Analysis I + II

Literature

- These lecture notes!
- Various books in the library with the word "Analysis" in the title. In particular the books by Förster.

Some standard logical symbols commonly used in mathematics

- "$a \in X$" means, X is a set, and a is an element of X.
- "$\emptyset$" is the empty set, which contains no elements.
- "$X \cup Y$" is the union of the sets X and Y. It is the set which contains the elements of X and also the elements of Y.
- "$X \cap Y$" is the intersection. It is the set consisting of the elements which are in both X and Y.
- "$X \setminus Y$" is the set difference. It is the set containing the elements of X which are not in Y.
- "$X \subset Y$" means that X is a subset of Y. All the elements of X are also elements of Y. Note that many people use the notation $X \subseteq Y$ to expressly say that equality $X = Y$ is also possible. But I will assume that when writing $X \subset Y$, the case $X = Y$ is also possible.
- "$\forall$" means "for all", as for example: "$\forall x$, $x \geq 0$". That means: "for all x, we have the condition $x \geq 0$".
- "$\exists$" means "there exists".
- "$P \Rightarrow Q$" means that P and Q are logical statements, and if P is true, then Q must also be true. (If P is false, then the combined statement "$P \Rightarrow Q$" is true, regardless of whether or not Q is true.)
- "$P \Leftrightarrow Q$" means that both $P \Rightarrow Q$ and also $Q \Rightarrow P$ are true. That is, P and Q are logically equivalent; they are simply different ways of saying the same thing. (Although often it is not immediately clear that this is the case. Thus we need to think about why it is true, constructing a proof.)

Contents

Chapter 1

Numbers, Arithmetic, Basic Concepts of Mathematics

In some branches of mathematics — for example geometry, or graph theory — numbers are only used as a tool for describing things which are not really numerical in themselves. On the other hand, one can say that the subject of these lectures — Analysis — is purely and simply the study of numbers. Thus it is *pure* mathematics, rather than *applied* mathematics.

But what are numbers?

Surely everybody will agree that the numbers we use to count things: 1,2,3,..., are the numbers which are "naturally" given to us by nature. So we use the symbol $\mathbb{N}$ to denote the set of all such "natural" numbers. When thinking about physical objects which we count using the natural numbers, it is useful to have the standard arithmetical operations: addition, subtraction, and multiplication. The other standard operation, namely division, is often not so natural. For example if an odd number of people are in a lecture, and there are two tutorial groups, then it is impossible that they be of equal size (assuming that all students are active participants!). But it is equally true that subtraction has its limitations. For example if there are 50 students in a given lecture, then it is not possible to have 51 of those students deciding that it is not worthwile to continue attending the lecture, and thus withdrawing.

Despite these sensible objections, some hundreds of years ago people decided to expand our system of natural numbers with various kinds of "imaginary", or non-natural numbers. For example the number zero, and the negative whole numbers: -1,-2,..., were considered to be sensible things to think about. Modern mathematicians use the symbol $\mathbb{Z}$ to denote the set of both positive and negative whole numbers, together with 0. One says that $\mathbb{Z}$ is the set of (real) *integers*. Then, of course, in order to allow division, we also have the set of *rational* numbers $\mathbb{Q}$.

To summarize then, we have the "usual" systems of numbers:

- The natural numbers $\mathbb{N} = \{1, 2, 3, 4, \dots\}$
- The whole numbers, or integers $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$

- The rational numbers $\mathbb{Q} = \{\frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{N}\}$

But often other systems of numbers are used as well, perhaps without even realizing that they are different from those dealt with above. For example if we look at a clock, we see that there are 60 minutes in an hour.[1] Thus if a lecture starts at 15 minutes, and the lecture lasts for 90 minutes, then it is obvious that there are not enough minutes on the clock to describe the situation completely. The clock counts up the minutes to 60, but then when reaching 60 it suddenly jumps back to the number 0. Therefore we see that as far as the clock is concerned, we have the equation

$$15 + 90 = 45.$$

In mathematics we write

$$45 \equiv 15 + 90 \bmod 60.$$

Perhaps the reason for using the strange symbol "$\equiv$", which has three horizontal lines, rather than the more usual "$=$", is to avoid having all those overly smart people constantly telling us that the equation $15 + 90 = 45$ is "wrong".[2]

