \epsilon, f(x_0 + \epsilon))$, then some parts of the function will lie above this line which cannot be true for a convex function. Of course this is not a proof yet, but it gives us the intuition for how to make the proof.

Assume that the bounded function f is discontiuous in the point x_0. This means that it has a jump in x_0 which means that

$$\lim_{\mu \to 0, with \mu > 0} f(x_0 - \mu) = b_- \tag{2.18}$$

$$\lim_{\mu \to 0, with \mu > 0} f(x_0 + \mu) = b_+ \tag{2.19}$$

$$\tag{2.20}$$

are different. For simplicity let us assume $b_- < b_+$ (an analogous proof works for $b_+ < b_-$). Now consider for some $\mu > 0$ the expression

$$\begin{aligned} b_+ &= \lim_{\mu \to 0, with \mu > 0} f(x_0 + 0.1\mu) \\ &= \lim_{\mu \to 0, with \mu > 0} f(\frac{1}{2}(x_0 - \mu) + \frac{1}{2}(x_0 + 1.2\mu)) \\ &\leq \lim_{\mu \to 0, with \mu > 0} (\frac{f(x_0 - \mu) + f(x_0 + 1.2\mu)}{2}) \\ &= \frac{b_- + b_+}{2} \end{aligned}$$

This implies that

$$b_+ \geq b_- \ . \tag{2.21}$$

But this is a contradiction because we had assumed that $b_+ < b_-$. This finishes the proof.

2.5 Differentiation

When you walk up a hill you would like to know how steep it is and as most hills are not uniformly steep, you would like to know how steep the mountain is in every point. In a gross, but useful simplification your path on the mountain can be described by a simple function $y = h(x)$, which I will refer to as the height. Surely a good approximation to steepness is to determine by how much the height changes when we increase x by some amount Δx, ie

$$\frac{h(x + \Delta x) - h(x)}{\Delta x} \ . \tag{2.22}$$

For any given finite Δx this is an approximation to the steepness in the point x. Therefore, the we define

$$h'(x) = \lim_{\Delta x \to 0} \frac{h(x + \Delta x) - h(x)}{\Delta x} \tag{2.23}$$

as the derivative of the function h in the point x. Clearly, a necessary condition for this limit to exist is that $\lim_{\Delta x \to 0} h(x + \Delta x) = h(x)$ which

is just the statement that the function h is continuous in the point x. However, this alone is not enough. Indeed, the differentiability of a function is a much stronger condition than the continuity of a function. This can be seen by the following example of a continuous function that is not differentiable in the point $x = 0$. Lets define the function

$$f(x) = |x| = \begin{cases} f(x) = x & \text{for } x \leq 0 \\ f(x) = -x & \text{for } x \geq 0 \end{cases} \tag{2.24}$$

which is plotted in figure .... Now let us try to differentiate the function.

Figure 2.6: This is a caption

This is easy for $x > 0$ where we simply get

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{x + \Delta x - x}{\Delta x} = 1 \tag{2.25}$$

and it is also easy for $x < 0$ where we simply get

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{-(x + \Delta x) - (-x)}{\Delta x} = -1 \tag{2.26}$$

But now we see the problem, because for $x = 0$ we have a problem because

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \tag{2.27}$$

will not converge, because in any interval around $x = 0$ there are positive and negative values of x for which the expression $\frac{f(x+\Delta x)-f(x)}{\Delta x}$ takes

the values 1 and -1 respectively. So, we do not have convergence and therefore the function cannot be differentiated in this point.

This is an example for a function that is continuous and can be differentiated everywhere except in a single point. There are much weirder functions which are continuous everywhere, but cannot be differentiated anywhere! Their construction, however, is also a bit more involved.

As I have showed you that there are functions that are continuous but are not everywhere differentiable, there are of course also functions that can be differentiated once but they may not be differentiated a second time. A simple example can be constructed from the example above. Namely, the function

$$f(x) = \begin{cases} f(x) = \frac{1}{2}x^2 & \text{for } x \leq 0 \\ f(x) = -\frac{1}{2}x^2 & \text{for } x \geq 0 \end{cases} \tag{2.28}$$

which is plotted in figure .... It is an easy exercise to check that at

Figure 2.7: This is a caption

the point $x = 0$ this function can be differentiated once but not twice. More generally, one can define a function that can be differentiated n times but not $n + 1$ times

$$f(x) = \begin{cases} f(x) = \frac{1}{n!}x^n & \text{for } x \leq 0 \\ f(x) = -\frac{1}{n!}x^n & \text{for } x \geq 0 \end{cases} \tag{2.29}$$

Again, it is an easy exercise to very this statement.

Theorem 42 *(Intermediate value theorem, Rolle's theorem) Be $a > b$ and $f : [a, b] \to \mathbb{R}$ that is differentiable in the interval $]a, b[$. The there exists an $x \in]a, b[$ such that*

$$f'(x) = \frac{f(a) - f(b)}{a - b} \tag{2.30}$$

Proof: Let us first simplify the setting a little bit, by considering the function $g(x) = f(x) - f(b) - \frac{f(a)-f(b)}{a-b}(x - b)$. the function f is differentiable exactly if g is differentiable. For the function g we have $g(a) = g(b) = 0$ and we have that

$$\frac{f(a) - f(b)}{a - b} = f'(x) = g'(x) + \frac{f(a) - f(b)}{a - b} \tag{2.31}$$

is equivalent to

$$g'(x) = 0. \tag{2.32}$$

So, it is enough to show the statement that for $a > b$ and $g : [a, b] \to \mathbb{R}$ that is differentiable in the interval $]a, b[$ and satisfies $g(a) = g(b) = 0$ there exists an $x \in]a, b[$ such that $g'(x) = 0$. The function g is differentiable and therefore continuous. This implies that the function is bounded on the interval $[a, b]$. Therefore the function possesses at least one local extremum and where, by definition, its derivative vanishes. This finishes the proof.

2.5.1 Convexity II and its application to inequalities

In our first section on convexity we had seen that the convexity of a function is closely related to the verification of inequalities that show up in many mathematical and physical problems. However, in these cases we had to know the truth of these inequalities to prove the convexity of a function. However, to make the idea of the convexity of a function useful for proving inequalities, we need independent criteria for proving the convexity of function. Once in the possession of such criteria, we can then turn the logic around and use convex functions to prove inequalities that are really difficult to prove without this tool. To this end we will try to connect differentiability with the convexity of a

function, thereby obtaining a very efficient way to decide the convexity (concavity) of a many functions.

Theorem 43 *Given a function f that it twice differentiable. Then the function is convex on an interval I exactly if $f''(x) > 0$ for all $x \in I$.*

Proof: I will not present this proof as it is not too illuminating and rather long. You can find it in most books on Real Analysis.

With this theorem we can now check very easily the convexity (concavity) of a large number of functions.

1. The function $f(x) = x^{2n}$ is convex for all $n \in \mathbb{N}$. This is verified very easily by differentiation the function twice. As a by-product we have also verified the correctness of the inequality

$$\left(\frac{x+y}{2}\right)^{2n} \leq \frac{x^{2n}+y^{2n}}{2} \tag{2.33}$$

2. The function $f(x) = \ln x$ is concave on the interval $]0, \infty]$. Again this is now almost trivial to check by differentiating the function twice. This also verifies the inequality: For all positive p and q with $\frac{1}{p} + \frac{1}{q} = 1$ and all positive x and y we have

$$x^{1/p} y^{1/q} \leq \frac{x}{p} + \frac{y}{q} \tag{2.34}$$

Proof: We use the concavity of the logarithm together with the monotonicity of the exponential function to see

$$\ln(\frac{1}{p}x + \frac{1}{q}y) \geq \frac{1}{p}\ln x + \frac{1}{q}\ln y \tag{2.35}$$

$$\Rightarrow \quad e^{\ln(\frac{1}{p}x+\frac{1}{q}y)} \geq e^{\frac{1}{p}\ln x+\frac{1}{q}\ln y} \tag{2.36}$$

$$\Rightarrow \quad \frac{1}{p}x + \frac{1}{q}y \geq x^{\frac{1}{p}} y^{\frac{1}{q}} \tag{2.37}$$

3. The function $f(x) = -x\log_2 x - (1-x)\log_2(1-x)$ is concave on the interval $]0, 1[$. This function is also called the entropy function and its concavity is an important physical property that allows its interpretation as a measure of disorder.

Exercises:

1) Consider a function that is convex on the interval $\mathcal{I}$. prove that for any set of positive numbers p_i that satisfy $\sum_{i=1}^{n} p_i = 1$ and for any $x_i \in \mathcal{I}$ we have

$$f(\sum_{i=1}^{n} p_i x_i) \leq \sum_{i=1}^{n} p_i f(x_i). \tag{2.38}$$

2) Prove that every function that is bounded and convex on an open interval $\mathcal{I}$ is continuous on the same interval.

2.5.2 Minimization of convex functions on convex sets.

Quite often in physics you are faced with the task of finding the absolute minimum of some function. The minimization of the potential energy of a system is an example which is important to find the equilibrium position of a particle in a potential. This problem is made difficult when there are many local minima. A convenient feature of a convex function on a simply connected interval is now that it does not have local minima. What is a local minimum? Given a function f we say the function has a relative minimum in the point x_i if there is an $\epsilon_i > 0$ such that for $x \in [x_i - \epsilon, x_i + \epsilon]$ we have $f(x) \geq f(x_0)$. The global minimum is the point x_{glob} such that $f(x_{glob}) \leq f(x)$ for all x.

Assume that for the convex function f defined on a convex set there is a local minimum x_1 for which $f(x_1) > f(x_{glob})$, ie the local minimum at x_1 is not the global minimum. Then from the fact that $f(x_1)$ is a local minimum and the convexity of f we conclude

$$f(x_1) \leq f(\lambda x_1 + (1-\lambda)x_{glob}) \leq \lambda f(x_1) + (1-\lambda) f(x_{glob}) \tag{2.39}$$

and for $\lambda \leq 1$ we find

$$f(x_1) \leq f(x_{glob}) \tag{2.40}$$

which is in contradiction to the assumption that $f(x_1) > f(x_{glob})$. Therefore, the function cannot have any local minimum that is smaller than the global minimum.

For convex function there is therefore a very simple method for finding the global minimum of the function which is called method of steepest descent. It works in the following way. Start in a randomly chosen point x_0. Determine the gradient of f in that point, ie $\frac{df}{dx}(x = x_0)$. Then determine the next point via $x_{n+1} = x_n - \Delta\frac{df}{dx}(x = x_n)$. If we have $f(x_{n+1}) < f(x_n)$ then continue. If not, then half the size of Δ, ie $\Delta \to \Delta/2$ and try again.