More generally, let $n \in \mathbb{N}$, and x, $y \in \mathbb{Z}$ be given. Then the expression

$$x \equiv y \bmod n$$

is defined to mean that the number $x - y$ is divisible by n. One writes $n|(x - y)$. That is to say, there exists some number $m \in \mathbb{Z}$ with $m \cdot n = (x - y)$.

For example we have that (15+90)−45 is 1 times 60, so that $15+90 \equiv 45 \bmod 60$ is true. Also $15 - 90 \equiv 45 \bmod 60$, since $60|(15 - 90) - 45$. On the other hand $15 + 90 \not\equiv 46 \bmod 60$, since $60 \nmid (15 + 90) - 46$.

If we do arithmetic according to the 60 minutes of the clock, then it can be said that we are doing "modular arithmetic", modulo 60. One writes $\mathbb{Z}/60\mathbb{Z}$ to denote this system of arithmetic with just 60 different numbers. It is usual to consider these 60 numbers to be the whole numbers from 0 to 59. In fact for any $n \in \mathbb{N}$, we can consider the system $\mathbb{Z}/n\mathbb{Z}$. Then, using the same convention, we could say that $\mathbb{Z}/n\mathbb{Z} = \{0, 1, 2, \ldots, n - 1\}$.

[1] This convention is due to the ancient Babylonians, whose number system was based on the number 60.

[2] Being even more overly smart, we could say that the expression $15 + 90 = 105$ is also "wrong", owing to the fact that the expression on the left-hand side, namely "15 + 90", is the description of two numbers and an arithmetical operation, whereas "105" is a pure number. And these are two different things. On the other hand, if — as in the usual convention — we agree to say that "15+90" is the number given by the result of the operation, then the expression is true. But then equally well, we could say that "15 + 90 mod 60" is also an arithmetical operation, and in this case it would make sense to say that the expression $45 = 15 + 90 \bmod 60$ is true. But then the expression $15 + 90 = 45 \bmod 60$ would be false.

1.1 The system $\mathbb{Z}/n\mathbb{Z}$ for $n = 60$

We have seen that in the system of modular arithmetic modulo 60, we have the equation

$$15 + 90 = 45.$$

Another way to think about this is to say that in the *usual* integer arithmetic of $\mathbb{Z}$ we have

$$15 + 90 = 105 = 1 \times 60 + 45.$$

In fact, given any integer x, and any natural number n, then we have two unique integers a and b, such that

$$x = an + b,$$

where $0 \leq b < n$. The number a is the result of the *whole number division* of x by n, and b is the *remainder* which results from this whole number division. The *operation* of finding the remainder when x is divided by n is denoted "x mod n". In particular then, we have the equation

$$45 = 105 \bmod 60.$$

Arithmetic generally has four operations: addition, subtraction, multiplication, and division. So let us say we have two numbers, x and y in our system $\mathbb{Z}/n\mathbb{Z}$. That is, we can assume that $0 \leq x, y < n$. Then in $\mathbb{Z}/n\mathbb{Z}$ we can simply *define* the sum of x and y to be

$$(x + y) \bmod n.$$

Similarly, the difference is

$$(x - y) \bmod n,$$

and the product is

$$(x \times y) \bmod n.$$

All of this is easy, since $x \pm y$ and $x \times y$ are always integers. However, what about division? The number $\frac{x}{y}$ is only occasionally an integer. And what do we do when $y = 0$?