This method will stop in some relative minimum, namely when $\frac{df}{dx}(x) = 0$. If the function is convex, then such a relative minimum is a global minimum and we have succeeded.

Note, that I have described the procedure for a function of a single variable. However, it straightforwardly generalizes to functions of many variables.

2.5.3 Newton's method

Now let us briefly come back to a topic that we had considered earlier, namely finding the square roots of the number a employing a sequence of numbers of the form

$$x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n}) \tag{2.41}$$

How did we come to this sequence in the first place. Well, it is a special case of a more general idea which goes back to Newton apparently and is an application of differentiability and works particularly well on convex functions. Newton asked himself the following question: "If I have a function $f(x)$ how can I find its zeros, ie those x that satisfy $f(x) = 0$?"

One possible way to arrive at a general procedure that will do this is the following argument. Given the function $f(x)$ we can expand it in a Taylor series which we break off after the first term.

$$f(x_{n+1}) \cong f(x_n) + (x_{n+1} - x_n)f'(x_n) \tag{2.42}$$

Now set the left hand side to 0, ie assume that x_{n+1} is a solution to $f(x) = 0$. Then solving for x_{n+1} we get

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{2.43}$$

Why could that work at all? If the function $f(x)$ is linear, ie $f(x) = ax$, then the above truncated Taylor expansion is actually exact. As a consequence, the iteration converges in a single step because $x_{n+1} = 0$. So we see that we have got a decent method for linear functions, which is of course not terribly interesting. However, now you can have some trust in the idea that if we are not too far from the solution of $f(x) = 0$ then the truncated Taylor series is a very good approximation and so the sequence of numbers that we generate from it can be useful. Furthermore, if you look at the Taylor expansion of a function around its zero you immediately see that in a close neighborhood of the zero the function is to a very good approximation linear because

$$f(x+\epsilon) = f(x)+\epsilon f'(x)+\frac{1}{2}\epsilon^2 f''(x)+\ldots = \epsilon f'(x)+\frac{1}{2}\epsilon^2 f''(x)+\ldots \quad (2.44)$$

Of course this is not any form of proof but it least it suggest that one should look into this in a bit more detail. I will not show you the proof with all the bells and whistles but will bring you a 'hand-waving' argument because the full prove for the convergence of the method is rather lengthy. What I am going to show you now is what I dreamed up on a sheet of paper rather quickly and illustrates how one proceeds as a theoretical physicist. Let me assume that I am already quite close to the true zero x which satisfies $f(x) = 0$. Furthermore, let me assume that

$$\frac{f''(x)}{f'(x)} = C < \infty \quad (2.45)$$

Then chose $\epsilon < \frac{1}{C}$. Now look at Newtons method starting with $x_n = x + \epsilon$ and employing the Taylor expansion

$$\begin{aligned} x_{n+1} &= x + \epsilon - \frac{f(x+\epsilon)}{f'(x+\epsilon)} && (2.46)\\ &\cong x + \epsilon - \frac{\epsilon f'(x) + \frac{1}{2}\epsilon^2 f''(x)}{f'(x) + \epsilon f''(x)} && (2.47)\\ &\cong x + \epsilon - \epsilon\frac{1 + \frac{1}{2}\epsilon\frac{f''(x)}{f'(x)}}{1 + \epsilon\frac{f''(x)}{f'(x)}} && (2.48)\\ &\cong x + \epsilon - \epsilon(1 + \frac{1}{2}\epsilon\frac{f''(x)}{f'(x)})(1 - \epsilon\frac{f''(x)}{f'(x)}) && (2.49) \end{aligned}$$

$$\cong \quad x - \frac{1}{2}\epsilon^2 \frac{f''(x)}{f'(x)} \tag{2.50}$$

Now we can see that the deviation from the true solution has decreased from ϵ to some quantity that is of the order of ϵ^2. If ϵ is very small then this means that we have come a lot closer to the true solution of $f(x) = 0$. Now wild hand-waving we have worked our way to reasonable amounts of hand-waving and we can be quite confident that there is will be a theorem that one can prove strictly and which ensures the convergence of the Newton method under certain assumptions. This is indeed the case and one example is

Theorem 44 *Let $f : [a, b] \to \mathbb{R}$ a twice differentiable convex function with $f(a) < 0$ and $f(b) > 0$. Let $x_0 \in [a, b]$ such that $f(x_0) > 0$ then the Newton sequence*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{2.51}$$

is monotonically falling and converges against the solution of $f(x) = 0$.

Proof: The full and strict proof of this theorem is about two pages long, so I will leave it out. Again it can be found in most books on Real Analysis.

Examples and Exercises:

- Verify that the function $f(x) = -\sin x$ on the interval $[0, \pi]$ satisfies the criteria of the above theorem. Write down Newtons iteration formula. Estimate the convergence rate, ie if you are an ϵ away from a solution show what power of ϵ the next value is away from the true zero.

- Let k be a natural number. Show that $x = \tan x$ has exactly one solution in the interval $](k - \frac{1}{2})\pi, (k + \frac{1}{2})\pi[$ which we call ξ. Prove that the sequence

$$x_0 = (k + \frac{1}{2})\pi \tag{2.52}$$

$$x_{n+1} = k\pi + arctanx_n \tag{2.53}$$

converges against ξ. Determine the ξ to a precision of 10^{-6} for $k = 1, 2, 3$.

2.6 Integration

In this section we will revisit the concept of integration which you have encountered already in the first term in the form of the inverse of differentiation. Here I will show you a different approach, namely the concept of Riemann integral. While this will lead us to the same answers it has two distinct advantages. Firstly, this approach is closely related to the concept of area, which allows its generalization to the integration of functions of more than just one variable which is pretty important in many areas of physics. Secondly, it will turn out that some functions are not integrable in the standard sense. However, the approach via Riemann integration actually motivates a more general approach to integration which is called Lebesgue integration.

2.6.1 Riemann integration

When Riemann thought about integration, he actually looked at it from a rather practical point of view. Let us draw the graph of a positive continuous function f on an interval $[a, b]$. Now there is a region that is delimited from above by the graph of the function, from below by the interval $[a, b]$ and on the sides by the vertical lines from $(x, y) = (a, f(a))$ and from $(x, y) = (b, f(b))$. This is shown in figure 2.8.

Figure 2.8: This is a caption

The region delimited in the way described above, will have a certain

area A and the idea is to denote this area as

$$A = \int_a^b f(x)dx. \tag{2.54}$$

Why could Riemann possibly think that this is a sensible approach in the first place? In particular it is not clear at all that this ties in nicely with the idea that the integral of a function f is just given by a function F such that $\frac{dF}{dx}(x) = f(x)$? You can gain some understanding for this by looking at some simple cases that you know from geometry. Let us consider first a very simple function, namely the constant function

$$f(x) = c > 0 \tag{2.55}$$

and draw this function on the interval $[a, b]$. Clearly, the area between the graph of the function and the x-axis is just a rectangle of width $b-a$ and height c, so that its area is give by $c(b-a)$. Its not quite clear how that connects to differentiation, but we can see that easily, when we consider the more general situation in which we draw the function on the interval $[a, x]$ where x is a variable. Then the volume of the area between the graph of the curve and the x-axis is given by

$$F(x) = c(x - a). \tag{2.56}$$

It is now straightforward to verify, that $\frac{dF}{dx}(x) = c = f(x)$. Therefore we can see that in this simple example the two concepts do indeed coincide.

However, what do you do if I give you the example $f(x) = \sin x$ on the interval $[0, \pi/2]$. How would you prove that indeed the area under the graph of the curve and the x-axis is given by $\cos 0 = 1$? As usual the best approach is to reduce the problem to something that we understand very well. In this case, this is the area under a constant function which we have studied above. Instead of a completely constant function let us consider functions that are called piecewise constant, ie a function $f(x)$ and a series of values $x_0 < x_1 < \ldots < x_n$ such that it takes the fixed value c_i on the open interval $]x_i, x_{i+1}[$. An example is given in figure 2.9. Quite evidently, it makes sense to say that the area under this function on the interval $[x_0, x_n]$ is given by

$$I = \sum_{k=0}^{n-1} c_k(x_{k+1} - x_k). \tag{2.57}$$

Figure 2.9: This is a caption

Obviously, only very few functions are stepwise constant and therefore we have to go a step further. Given that we wish to define the integral of a function on an interval as the area between the graph and the x-axis, what we really need to do is to approximate this area. How could we do this? One possibility is to find upper bounds and lower bounds which we successively improve such that eventually they coincide. We will achieve exactly this employing stepwise constant functions.

Consider a function $f(x)$ that is defined on the interval $[a, b]$ and that is bounded both from above and below on that interval, ie there are number c and C such that $c \leq f(x) \leq C$ for all $x \in [a, b]$. Then we can clearly say that the area A under the graph of $f(x)$ is smaller than $C(b-a)$ and is larger than $c(b-a)$. Therefore, we have found an upper and lower bound on the area A. Of course, these two bounds will not normally coincide, unless the function $f(x)$ is constant. How can we improve on these bounds? One possible approach would be one that you have encountered already on various occasions. Why don't we split the interval $[a, b]$ into two halves, namely

$$\mathcal{I}_1 = [a, \frac{a+b}{2}] \text{ and } \mathcal{I}_2 = [\frac{a+b}{2}, b] \tag{2.58}$$

For each of these intervals we determine the maximum and the minimum of the function $f(x)$. For the interval $\mathcal{I}_1$ these are c_1 and C_1. For the interval $\mathcal{I}_2$ these are c_2 and C_2. Then we have new upper bounds

$$c_1 \frac{b-a}{2} + c_2 \frac{b-a}{2} \leq A \leq C_1 \frac{b-a}{2} + C_2 \frac{b-a}{2} \tag{2.59}$$

or equivalently

$$\frac{c_1 + c_2}{2}(b-a) \leq A \leq \frac{C_1 + C_2}{2}(b-a)\,. \tag{2.60}$$

Now we can continue this procedure, by subdividing both intervals into two equal halves. After n successive divisions we end up with 2^n intervals, each of length $\frac{b-a}{2^n}$. On each of these intervals we determine the largest and the smallest value of the function $f(x)$. For the $i-th$ interval these are denoted c_i and C_i so that we end up with the bounds

$$A_n^{low} = \frac{\sum_{i=1}^{2^n} c_i}{2^n}(b-a) \leq A \leq \frac{\sum_{i=1}^{2^n} c_i}{2^n}(b-a) = A_n^{up}\,. \tag{2.61}$$

In this way the sequence of successive upper bounds A_n^{up} is non-increasing while the sequence of successive lower bounds A_n^{low} is non-decreasing. As both sequences are bounded, they must both converge. Let us denote the limiting values by

$$\begin{aligned} \lim_{n\to\infty} A_n^{up} &= \int_a^{*b} f(x)dx, \\ \lim_{n\to\infty} A_n^{low} &= \int_{*a}^{b} f(x)dx. \end{aligned}$$

These two limits are also called upper and lower integral. The big question is now, whether they will actually converge to the same value or not.