The solution to this problem is to think of division as being the problem of solving a simple equation. Thus the number $\frac{x}{y}$ is really the solution z of the equation

$$z \times y = x.$$

For example, what is $\frac{1}{7}$ in our modular arithmetic modulo 60? That is, the problem is to find some number z with $0 \leq z < 60$, such that

$$1 = (z \times 7) \bmod 60.$$

The answer? It is $z = 43$, since $43 \times 7 = 301$, and $1 = 301 \bmod 60$.

On the other hand, what is $\frac{1}{2}$ modulo 60? That is, let z be such that

$$1 = (z \times 2) \bmod 60.$$

What is z? The answer is that there is *no* answer! That is to say, the number $\frac{1}{2}$ *does not exist* in the modular arithmetic modulo 60. The reason for this is that for all z we always have $z \times 2$ being an even number, yet since 60 is also an even number, it must be that the equation $1 = y \bmod 60$ can only have a solution when y is an *odd* number.

1.2 Equivalence relations, equivalence classes

Definition. *Let M be a set. The set of all pairs of elements of M is denoted by $M \times M$. Thus*

$$M \times M = \{(a, b) : a, b \in M\}.$$

This is called the Cartesian product *of M with itself.*[3] *An equivalence relation "$\sim$" on M is a subset of $M \times M$. Given two elements a, $b \in M$, we write $a \sim b$ to denote that the pair (a, b) is in the subset. For an equivalence relation, we must have:*

1. *$a \sim a$, for all $a \in M$ (reflectivity)*
2. *if $a \sim b$, then we also have $b \sim a$ (symmetry)*
3. *if $a \sim b$ and $b \sim c$ the we also have $a \sim c$ (transitivity)*

If $a \sim b$, then we say that "a is equivalent to b".

Examples

1. Given any set M, the most trivial possible equivalence relation is simply equality. Namely $a \sim b$ only when $a = b$.
2. In $\mathbb{Z}$, the set of integers, let us say that for two integers a and b, we have $a \sim b$ if and only if $a - b$ is an *even* number. Then this is an equivalence relation on $\mathbb{Z}$.
3. Again in $\mathbb{Z}$, this time take some natural number $n \in \mathbb{N}$. Now we define a to be equivalent to b if and only if there exists some further number $x \in \mathbb{Z}$ with

   $$a - b = xn.$$

   That is, the difference $a - b$ is divisible by n. And again, this is an equivalence relation on $\mathbb{Z}$.

   (Obviously, the example 2 is just a special case of example 3. In fact, it is the equivalence relation which results when we take $n = 2$.)

[3] More generally, if X and Y are two different sets, then the Cartesian product $X \times Y$ is the set of all pairs (x, y), with $x \in X$ and $y \in Y$.

Definition. *Given a set M with an equivalence relation $\sim$, then we have M being split up into* equivalence classes. *For each $a \in M$, the equivalence class containing a is the set of all elements of M which are equivalent to a. The equivalence class containing a is usually denoted by $[a]$. Therefore*

$$[a] = \{x \in M : x \sim a\}.$$

Note that if we have two equivalence classes $[a]$ and $[b]$ such that their intersection is not empty

$$[a] \cap [b] \neq \emptyset,$$

then we must have $[a] = [b]$. To see this, assume that $x \in [a] \cap [b]$. Then $x \sim a$ and $x \sim b$. But $x \sim a$ means that $a \sim x$, since the equivalence relation is symmetric. Then $a \sim b$ since it is transitive. If then $y \in [b]$, then we have $y \sim b$. But also $b \sim a$, and so using the transitivity of the equivalence relation again, we have $y \sim a$. Thus $y \in [a]$. So this shows that $[b]$ is contained in $[a]$. i.e. $[b] \subseteq [a]$. A similar argument shows that also $[a] \subseteq [b]$. Therefore we have shown that:

Theorem 1.1. *Given an equivalence relation $\sim$ on a set M, then the equivalence relation splits M into a set of* disjoint *equivalence classes.*