Definition: We call a function $f(x)$ integrable on the interval $[a,b]$ if the upper and lower bounds A_n^{up} and A_n^{low} converge to the same value, ie

$$\int_a^{*b} f(x)dx = \int_{*a}^{b} f(x)dx$$

In this case we define the definite integral of $f(x)$ on the interval $[a,b]$ by

$$\int_a^{*b} f(x)dx = \int_{*a}^{b} f(x)dx \equiv \int_a^b f(x)dx$$

which equals the area between the graph of the curve and the x-axis.

Example I: Let us consider the function $f(x) = x$ on the interval $[0, 1]$. Let us assume that we have split the interval into 2^n pieces. Then for any interval $[x_k, x_{k+1}]$ we have that the largest value of the function is x_{k+1} while the smallest value is x_k. Therefore we obtain for the area

$$\sum_{k=0}^{2^n-1} \frac{k}{2^n}\frac{1}{2^n} \leq A \leq \sum_{k=1}^{2^n} \frac{k}{2^n}\frac{1}{2^n} \tag{2.62}$$

from which it follows that

$$\frac{1}{2}\frac{2^n-1}{2^n} \leq A \leq \frac{1}{2}\frac{2^n+1}{2^n}. \tag{2.63}$$

Taking the limit $n \to \infty$ on both sides, we find

$$\frac{1}{2} \leq A \leq \frac{1}{2}, \tag{2.64}$$

so that upper and lower bound converge to the same value and we can write

$$\int_0^1 x dx = \frac{1}{2}. \tag{2.65}$$

Example II: You may wonder whether there are functions that are not integrable. Consider the example

$$f(x) = \begin{cases} 1 & \text{x = rational} \\ 0 & \text{x = irrational} \end{cases} . \tag{2.66}$$

Clearly, on any interval, no matter how small, this function is bounded from above by 1 and from below by 0 and you cannot improve on these bounds. Therefore, we find that

$$\int_a^{*b} f(x)dx = 1 > 0 = \int_{*a}^b f(x)dx.$$

Therefore, the upper and lower integrals are different and therefore this function is not integrable in the sense of Riemann. On the other hand

it seems quite reasonable to attribute the value zero to this integral because the function that we are integrating is practically always zero. In fact, if you chose a number randomly, then it will always be irrational and as you evaluate the function, you will find it to be zero. What you need to deal with this question is a generalization of Riemanns approach, that is not only dealing with a decomposition of the integration interval into simple intervals, but into more general sets. Then you have to develop a way to measure the size of the set. This is done in an approach that is called measure theory and is indeed the basic building block in the more general framework developed by Lebesgue. Indeed, as a Lebesque integral the function considered above becomes integrable and the integral takes the value 0. Fortunately, for many applications in physics you do not need this more advanced theory of integration. However, if you intend to study quantum field theory, then it may be useful.

On the other hand there are large classes of functions that are integrable.

Theorem: Every continuous function on a closed interval $[a, b]$ is integrable.

Proof: For a continuous function we have that $\lim_{x\to x_0} f(x) = f(x_0)$. This has to be true for any point $x_0 \in [a, b]$. Now we realize that this means that the function cannot vary too wildly. Indeed, consider that we split the interval $[a, b]$ into 2^n equally sized intervals, via repeated halving of the interval. Then for given $\epsilon > 0$ there is always an n such that in *each* of these intervals $maxf(x) - minf(x) \leq \epsilon$. If this were not the case, then there would be a point x_0 which belongs to a sequence of ever smaller intervals and in each of these intervals the function $f(x)$ would vary by more than ϵ. That would imply that $\lim_{x\to x_0} f(x)$ would not exist. This is in contradiction to the continuity of the function.

If we now compute the sequences of upper and lower bounds A_n^{up} and A_n^{low}, we observe that for every $\epsilon > 0$ there is an n such that $A_n^{up} - A_n^{low} \leq \epsilon(b - a)$. Therefore,

$$\int_a^{*b} f(x)dx - \int_{*a}^{b} f(x)dx = \lim_{n\to\infty} (A_n^{up} - A_n^{low}) = 0 \tag{2.67}$$

which implies integrability and completes the proof.

There are also discontinuous functions that are integrable. For example any monotonically increasing bounded function is integrable. Proof = Exercise.

Now let us establish the connection between Riemann integrals and differentiability. To this end, we need to consider the integral of a continuous function $f(x)$ over an interval $[a, x]$ where x is variable, ie we consider the function

$$F(x) = \int_a^x f(x)dx. \tag{2.68}$$

Now let us differentiate this function, ie consider

$$\frac{dF}{dx}(x_0) = \lim_{\Delta x \to 0} \frac{F(x_0 + \Delta x) - F(x_0)}{\Delta x} \quad = \quad \lim_{\Delta x \to 0} \frac{1}{\Delta x} \int_{x_0}^{x_0+\Delta x} f(x)dx.$$

Now because the function f is continuous we have that for every $\epsilon > 0$ there is a Δx such that for all $x \in [x_0, x_0 + \Delta x]$ we have that $|f(x) - f(x_0)| \leq \epsilon$. Then,

$$\Delta x(f(x_0) - \epsilon) \leq \int_{x_0}^{x_0+\Delta x} f(x)dx \leq \Delta x(f(x_0) + \epsilon) \tag{2.69}$$

and as a consequence

$$\frac{dF}{dx}(x_0) = f(x_0).$$

This establishes the connection between Riemann integration and differentiation and constitutes the so-called main theorem of calculus. From this definition we also obtain

$$\int_a^b f(x)dx = F(b) - F(a). \tag{2.70}$$

To see that it is really important to assume that the function f is continuous. Let us consider the function

$$f(x) = \begin{cases} 0 & \text{for } 0 \leq x < 1 \\ 1 & \text{for } 1 \leq x \leq 2 \end{cases} \tag{2.71}$$

Now we can clearly compute the function

$$F(x) = \int_0^x f(x)dx = \begin{cases} 0 & \text{for } 0 \leq x < 1 \\ x - 1 & \text{for } 1 \leq x \leq 2 \end{cases} \tag{2.72}$$

Can you differentiate this function. Certainly, you will have a problem at $x = 1$ as

$$\lim_{\Delta x \to 0, \Delta x > 0} \frac{F(1 + \Delta x) - F(1)}{\Delta x} = 1 \tag{2.73}$$

$$\lim_{\Delta x \to 0, \Delta x < 0} \frac{F(1 + \Delta x) - F(1)}{\Delta x} = 0\,. \tag{2.74}$$

So, obviously they are not the same.

Now let us consider two situations which have not been included in the above considerations but which nevertheless are very important in practice. In the discussions above I always considered the situation in which we wish to integrate a function in a closed and finite interval such as $[a, b]$. However, there may very well be situations in which we wish to integrate over an infinitely large interval, e.g. $[a, \infty[$ or even $]-\infty, \infty[$. How could we define such integrals? A quite natural way is to consider these integrals as limits of definite integrals, ie

$$\int_a^\infty f(x)dx = \lim_{R \to \infty} \int_a^R f(x)dx, \tag{2.75}$$

$$\int_{-\infty}^\infty f(x)dx = \lim_{R,S \to \infty} \int_{-R}^S f(x)dx \tag{2.76}$$

where in the second integral the limiting processes have to be carried out independently. If the limit only exists under the assumption $R = S$, then

$$P\!-\!\int_{-\infty}^\infty f(x)dx = \lim_{R \to \infty} \int_{-R}^R f(x)dx \tag{2.77}$$

where the $P-$ in front of the integral indicates the restricted limiting process and is called principal value.

Whether the left hand side of this equation makes any sense then simply depends on the existence of the limit on the right hand side.

Quite clearly, for the existence of this limit it will not be sufficient to have a continuous function as the simple example $f(x) = c > 0$ shows. For such a function the limit does not exist as it diverges to ∞. Therefore, the function that you integrate must go to zero sufficiently rapidly when its argument grows to infinity.

There is another situation in which we will have to define the value of our integrals as a limit. This situation can arise when we wish to integrate a function $f(x)$ on an interval $[a, b]$ on which the function is not bounded. An example would be $f(x) = \frac{1}{\sqrt{x}}$ on the interval $]0, 1]$. This is not a closed interval because the function is not defined in the point $x = 0$. Again, if we wish to make sense of this integral, the simplest approach is by taking limits, namely by stating

$$\int_a^b f(x)dx = \lim_{R\to b, R<b} \int_a^R f(x)dx, \tag{2.78}$$

$$\int_a^b f(x)dx = \lim_{R\to a, R>a} \int_R^b f(x)dx. \tag{2.79}$$

For the example $f(x) = 1/\sqrt{x}$ we can then obtain

$$\int_0^1 \frac{1}{\sqrt{x}}dx = \lim_{R\to 0, R>0} \int_R^1 \frac{1}{\sqrt{x}}dx = \lim_{R\to 0, R>0} (2\sqrt{1} - 2\sqrt{R}) = 2. \tag{2.80}$$

From now on, when we write an integral which is either over an infinite interval or which is over a function that is not defined at the end points, we actually mean to write limits as above. Indeed, there is one further possibility, namely a function that diverges inside the integration interval. An example is given by

$$f(x) = \frac{1}{\sqrt{|x|}} \tag{2.81}$$

over the interval $[-1, 1]$. In that case we define

$$\int_{-1}^1 f(x)dx = \lim_{R\to 0, R>0} [\int_{-1}^{-R} f(x)dx + \int_R^1 f(x)dx]. \tag{2.82}$$

2.6.2 The integral comparison criterion

Now let me use what we have learnt here to derive a very useful method to determine the convergence of some infinite series that is otherwise extremely difficult to decide. In an example I had shown you how to work out the integral of the function $f(x) = x$. The task was simplified considerably by the fact that $f(x) = x$ is monotonically growing. As a consequence we could determine the largest and smallest value in each interval very easily. They were assumed by the function at the end points of the intervals. Let us see whether we can make use of the monotonicity for general functions. As we would like to learn something about infinite series, we will need to consider functions that are defined on the interval $[1, \infty[$ and which are monotonically decreasing on that interval as we could otherwise not hope for convergence in the first place. We also need to assume that the function only takes positive values. Now we know that we can always find a lower and an upper bound on the integral of f on that interval. We simply divide the interval $[1, \infty[$ into intervals of unit length $[n, n+1]$ where n runs through the natural numbers. Then we find

$$\sum_{n=1}^{R-1} f(n) \geq \int_1^R f(x)dx \geq \sum_{n=1}^{R-1} f(n+1). \tag{2.83}$$

If the limit of the integral for $R \to \infty$ exists, then the series on the right hand side is bounded from above and because it is growing monotonically it converges as well. If on the other hand we know for some reason that the series converges in this limit then we know that the integral is bounded by the left hand side of the inequality and therefore it exists. Therefore we have the following

Integral comparison criterion: Given a function $f : [0, \infty[\to \mathbb{R}_+$ that is monotonically decreasing, then the series $\sum_{n=1}^{\infty} f(n)$ converges exactly if $\int_1^{\infty} f(x)dx$ exists. If the integral exists then we have

$$\sum_{n=1}^{\infty} f(n) \geq \int_1^{\infty} f(x)dx \geq \sum_{n=2}^{\infty} f(n). \tag{2.84}$$

This is a very useful criterion indeed, as the following examples will show you.

1. The series
$$\sum_{n=1}^{\infty} \frac{1}{n} \tag{2.85}$$
diverges. Of course you know this already because I have shown you earlier in the lecture. But let us see this by considering the corresponding integral
$$\lim_{R\to\infty} \int_1^R \frac{1}{x} dx. \tag{2.86}$$
This can be evaluated realizing that
$$\frac{d}{dx} lnx = \frac{1}{x} \tag{2.87}$$
so that
$$\lim_{R\to\infty} \int_1^R \frac{1}{x} dx = \lim_{R\to\infty} lnR = \infty. \tag{2.88}$$
From the fact that the integral does not exist we can, for the integral comparison theorem, conclude that the corresponding series does not converge.

2. Now consider the series
$$\sum_{n=2}^{\infty} \frac{1}{nlnn} \tag{2.89}$$
to show that it also diverges. This can be seen by considering the corresponding integral
$$\lim_{R\to\infty} \int_2^R \frac{1}{xlnx} dx. \tag{2.90}$$
This doesn't look much easier until we realize that by the chain rule
$$\frac{d}{dx} ln(lnx) = \frac{1}{lnx} \frac{dlnx}{dx} = \frac{1}{xlnx} \tag{2.91}$$
so that
$$\lim_{R\to\infty} \int_2^R \frac{1}{xlnx} dx = \lim_{R\to\infty} ln(lnR) = \infty. \tag{2.92}$$

From the fact that the integral does not exist we can, for the integral comparison theorem, conclude that the corresponding series does not converge.

3. Now consider

$$\sum_{n=2}^{\infty} \frac{1}{n(lnn)^2} \tag{2.93}$$

to show it converges. This can be seen by considering the corresponding integral

$$\lim_{R\to\infty} \int_2^R \frac{1}{x(lnx)^2} dx. \tag{2.94}$$

Now realize that by the chain rule

$$\frac{d}{dx}\frac{1}{lnx} = -\frac{1}{(lnx)^2}\frac{dlnx}{dx} = -\frac{1}{x(lnx)^2} \tag{2.95}$$

so that

$$\lim_{R\to\infty} \int_2^R \frac{1}{x(lnx)^2} dx = \lim_{R\to\infty} (\frac{1}{2(ln2)^2} - \frac{1}{R(lnR)^2}) = \frac{1}{2(ln2)^2}. \tag{2.96}$$

From the fact that the integral exists we can, by the integral comparison theorem, conclude that the corresponding series converges.

Imagine someone would have asked you to prove convergence or divergence of the series without this criterion.

2.6.3 Interchanging Limites

In the lectures so far you have learned a lot about various limites. They showed up in sequences, series, in the definitions of real numbers, continuity, differentiability and the integrability of functions. Sometimes you may encounter two limiting procedures at the same time. An example were integrals with infinite boundaries which we defined for example as

$$\int_0^\infty f(x)dx := \lim_{R\to\infty} \int_0^R f(x)dx\,. \tag{2.97}$$

Therefore we first evaluate an integral over a finite interval, which in itself involves a limiting process, and subsequently take the limit $R \to \infty$. Here it is clearly stated in which order to take the limit, but could one perhaps interchange the limits? This is a question that shows up in many different contexts and often the order of the limites is important. In mathematics it will have to be defined while in physics it may be imposed by the experimental situation. In the following I want to show you some examples in which it is dangerous to interchange the limites and I also would like to give you conditions when it is perfectly acceptable to interchange the limites, ie the results do not depend on the order in which the limit is taken.

Let us consider a function of a single variable and see whether problems can arise. Take a really simple example and consider on the interval $[0, 1]$ the function

$$f_n(x) = \begin{cases} nx^{n-1} & \text{for } 0 \le x < 1 \\ 0 & \text{for } x = 1 \end{cases} \tag{2.98}$$

Now we find that

$$\lim_{n\to\infty} f_n(x) = 0\,. \tag{2.99}$$

So quite clearly the function converges for every value of the argument x to zero. Now let us evaluate the integral of the function $f_n(x)$. We find

$$\int_0^1 f_n(x)dx = 1\,. \tag{2.100}$$

Now taking the limit we find

$$\lim_{n\to\infty} \int_0^1 f_n(x)dx = 1 \neq 0 = \int_0^1 0dx = \int_0^1 \lim_{n\to\infty} f_n(x)dx\,. \tag{2.101}$$

So, as a consequence we learn that for such a harmless function as $f_n(x)$ we see that we cannot interchange the limit in n with the integration.

So what is going on here? When we look at the function $f_n(x)$ as a whole for various values of n then we see that very near $x = 1$ then function becomes larger and larger with increasing n, but that the region where the function is large becomes narrower and narrower. In the integration we therefore have an area that becomes narrower and higher in such a way that it always gives a finite contribution. Indeed, in the limit $n \to \infty$ we have an area that corresponds to $0 \cdot \infty$ and as you learnt in the earlier lectures this is a tricky quantity.

In summary you observe the interesting phenomenon that although for every fixed x the function $f_n(x)$ converges there is an increasingly narrow region for which the function becomes arbitrarily large and it is exactly this behaviour that leads to problems in the interchanging of the limit with the integration. This is the motivation to introduce a stronger notion of convergence than the previously considered notion of pointwise convergence, ie convergence for a fixed x. Indeed, what we would like to have is that the function converges uniformly. Let me give a formal definition.

Definition 45 *A sequence of function $f_n(x)$ defined on an interval $\mathcal{I}$ converges uniformly to a function $f(x)$ if*

$$\lim_{n\to\infty} sup_{x\in\mathcal{I}}\{|f_n(x) - f(x)|\} = 0 \tag{2.102}$$

ie if for all $\epsilon > 0$ there is an N such that for all $n \geq N$ and all $x \in \mathcal{I}$ we have $|f_n(x) - f(x)| \leq \epsilon$.

Here $sup_{x\in\mathcal{I}} f(x)$ denotes the largest value of $f(x)$ on the interval $\mathcal{I}$.

The great thing about uniform convergence is that once you have established it, then you can be relatively relaxed about interchanging the limites. The following statements hold true

Theorem 46 *Assume that in the interval $[a, b]$ the sequence $f_n(x)$ converges uniformly to $f(x)$ and assume that the functions $f_n(x)$ are integrable for any n, then $f(x)$ is also integrable and we have*

$$\int_a^b \lim_{n\to\infty} f_n(x)dx = \lim_{n\to\infty} \int_a^b f_n(x)dx \tag{2.103}$$

Proof: Consider the integral $\int_a^b f(x)dx$ and form upper and lower bounds on it. We do this by splitting the interval $[a,b]$ into 2^k equally long intervals of the form $\mathcal{I}_l = [a+(b-a)l/2^k, a+(b-a)(l+1)/2^k]$.

$$\sum_{l=0}^{2^k-1} min_{x\in\mathcal{I}_l} f(x)\frac{b-a}{2^k} \leq \int_a^b f(x)dx \leq \sum_{l=0}^{2^k-1} max_{x\in\mathcal{I}_l} f(x)\frac{b-a}{2^k}$$

From the uniform convergence of $f_n(x)$ to $f(x)$ we know that for all $\epsilon > 0$ there is an N such that for all $n > N$ and all x we have $|f_n(x) - f(x)| < \epsilon$. Therefore we find

$$\sum_{l=0}^{2^k-1} min_{x\in\mathcal{I}_l}(f_n(x)-\epsilon)\frac{b-a}{2^k} \leq \sum_{l=0}^{2^k-1} min_{x\in\mathcal{I}_l} f(x)\frac{b-a}{2^k}$$
$$\sum_{l=0}^{2^k-1} min_{x\in\mathcal{I}_l} f(x)\frac{b-a}{2^k} \leq \sum_{l=0}^{2^k-1} min_{x\in\mathcal{I}_l}(f_n(x)+\epsilon)\frac{b-a}{2^k}$$

Now we can take the limit $k\to\infty$ to find

$$\int_a^b f_n(x)dx - \epsilon(b-a) \leq \int_a^b f(x)dx \leq \int_a^b f_n(x)dx + \epsilon(b-a)$$

Now we know from uniform convergence again that in the limit $n\to\infty$ the value of ϵ can be made arbitrarily small so that we find

$$\lim_{n\to\infty}\int_a^b f_n(x)dx \leq \int_a^b f(x)dx \leq \lim_{n\to\infty}\int_a^b f_n(x)dx$$

which implies

$$\lim_{n\to\infty}\int_a^b f_n(x)dx = \int_a^b f(x)dx$$

which was the statement that we wanted to prove.

Therefore we can interchange the limit with the integration in this case. Let us now consider when we can interchange the limits with respect to n and x.

$$f_n(x) = \frac{x^2}{x^2 + (\frac{1}{n})^2} \tag{2.104}$$

We find

$$\lim_{x\to 0}(\lim_{n\to\infty}\frac{x^2}{x^2+(\frac{1}{n})^2}) = \lim_{x\to 0} 1 = 1 \tag{2.105}$$

So, here the limit exists. If we take the limit the other way around the we find

$$\lim_{n\to\infty}(\lim_{x\to 0}\frac{x^2}{x^2+(\frac{1}{n})^2}) = \lim_{n\to\infty} 0 = 0 \tag{2.106}$$

So, again, we make the very important observation that the limites cannot necessarily be interchanged. Of course there may be compelling physical reasons for choosing a particular ordering for the limits but in the absence of such reasons, one has to take great care with multiple limits. Again the problem lies in the lack of uniform convergence. Indeed, we have

Theorem 47 *Given functions $f_n(x)$ defined on the interval $\mathcal{I}$ assume that $f_n(x)$ converges uniformly to $f(x)$ and assume that $\lim_{x\to a} f_n(x) = c_n$ exists for all n, then also $\lim_{n\to\infty} c_n$ and $\lim_{x\to a} f(x)$ exist and both limits are equal, or in other words*

$$\lim_{x\to a}\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty}\lim_{x\to a} f_n(x) \tag{2.107}$$

Proof: The proof is an exercise.

There is another rather important combination of operations that involve limiting processes. These are differentiation and integration. Often it turns out to be useful to introduce an extra parameter in your integral and differentiate with respect to it. Or perhaps the physical situation that you have depends on various parameters. If for example your function is of the form $f(x,t) = tx$ where t is for example time, then we have that

$$\int_0^1 f(x,t)dx = \frac{t}{2}\,. \tag{2.108}$$

Now you may want to differentiate with respect to this extra parameter t, ie

$$\frac{d}{dt}\int_0^1 f(x,t)dx = \frac{1}{2} = \int_0^1 \frac{d}{dt}f(x,t)dx\,. \tag{2.109}$$

So, in this case you can interchange the order of differentiation and integration. But in general this is not the case. Lets see whether we can guess what is needed. Let us assume a function of two variables $f(x,t)$ and define $F(t) = \int_a^b f(x,t)dx$. Then we have

$$\begin{aligned}
\frac{dF}{dt}(t) &= \lim_{\Delta t\to\infty} \frac{F(t+\Delta t) - F(t)}{\Delta t} && (2.110)\\
&= \lim_{\Delta t\to\infty} \frac{1}{\Delta t}\left(\int_a^b f(x,t+\Delta t)dx - \int_a^b f(x,t)dx\right) && (2.111)\\
&= \lim_{\Delta t\to\infty} \int_a^b \frac{f(x,t+\Delta t)dx - f(x,t)}{\Delta t}dx && (2.112)\\
&\doteq \int_a^b \lim_{\Delta t\to\infty} \frac{f(x,t+\Delta t)dx - f(x,t)}{\Delta t}dx && (2.113)\\
&\doteq \int_a^b \frac{df(x,t)}{dt}dx && (2.114)
\end{aligned}$$

To be able to say that we have equality all the way through, then we need to have that the limit can be taken under the integral and that the function $f(x,t)$ can actually be differentiated in the parameter t and finally the result should be an integrable function in x (these hoped for equalities are indicated by $\doteq$). Indeed, one can show the following

Theorem 48 *(Leibniz rule) If* $\frac{df(x,t)}{dt}$ *exists and* $f(x,t)$ *as well as* $\frac{df(x,t)}{dt}$ *are continuous in both* x *and* t*, then we have*

$$\frac{d}{dt}\int_a^b f(x,t)dx = \int_a^b \frac{df(x,t)}{dt}dx \qquad (2.115)$$

Example: It is well known that $\int_{-\infty}^{\infty} e^{-x^2} = \sqrt{\pi}$ Consider the integral

$$F_n = \int_{-\infty}^{\infty} x^{2n} e^{-x^2} dx\,. \qquad (2.116)$$

One useful way of evaluating this integral is by introducing an extra parameter t such that

$$F_n(t) = \int_{-\infty}^{\infty} x^{2n} e^{-tx^2} dx\,. \qquad (2.117)$$

Now we can differentiate this function n times with respect to t to find using Leibniz' rule

$$F_n(t) = (-1)^n \frac{d^n}{dt^n} \int_{-\infty}^{\infty} e^{-tx^2} dx = (-1)^n \sqrt{\pi} \frac{d^n}{dt^n} t^{-\frac{1}{2}} \tag{2.118}$$

As a consequence we find that

$$F_n(t) = \sqrt{\pi} (\prod_{k=1}^{n} \frac{2k-1}{2}) t^{-(2n+1)/2} \tag{2.119}$$

and then

$$\int_{-\infty}^{\infty} x^{2n} e^{-x^2} dx = \lim_{t \to \infty} F_n(t) = \sqrt{\pi} (\prod_{k=1}^{n} \frac{2k-1}{2}) \tag{2.120}$$

This is certainly a lot easier than doing integration by parts n times and illustrates that it may be useful to insert some parameter into an integral as long as one pays attention to the possibility that one may not be able to interchange limites in some cases.

Chapter 3

Vectors and Matrices

In the first part of this section I will briefly recap basic notions of vector and matrices. I will then put them to work when we are talking about markov processes, entropy and disorder.

3.1 Vectors

In the first term you got to know column vectors with two components such as

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{3.1}$$

with some number x_1 and x_2. Of course this can also be generalized to any number of components, ie to column vectors with n components such as

$$\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} . \tag{3.2}$$

The laws for the addition and of two vectors and for the multiplication of a vector by a number are analogous to those for vectors with 2 components, ie

$$\vec{x} + \vec{y} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix} \tag{3.3}$$

and

$$\lambda\vec{x} = \lambda \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \lambda x_1 \\ \vdots \\ \lambda x_n \end{pmatrix} . \tag{3.4}$$

With these rules you have the usual properties for vectors.

1. We call V a set of vectors and $\vec{a}$ is a vector if it is element of V. The we have

   (a) $\forall \vec{a}, \vec{b} \in V \quad \Rightarrow \vec{a} + \vec{b} \in V.$ (closure)

   (b) $\forall \vec{a}, \vec{b}, \vec{c} \in V \Rightarrow \vec{a} + (\vec{b} + \vec{c}) = (\vec{a} + \vec{b}) + \vec{c}.$ (associative)

   (c) $\exists \vec{0} \in V$ so that $\forall \vec{a} \in V \Rightarrow \vec{a} + \vec{0} = \vec{a}.$ (zero)

   (d) $\forall \vec{a} \in V : \exists(-\vec{a}) \in V$ so that $\vec{a} + (-\vec{a}) = \vec{0}.$ (inverse)

   (e) $\forall \vec{a}, \vec{b} \in V \Rightarrow \vec{a} + \vec{b} = \vec{b} + \vec{a}.$ (Abelian)

2. The Scalar multiplication satisfies

   (a) $\forall \alpha \in \mathbb{R}, \vec{x} \in V \quad \Rightarrow \alpha\vec{x} \in V$

   (b) $\forall \vec{x} \in V \quad \Rightarrow 1 \cdot \vec{x} = \vec{x}$ (unit)

   (c) $\forall c, d \in \mathbb{R}, \vec{x} \in V \quad \Rightarrow (c \cdot d) \cdot \vec{x} = c \cdot (d \cdot \vec{x})$ (associative)

   (d) $\forall c, d \in \mathbb{R}, \vec{x}, \vec{y} \in V \Rightarrow c \cdot (\vec{x} + \vec{y}) = c \cdot \vec{x} + c \cdot \vec{y}$ and $(c + d) \cdot \vec{x} = c \cdot \vec{x} + c \cdot \vec{y}.$ (distributive)

Indeed, usually one says that any set of objects V for which you have defined addition and scalar multiplication that satisfiy the above properties, we call a set of vectors. In that way, even matrices or functions can be viewed as vectors. Examples are

1. *The set of real functions of one variable* $f : \mathbf{R} \to \mathbf{R}$
   The group operations are defined as

   $$\begin{aligned} (f_1 + f_2)(x) &:= f_1(x) + f_2(x) \\ (c \cdot f)(x) &:= c \cdot f(x) \end{aligned}$$

   Again it is easy to check that all the properties of a complex vector space are satisfied.

2. *$n \times n$ matrices*
   The elements of the vector space are

$$M = \begin{pmatrix} m_{11} & \dots & m_{1n} \\ \vdots & \ddots & \vdots \\ m_{n1} & \dots & m_{nn} \end{pmatrix} , \tag{3.5}$$

   where the m_{ij} are arbitrary real numbers. The addition and scalar multiplication are defined as

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} + \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{nn} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & \dots & a_{1n}+b_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1}+b_{n1} & \dots & a_{nn}+b_{nn} \end{pmatrix} ,$$

$$c \cdot \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} c \cdot a_{11} & \dots & c \cdot a_{1n} \\ \vdots & \ddots & \vdots \\ c \cdot a_{n1} & \dots & c \cdot a_{nn} \end{pmatrix} .$$

   Again it is easy to confirm that the set of real $n \times n$ matrices with the rules that we have defined here forms a vector space. Note that we are used to consider matrices as objects acting **on** vectors, but as we can see here we can also consider them as elements (vectors) of a vector space themselves.

3.2 Matrices

Vectors can be transformed. For example we can stretch them or rotate them. Many operations are possible, but of particular significance in physics are linear operations which are described by matrices. An example is

$$\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \lambda a_1 \\ \lambda a_2 \end{pmatrix} \tag{3.6}$$

which stretches a vector by a factor or

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} a_2 \\ -a_1 \end{pmatrix} \tag{3.7}$$

which rotates a vector by 90 degrees. We call matrices linear maps because the have the property that

$$\mathbf{M}(\lambda\vec{x} + \mu\vec{y}) = \lambda\mathbf{M}\vec{x} + \mu\mathbf{M}\vec{y} \tag{3.8}$$

which means that one can first add the vectors and then apply the transformation or one can first apply the transformation on the individual vectors and then add the results. All this is again true for n dimensions as well.

3.3 Eigenvalues, eigenvectors, singular values

When a matrix is applied to a vector the result is generally a different vector. However, for any matrix there will be some vectors that are special in the sense that upon multiplication with that matrix they will not change their orientation but only their length. Such vectors are called eigenvectors and the stretching factor is called the eigenvalues. Both are determined as solution to the equation

$$\mathbf{M}\vec{x} = \lambda\vec{x} \tag{3.9}$$

or equivalently

$$(\mathbf{M} - \lambda\mathbb{1})\vec{x} = 0\,. \tag{3.10}$$

Note that the nullvector that is the vector whose components are all zero does not qualify as an eigenvector. The way to proceed to compute the eigenvalues is to solve first the characteristic polynomial $det(\mathbf{M} - \lambda\mathbb{1}) =$ and then for every solution you compute the vector that solves $\mathbf{M}\vec{x} = \lambda\vec{x}$. Note that usually one normalizes the eigenvectors such that the squares of its components add up to 1.

If for an $n \times n$ matrix we can find n independent eigenvectors, then we can form the matrix $\mathbf{B}$ whose columns are the various eigenvectors. Let us assume that a matrix $\mathbf{M}$ has eigenvalues $\{\lambda_i\}$ and corresponding eigenvectors $\vec{x}_i\}$, then we have that

$$\mathbf{B} = (\vec{x}_1, \ldots, \vec{x}_n) \tag{3.11}$$

and we find that

$$\mathbf{B}^{-1}\mathbf{M}\mathbf{B} = \mathbf{D} \tag{3.12}$$

where $\mathbf{D}$ is a diagonal matrix whose diagonal entries are the eigenvalues of $\mathbf{M}$. This can be seen by applying the left hand side to the various basis vectors

$$\vec{e}_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.13}$$

where only the $i-th$ component of the vector is equal to 1 while all others vanish. Then we have that

$$\mathbf{B}^{-1}\mathbf{M}\mathbf{B}\vec{e}_i = \mathbf{B}^{-1}\mathbf{M}\vec{x}_i = \mathbf{B}^{-1}\lambda_i\vec{x}_i = \lambda_i\vec{e}_i \tag{3.14}$$

This is exactly the same action as that of the diagonal matrix $\mathbf{D}$ which together with linearity shows that the actions of the two matrices re the same for all vectors.

3.4 Functions of matrices

Quite frequently it will be necessary to compute a function of a matrix. In some cases it is straightforward to define what is meant by the function of a matrix. For example if we are given the function $f(x) = x^2$ then it is clear that for a matrix $\mathbf{A}$ we define $f(\mathbf{A} = \mathbf{A}^2$. More generally however, it may not be quite so clear how to compute functions of a matrix and indeed, there are two ways which are useful in different regimes. Fortunately they do not contradict each other in the situations where both can be applied.

Definition 49 *Given an operator $\hat{A}$ with eigenvalues a_i, and eigenvectors $\vec{a}_i$ such that there is a matrix $\mathbf{B}$ that allows to diagonalize the matrix $\mathbf{A}$, ie we have $\mathbf{A} = \mathbf{B}\mathbf{D}\mathbf{B}^{-1}$ with a diagonal matrix $\mathbf{D}$. Further*

have a function $f : \mathbf{R} \to \mathbf{R}$ *that maps real numbers into real numbers then we define*

$$f(\hat{A}) := \mathbf{B} f(\mathbf{D}) \mathbf{B}^{-1} \tag{3.15}$$

where $f(\mathbf{D})$ *is the matrix whose diagonal elements are given by* $f(a_i)$ *when the* a_i *are the diagonal elements of* $\mathbf{D}$*.*

Example: Given a diagonal matrix

$$\mathbf{D} = \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix} \tag{3.16}$$

then

$$f(\mathbf{D}) = \begin{pmatrix} f(d_1) & 0 & 0 \\ 0 & f(d_2) & 0 \\ 0 & 0 & f(d_3) \end{pmatrix} \tag{3.17}$$

Obviously the above way of defining the function of a matrix is only good when you can diagonalize the matrix. When this is impossible, then not all is lost. Indeed, if the function can be expanded into a Taylor series then there is another way of defining the function of a matrix.

Definition 50 *Given a function* $f : \mathbf{R} \to \mathbf{R}$ *that can be expanded into a power series*

$$f(z) = \sum_{i=0}^{\infty} f_i z^i \tag{3.18}$$

then we define

$$f(\hat{A}) = \sum_{i=0}^{\infty} f_i A^i \ . \tag{3.19}$$

Note also the definition of the derivative of a function of a matrix.

Definition 51 *The derivative of an operator function* $f(\hat{A})$ *is defined via* $g(z) = \frac{df}{dz}(z)$ *as*

$$\frac{df(\hat{A})}{d\hat{A}} = g(\hat{A}) \ . \tag{3.20}$$

Let us see whether the two definitions Def. 49 and 50 coincide for operators with complete set of eigenvectors and functions that can be expanded into a power series given in Eq. (3.18).

$$\begin{aligned} f(\hat{A}) &= \sum_{k=1}^{\infty} f_k A^k \\ &= \sum_{k=1}^{\infty} f_k (\mathbf{B}\mathbf{D}\mathbf{B}^{-1})^k \\ &= \sum_{k=1}^{\infty} f_k \mathbf{B}\mathbf{D}^k\mathbf{B}^{-1} \\ &= \mathbf{B}(\sum_{k=1}^{\infty} f_k \mathbf{D}^k)\mathbf{B}^{-1} \\ &= \mathbf{B}f(\mathbf{D})\mathbf{B}^{-1} \end{aligned} \tag{3.21}$$

Exercise:
1) Show that for any orthogonal operator $\hat{U}$, ie an operator that has the property $\hat{U}\hat{U}^T = \mathbb{1}$ we have $f(\hat{U}^\dagger \hat{A}\hat{U}) = \hat{U}^\dagger f(\hat{A})\hat{U}$

Proof: We use the fact that $\hat{U}\hat{U}^\dagger = \mathbb{1}$ to find

$$\begin{aligned} f(\hat{U}^\dagger \hat{A}\hat{U}) &= \sum_{k=0}^{\infty} f_k (\hat{U}^\dagger \hat{A}\hat{U})^k \\ &= \sum_{k=0}^{\infty} f_k \hat{U}^\dagger \hat{A}^k \hat{U} \\ &= \hat{U}^\dagger (\sum_{k=0}^{\infty} f_k \hat{A}^k)\hat{U} \\ &= \hat{U}^\dagger f(\hat{A})\hat{U} \quad . \end{aligned}$$

3.5 Markov processes

You may have heard a few times already statements like: "Disorder can only increase in time." In the following sections I would like to

illuminate this statement critically. Indeed, I would like to show you when this statement is correct. To this end I will need to define what is meant by disordered and how one can quantify disorder.

Given are chairs that are numbered from 1 to N and a probability distribution for finding a person on one of those chairs, ie $p(1)$ is the probability that the person is sitting on chair 1. Of course, as any decent probability, we have to have that $\sum_{i=1}^{N} p(i) = 1$ and all the $1 \leq p(i) \geq 0$. Now consider the following game. Assume that after a fixed amount of time the person throws a coin and with probability $1/2$ he remains in his chair and with probability $1/2$ he moves from chair k to the next chair $k+1$. if he is in chair N then he moves to chair 1. Given the probability distribution after step n, what is the probability distribution after step $n+1$? The answer can be captured in a matrix via

$$\vec{p}_{n+1} = \begin{pmatrix} \frac{1}{2} & 0 & \dots & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \ddots & 0 & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots \\ 0 & \ddots & \ddots & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \vec{p}_n \tag{3.22}$$

where the vector $\vec{p}_n$ stands for the column vector containing the probability distribution at step n. In general, for whatever set of rules that I invent, I will find a matrix $\mathbf{M}$ such that

$$\vec{p}_{n+1} = \mathbf{M}\vec{p}_n \tag{3.23}$$

Of course, not any matrix will do. Indeed, we have to make sure that a probability distribution is mapped again into a probability distribution. In other words, if $\sum_{i=1}^{N} p_n(i) = 1$ and all the $p_n(i) \geq 0$ then also $\sum_{i=1}^{N} p_{n+1}(i) = 1$ and all the $p_{n+1}(i) \geq 0$. As this has to be true for any valid probability distribution $\vec{p}_n$, we find the condition that $\sum_{i=1}^{N} M_{ij} = 1$ for any j because this ensures that

$$\sum_{i=1}^{N} p_{n+1}(i) = \sum_{i=1}^{N}\sum_{j=1}^{N} M_{ij}p_n(j) = \sum_{j=1}^{N}\sum_{i=1}^{N} M_{ij}p_n(j) = \sum_{j=1}^{N} p_n(j) = 1 \tag{3.24}$$

Definition 52 *A $n \times n$ matrix M is called a stochastic matrix if for*

every j we have that $\sum_{i=1}^{N} M_{ij} = 1$, ie all the columns of the matrix sum up to one.

Using the uniform probability distribution for N events defined by

$$\vec{e} = \frac{1}{N} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \tag{3.25}$$

we can write this as

$$\mathbf{M}^T \vec{e} = \vec{e} \tag{3.26}$$

Now the big question is, what happens when we consider the limit

$$\lim_{n \to \infty} \vec{p}_n \tag{3.27}$$

Before we look at this question in general, let us consider two special cases, that illuminate quite well what can happen.

Example I: To reduce the mathematical complications as much as possible, let us consider probability distributions with just two possible events and consider

$$\begin{pmatrix} p_{n+1}(1) \\ p_{n+1}(2) \end{pmatrix} = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix} \begin{pmatrix} p_n(1) \\ p_n(2) \end{pmatrix} \equiv \mathbf{M} \begin{pmatrix} p_n(1) \\ p_n(2) \end{pmatrix} \tag{3.28}$$

To determine

$$\vec{p}_n = \mathbf{M}^n \vec{p}_0 \tag{3.29}$$

we have to determine the eigenvectors and eigenvalues of $\mathbf{M}$. We find the eigenvalues

$$\lambda_1 = 1 \text{ and } \lambda_2 = \frac{1}{2} \tag{3.30}$$

and the corresponding eigenvectors

$$\vec{x}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \text{ and } \vec{x}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{3.31}$$

so that we have

$$\vec{p}_n = \mathbf{M}^n \vec{p}_0 \tag{3.32}$$

$$= \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \left(\frac{1}{2}\right)^n \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \vec{p}_0 \tag{3.33}$$

Therefore the limit is

$$\lim_{n\to\infty}\vec{p}_n = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\vec{p}_0 = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} \tag{3.34}$$

Therefore the limiting distribution is uniform, ie both possibilities 1 and 2 are equally likely.

Example II: Now consider a slightly different matrix $\mathbf{M}$, ie

$$\begin{pmatrix} p_{n+1}(1) \\ p_{n+1}(2) \end{pmatrix} = \begin{pmatrix} \frac{3}{4} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} \end{pmatrix}\begin{pmatrix} p_n(1) \\ p_n(2) \end{pmatrix} \equiv \mathbf{M}\begin{pmatrix} p_n(1) \\ p_n(2) \end{pmatrix} \tag{3.35}$$

To determine

$$\vec{p}_n = \mathbf{M}^n\vec{p}_0 \tag{3.36}$$

we have to determine the eigenvectors and eigenvalues of $\mathbf{M}$. We find the eigenvalues

$$\lambda_1 = 1 \text{ and } \lambda_2 = \frac{1}{4} \tag{3.37}$$

and the corresponding eigenvectors

$$\vec{x}_1 = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 \\ 1 \end{pmatrix} \text{ and } \vec{x}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{3.38}$$

so that we have

$$\vec{p}_n = \mathbf{M}^n\vec{p}_0 \tag{3.39}$$

$$= \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & \left(\frac{1}{4}\right)^n \end{pmatrix}\begin{pmatrix} \frac{\sqrt{5}}{3} & \frac{\sqrt{5}}{3} \\ \frac{\sqrt{2}}{3} & -\frac{\sqrt{8}}{3} \end{pmatrix}\vec{p}_0 \tag{3.40}$$

Therefore the limit is

$$\lim_{n\to\infty}\vec{p}_n = \begin{pmatrix} \frac{2}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix}\vec{p}_0 = \begin{pmatrix} \frac{2}{3} \\ \frac{1}{3} \end{pmatrix} \tag{3.41}$$

Therefore the limiting distribution is *not* uniform, ie both possibilities 1 and 2 are equally likely.

It is clear that there are matrices $\mathbf{M}$ such that the limiting distribution is uniform and there are others where it is not uniform. In particular in example II we observe that, even if we start with the uniform distribution, we obtain in the limit a non-uniform distribution. In following lectures we will see that the uniform distribution is the most disordered disribution of all and as a consequence we observe that there are processes in which disorder decreases with increasing n.

The question is now which structure of $\mathbf{M}$ ensures that the limiting distribution is actually uniform. The answer to that question requires the concept of *doubly stochastic* matrices.

Definition 53 *A $n \times n$ matrix M is called a doubly stochastic matrix if for every j we have that $\sum_{i=1}^{N} M_{ij} = 1 = \sum_{i=1}^{N} M_{ji}$, ie all the columns as well as all of the rows of the matrix sum up to one. In other words $\mathbf{M}\vec{e} = \vec{e}$ and $\mathbf{M}^T\vec{e} = \vec{e}$.*

The fact that the matrix $\mathbf{M}$ is doubly stochastic is not sufficient to ensure that the limiting distribution is uniform irrespective of the initial distribution. This becomes obvious from the trivial example

$$\mathbf{M} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{3.42}$$

for which

$$\lim_{n\to\infty} \vec{p}_n \;\; = \;\; \vec{p}_0 \, . \tag{3.43}$$

for any choice of $\vec{p}_0$. The problem however is quite clear, namely the matrix $\mathbf{M}$ has more than one eigenvector to the eigenvalue 1. Indeed, we find

Definition 54 *For a doubly stochasrtic matrix M which has exactly one eigenvector corresponding to the eigenvalue 1, then we finnd that*

$$\lim_{n\to\infty} \vec{p}_n = \vec{e} \tag{3.44}$$

ie the limiting distribution is uniform.

It is under these conditions that the disorder in the distribution is increasing. But how do we actually quantify disorder?

Chapter 4

Entropy, disorder and information

Disorder may be understood as the lack of information, eg quantified by the average number of questions that we may have to ask to identify an object or to find it. Therefore let me start by trying to build an intuitive understanding of the concept of classical information. A more quantitative approach will be taken in section 4.1, but for the full blown mathematical apparatus I have to refer you to textbooks, e.g. Cover and Thomas' book 'Elements of information theory'.

Imagine that you are holding an object, be it an array of cards, geometric shapes or a complex molecule and we ask the following question: *what is the information content of this object?* To answer this question, we introduce another party, say a friend, who shares some background knowledge with us (e.g. the same language or other sets of prior agreements that make communication possible at all), but who does not know the state of the object. We define the *information content* of the object as the size of the set of instructions that our friend requires to be able to identify the object, or better the state of the object. For example, assume that the object is a spin-up particle and that we share with the friend the background knowledge that the spin is oriented either upwards or downwards along the z direction with equal probability (see fig. 4.1 for a slightly more involved example). In this case, the only instruction we need to transmit to another party to let him recreate the state is whether the state is spin-up $\uparrow$ or spin-down

Figure 4.1: An example for a decision tree. Two binary choices have to be made to identify the shape (triangle or square) and the orientation (horizontal or rotated). In sending with equal probability one of the four objects, one therefore transmits 2 bits of information.

↓. This example shows that in some cases the instruction transmitted to our friend is just a choice between two alternatives. More generally, we can reduce a complicated set of instructions to n binary choices. If that is done we readily get a measure of the information content of the object by simply counting the number of binary choices. In classical information theory, a variable which can assume only the values 0 or 1 is called a *bit*. Instructions to make a binary choice can be given by transmitting 1 to suggest one of the alternative (say arrow up ↑) and 0 for the other (arrow down ↓).

To sum up, we say that n bits of information can be encoded in a system when instructions in the form of n binary choices need to be transmitted to identify or recreate the state of the system. In the following we will turn this idea into a more precise form.

4.1 Quantifying classical information

In 1948 Shannon developed a rigorous framework for the description of information and derived an expression for the information content of the message which depends on the probability of each letter occurring and results in the Shannon entropy. We will illustrate Shannon's reasoning in the context of the example above. Shannon invoked the law of large numbers and stated that, if the message is composed of N letters where N is very large, then the *typical* messages will be composed of Np_1 1's and Np_0 0's. For simplicity, we assume that N is 8 and that p_1 and p_0 are $\frac{1}{8}$ and $\frac{7}{8}$ respectively. In this case the typical messages are the 8 possible sequences composed of 8 binary digits of which only one is equal to 1 (see left side of figure 4.2). As the length of the message

Figure 4.2: The idea behind classical data compression. The most likely sequences are relabeled using fewer bits while rare sequences are discarded. The smaller number of bits still allows the reconstruction of the original sequences with very high probability.

increases (i.e. N gets large) the probability of getting a message which is all 1's or any other message that differs significantly from a typical sequence is negligible so that we can safely ignore them. But how many distinct typical messages are there? In the previous example the answer was clear: just 8. In the general case one has to find in how many ways the Np_1 1's can be arranged in a sequence of N letters? Simple

combinatorics tells us that the number of distinct typical messages is

$$\left(\begin{array}{c} N \\ Np_1 \end{array} \right) = \frac{N!}{(Np_1)!(Np_0)!} \tag{4.1}$$

and they are all equally likely to occur. Therefore, we can label each of these possible messages by a binary number. If that is done, the number of binary digits I we need to label each typical message is equal to $log_2 \frac{N!}{Np_1!Np_0!}$. In the example above each of the 8 typical message can be labeled by a binary number composed by $I = log_2 8 = 3$ digits (see figure 4.2). It therefore makes sense that the number I is also the number of bits encoded in the message, because Alice can unambiguously identify the content of each typical message if Bob sends her the corresponding binary number, provided they share the background knowledge on the labeling of the typical messages. All other letters in the original message are really redundant and do not add any information! When the message is very long almost any message is a typical one. Therefore, Alice can reconstruct with arbitrary precision the original N bits message Bob wanted to send her just by receiving I bits. In the example above, Alice can compress an 8 bits message down to 3 bits. Though, the efficiency of this procedure is limited when the message is only 8 letters long, because the approximation of considering only typical sequences is not that good. We leave to the reader to show that the number of bits I contained in a large N-letter message can in general be written, after using Stirling's formula, as

$$I = -N(p_1 log_2 p_1 + p_0 log_2 p_0) \ . \tag{4.2}$$

If we plug the numbers $\frac{1}{8}$ and $\frac{7}{8}$ for p_0 and p_1 respectively in equation 4.2, we find that the information content per symbol $\frac{I}{N}$ when N is very large is approximately 0.5436 bits. On the other hand, when the binary letters 1 and 0 appear with equal probabilities, then compression is not possible, i.e. the message has no redundancy and each letter of the message contains one full bit of information per symbol. These results match nicely the intuitive arguments given above.

Equation 4.2 can easily be generalized to an alphabet of n letters ρ_i each occurring with probabilities p_i. In this case, the average information in bits transmitted per symbol in a message composed of a large number N of letters is given by the Shannon entropy:

$$\frac{I}{N} = H\{p_i\} = -\sum_{i=1}^{n} p_i log p_i \ . \tag{4.3}$$

We remark that the information content of a complicated classical system composed of a large number N of subsystems each of which can be in any of n states occurring with probabilities p_i is given by $N \times H\{p_i\}$.

4.2 Elements of the theory of majorization

In the previous section we have derived the entropy as a sensible measure for quantifying disorder via the idea of disorder being characterized by a lack of information. In this section I am going to refine these ideas somewhat. Instead of comparing two probability distributions only via their entropy, we are now looking at a whole range of functions of these probability distributions and compare those numbers. This approach runs under the title of the theory of majorization. This approach will give us a refined picture of disorder of probability distributions. Furthermore, in recent years it has emerged that the theory of majorization is of central importance in quantum mechanics and in particular in quantum information theory.

Let us consider probability distributions that are ordered in descending order, ie $p(1) \geq p(2) \geq \ldots \geq p(n)$ and $q(1) \geq q(2) \geq \ldots \geq q(n)$ then we say that p is majorized by q, or in formulas, $\vec{p} \prec \vec{q}$, if for any value of k that satisfies $1 \leq k \leq n-1$ we have

$$\sum_{i=1}^{k} p(i) \leq \sum_{i=1}^{k} q(i) \tag{4.4}$$

and

$$\sum_{i=1}^{n} p(i) \leq \sum_{i=1}^{n} q(i) \tag{4.5}$$

More generally, for two vectors $\vec{p}$ and $\vec{q}$ that are not yet in non-increasing order, one first orders them and then checks the conditions above.

The relation to disorder is that it will make sense to say that if $\vec{x} \prec \vec{y}$ then $\vec{x}$ is more disordered than $\vec{y}$.

Examples: Consider the distributions

$$\vec{x} = \begin{pmatrix} 0.4 \\ 0.4 \\ 0.1 \\ 0.1 \end{pmatrix} \quad \vec{y} = \begin{pmatrix} 0.5 \\ 0.25 \\ 0.25 \\ 0 \end{pmatrix} \tag{4.6}$$

Then it is straightforward to check that we do not have that $\vec{x} \prec \vec{y}$ because

$$\begin{aligned} 0.4 &\leq 0.5 && (4.7)\\ 0.4 + 0.4 &\geq 0.5 + 0.25 && (4.8)\\ 0.4 + 0.4 + 0.1 &\leq 0.5 + 0.25 + 0.25 && (4.9)\\ 0.4 + 0.4 + 0.1 + 0.1 &= 0.5 + 0.25 + 0.25 && (4.10) \end{aligned}$$

The entropies corresponding to the two distributions are different. Indeed one finds that $H(\vec{x}) = 1.7219 \geq H(\vec{y}) = 1.5$. As a consequence we observe that the relation $\prec$ is not a total order in the sense that sometimes we neither that $\vec{x} \prec \vec{y}$ nor $\vec{y} \prec \vec{x}$ while we can always say that either $H(\vec{x}) \geq H(\vec{y})$ or $H(\vec{x}) \leq H(\vec{y})$.

Consider the distributions

$$\vec{x} = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \quad \vec{y} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix} \tag{4.11}$$

Then it is straightforward to check that $\vec{x} \prec \vec{y}$ because

$$\begin{aligned} \frac{1}{3} &\leq \frac{1}{2} && (4.12)\\ \frac{1}{3} + \frac{1}{3} &\leq \frac{1}{2} + \frac{1}{2} && (4.13)\\ \frac{1}{3} + \frac{1}{3} + \frac{1}{3} &= \frac{1}{2} + \frac{1}{2} && (4.14) \end{aligned}$$

Quite clearly, the distribution $\vec{x}$ would be considered as being more disordered than the distribution of $\vec{y}$, a viewpoint that is confirmed by computing the entropies $H(\vec{x}) = 1.5850 \geq H(\vec{y}) = 1$. Indeed, we will soon learn that $\vec{x} \prec \vec{y}$ implies $H(\vec{x}) \geq H(\vec{y})$

The fact that the idea of majorization is related to that of disorder is made more clear by the following. Take a probability vector $\vec{x}$ and interchange two of the elements to obtain $\vec{x}'$ and chose any $\lambda \in [0, 1]$ then

$$\lambda\vec{x} + (1 - \lambda)\vec{x}' \prec \vec{x} \tag{4.15}$$

Quite clearly, the probability distribution on the left hand side has been obtained from $\vec{x}$ by jumbling things a bit more and it can therefore be rightfully be said that the left hand side represents a more disordered situation than the right hand side. Instead of proving the above relation directly, I will prove a somewhat more general statement in terms of doubly stochastic matrices which includes, as a special case, the above relation.

Indeed, we have

Theorem 55 *Given a doubly stochastic matrix* $\mathbf{M}$ *and the Markov process*

$$\vec{p}_{k+1} = \mathbf{M}\vec{p}_k \tag{4.16}$$

then we have for all values of k *that*

$$\vec{p}_{k+1} \prec \vec{p}_k \tag{4.17}$$

ie in the sense of majorization, the probability distribution becomes more and more disordered.

Proof: To simplify notation, we will assume in the following that both vectors $\vec{p}_{k+1}$ and $\vec{p}_k$ are in non-increasing order, ie $p_{k+1}(1) \geq p_{k+1}(2) \geq \ldots \geq p_{k+1}(n)$ as well as $p_k(1) \geq p_k(2) \geq \ldots \geq p_k(n)$. If this is not the case, then one can always achieve this by reordering the vectors. This does also lead to a rearrangement of the matrix $\mathbf{M}$ but it does not affect the property that $\mathbf{M}$ is doubly stochastic.

In the following we will use the facts that

$$0 \leq \sum_{i=1}^{r} M_{ij} \leq 1 \quad \sum_{j=1}^{n}\sum_{i=1}^{r} M_{ij} = r \tag{4.18}$$

which both follow from the doubly stochasticity of the matrix $\mathbf{M}$. Now consider

$$\begin{aligned}
\sum_{i=1}^{r} p_{k+1}(i) - \sum_{i=1}^{r} p_k(i) &= \sum_{i=1}^{r}\sum_{j=1}^{n} M_{ij} p_k(j) - \sum_{i=1}^{r} p_k(i) \\
&= \sum_{j=1}^{n} \left(\sum_{i=1}^{r} M_{ij} \right) p_k(j) - \sum_{i=1}^{r} p_k(i) \\
&= \sum_{j=1}^{n} \left(\sum_{i=1}^{r} M_{ij} \right) p_k(j) - \sum_{i=1}^{r} p_k(i) \\
&\quad + p_k(r) \left(r - \sum_{j=1}^{n}\sum_{i=1}^{r} M_{ij} \right) \\
&= \sum_{j=1}^{n} \left(\sum_{i=1}^{r} M_{ij} \right) (p_k(j) - p_k(r)) - \sum_{i=1}^{r} (p_k(i) - p_k(r)) \\
&= \sum_{j=1}^{r} \left[\left(\sum_{i=1}^{r} M_{ij} \right) - 1 \right] (p_k(j) - p_k(r)) \\
&\quad + \sum_{j=r+1}^{n} \left(\sum_{i=1}^{r} M_{ij} \right) (p_k(i) - p_k(r)) \\
&\leq 0
\end{aligned}$$

where we have used in the last step that $p_k(1) \geq p_k(2) \geq \ldots \geq p_k(n)$. Therefore we have shown that for any r

$$\sum_{i=1}^{r} p_{k+1}(i) \leq \sum_{i=1}^{r} p_k(i)$$

and therefore

$$\vec{p}_{k+1} \prec \vec{p}_k$$

This finishes the proof.

Now we realize that the example I gave above is just a special case of this theorem. If we are given the probability vector $\vec{x}$ and we interchange the elements j and k to obtain $\vec{x}'$ and then form

$$\lambda\vec{x} + (1-\lambda)\vec{x}' \tag{4.19}$$

we can write this as well as

$$\lambda\vec{x} + (1-\lambda)\vec{x}' = \mathbf{M}\vec{x} \tag{4.20}$$

where the matrix $\mathbf{M}$ has non-zero entries only when i takes all values from 1 to n except j and k when we have $M_{ii} = 1$. Furthermore we have $M_{jj} = \lambda$ and $M_{kk} = \lambda$ and $M_{jk} = 1 - \lambda = M_{kj}$. This matrix is easily verified to be doubly stochastic so that the above theorem applies and we find

$$\lambda\vec{x} + (1-\lambda)\vec{x}' = \mathbf{M}\vec{x} \prec \vec{x}. \tag{4.21}$$

Note that indeed, there is an even closer relation between majorization and doubly stochastic matrices. In fact, we have $\vec{x} \prec \vec{y}$ exactly if there is a doubly stochastic matrix such that $\vec{x} = \mathbf{M}\vec{y}$. The proof of this statement is a little tedious, but can be found for example in the books of Horn and Johnson.

So, we have seen that in the sense of majorization a Markov process governed by a doubly stochastic matrix leads to an increase in disorder in every step. But of course we would also like to see whether the concept of majorization and of entropy are actually compatible with each other, ie if a probability distribution $\vec{p}$ is more disordered than $\vec{q}$ according to majorization we hope that the same is true for the entropy. The aim of the next two small theorems is exactly the verification of this statement. To this end we will employ some ideas from the theory of convex functions.

Lemma 56 *Given a concave function $f(x)$ then we have for $x \leq y$ that*

$$\frac{f(y) - f(x)}{y - x} \leq f'(x) \tag{4.22}$$

Proof: Use the defining property of concavity, namely that

$$f(\lambda y + (1-\lambda)x) \geq \lambda f(y) + (1-\lambda)f(x) \tag{4.23}$$

for $\lambda \in [0,1]$. From this is follows that

$$(y-x)\frac{f(x+\lambda(y-x))-f(x)}{\lambda(y-x)} \geq f(y)-f(x) \tag{4.24}$$

Now we take the limit λ towards 0 to find

$$(y-x)f'(x) \geq f(y)-f(x) \tag{4.25}$$

from which we find

$$f'(x) \geq \frac{f(y)-f(x)}{y-x} \tag{4.26}$$

which in turn finishes the proof.

Example: This is a very useful property which allows us to find all sorts of inequalities. Indeed, if we consider

$$f(x) = -x\log_2 x$$

then we find

$$f'(x) = -1-\log_2 x$$

. Then we use the lemma to find

$$-1-\log_2 x \geq \frac{-y\log_2 y + x\log_2 x}{y-x} \tag{4.27}$$

which leads to

$$y(\log_2 y - \log_2 x) \geq y-x. \tag{4.28}$$

For probability distributions $\vec{p}$ and $\vec{q}$ we therefore find

$$\sum_i p(i)\left(\log_2 p(i) - \log_2 q(i)\right) \geq 0. \tag{4.29}$$

Now we wish to prove

Lemma 57 *Given two probability distributions $\vec{p}$ and $\vec{q}$ such that $\vec{p} \prec \vec{q}$ then we have that $H(\vec{p}) \geq H(\vec{q})$.*

Proof: Consider

$$\begin{aligned}
\sum_{i=1}^{n} p(i)\log_2 p(i) &= p(1)\left(\log_2 p(1) - \log_2 p(2)\right) \\
&\quad +(p(1)+p(2))\left(\log_2 p(2) - \log_2 p(3)\right) \\
&\quad +(p(1)+p(2)+p(3))\left(\log_2 p(3) - \log_2 p(4)\right) \\
&\quad \vdots \\
&\quad +(p(1)+\ldots+p(n-1))\left(\log_2 p(n-1) - \log_2 p(n)\right) \\
&\quad +(p(1)+\ldots+p(n))\log_2 p(n) \\
&\leq q(1)\left(\log_2 p(1) - \log_2 p(2)\right) \\
&\quad +(q(1)+q(2))\left(\log_2 p(2) - \log_2 p(3)\right) \\
&\quad +(q(1)+q(2)+q(3))\left(\log_2 p(3) - \log_2 p(4)\right) \\
&\quad \vdots \\
&\quad +(q(1)+\ldots+q(n-1))\left(\log_2 p(n-1) - \log_2 p(n)\right) \\
&\quad +(q(1)+\ldots+q(n))\log_2 p(n) \\
&= \sum_{i=1}^{n} p(i)\log_2 q(i) \\
&\leq \sum_{i=1}^{n} p(i)\log_2 p(i)
\end{aligned}$$

where the last step used the inequality proven in the example. This finishes the proof.

Now the proof of the final theorem is rather simple.

Theorem 58 *Given a doubly stochastic matrix* $\mathbf{M}$ *and the Markov process*

$$\vec{p}_{k+1} = \mathbf{M}\vec{p}_k \tag{4.30}$$

then we have for all values of k *that*

$$H(\vec{p}_{k+1}) \geq H(\vec{p}_k) \tag{4.31}$$

ie in the sense of entropy, the probability distribution becomes more and more disordered.

Proof: From theorem 55 we find that

$$\vec{p}_{k+1} \prec \vec{p}_k \tag{4.32}$$

and employing the above lemma we find

$$H(\vec{p}_{k+1}) \geq H(\vec{p}_k). \tag{4.33}$$

This finishes the proof.

With the above I have shown you under what circumstances we can expect to obtain a behaviour that corresponds to the second law of thermodynamics, which very loosely spoken state that the entropy in an isolated system can never decrease. We have achieved this connection by making heavy use of the ideas of doubly stochastic maps, majorization and convex functions all of which are very important tools in thermodynamics and, as has been realized very recently, also in quantum mechanics.

If you wish to learn more about majorization, which is a beautiful theory in itself, have a look at the books by Horn and Johnson as well as the one by Marshall and Olkin 'Inequalities: Theory of majorization and its applications' and perhaps the book by Bhatia on 'Matrix Analysis' .