1.3 The system $\mathbb{Z}/n\mathbb{Z}$ revisited

In fact, rather than thinking about $\mathbb{Z}/n\mathbb{Z}$ as the set of numbers $\{0, \ldots, n-1\}$, it is more usual to say that $\mathbb{Z}/n\mathbb{Z}$ is the set of equivalence classes with respect to the equivalence relation given by $x \sim y$ if and only if $x - y$ is divisible by n. Thus

$$\mathbb{Z}/n\mathbb{Z} = \{[0], \ldots, [n-1]\}.$$

As we have seen, it is more usual to write

$$x \equiv y \bmod n,$$

rather than $x \sim y$ when describing this equivalence relation. One says that "x is congruent to y modulo n". It is easy to see that if two numbers x, $y \in \mathbb{Z}$ are given, then we have $x \equiv y \bmod n$ if, and only if, the remainder when x is divided by n is equal to the remainder when y is divided by n. That is, thinking of "mod" as an *operation* in the arithmetic of $\mathbb{Z}$, then we have $x \equiv y \bmod n$ if, and only if,

$$x \bmod n = y \bmod n.$$

Addition and multiplication in $\mathbb{Z}/n\mathbb{Z}$ are given by the simple rules

$$[x] + [y] = [x + y]$$

and

$$[x] \times [y] = [x \times y],$$

for any two numbers x, $y \in \mathbb{Z}$.

But we must be careful! It is necessary to check that these operations are *well-defined*. What does this mean?

Let us say that we have two different numbers x and x' in $\mathbb{Z}$ which are equivalent to one another. That is, we have $x \equiv x' \bmod n$. But then, since both x and x' are in the same equivalance class, we must have

$$[x] = [x'].$$

Similarly, if we have two numbers y and y' with $y \equiv y' \bmod n$, then we have

$$[y] = [y'].$$

To say that the addition operation, as we have defined it above, is well-defined, means that we must show that for arbitrary such x, x', y, and y', we always have

$$[x] + [y] = [x + y] = [x' + y'] = [x'] + [y'].$$

But this is clear, since

$$[x] = [x'] \quad \Rightarrow \quad x \equiv x' \bmod n \quad \Rightarrow \quad n|(x - x')$$

and

$$[y] = [y'] \quad \Rightarrow \quad y \equiv y' \bmod n \quad \Rightarrow \quad n|(y - y').$$

Therefore

$$\begin{aligned} n|(x - x') + (y - y') &\Rightarrow n|(x + y) - (x' + y') \\ &\Rightarrow (x + y) \equiv (x' + y') \bmod n \\ &\Rightarrow [x + y] = [x' + y']. \end{aligned}$$

It is now a simple exercise to show that multiplication is also well-defined in the arithmetic of $\mathbb{Z}/n\mathbb{Z}$.

But we are still left with the problem of division in $\mathbb{Z}/n\mathbb{Z}$. That is, given a, $b \in \mathbb{Z}$, does there exist an $x \in \mathbb{Z}$ such that $ax \equiv b \bmod n$?

1.4 The greatest common divisor function

To solve this equation, we first need to think about greatest common divisors.

Definition. *Let x, $y \in \mathbb{Z}$. Then we say that x is a* divisor *of y if there exists $z \in \mathbb{Z}$ with $y = xz$. Given two numbers a, $b \in \mathbb{Z}$, the number d is a* common divisor *of a and b if d is a divisor of both a and b. The* greatest common divisor *of a and b, is denoted by $gcd(a, b)$.*

Obviously, every integer is a divisor of the number zero. Furthermore, if x divides y, then obviously x also divides $-y$. Thus we can restrict our thinking to the integers which are either zero, or else positive. Given two integers a and b, not both zero, then obviously the number 1 is a common divisor. Therefore we always have $gcd(a, b) \geq 1$.

Theorem 1.2. *Given any two integers a and b, not both zero, then there exist two further integers x and y, such that*

$$xa + yb = gcd(a, b).$$

Proof. If one of the integers is zero, say $a = 0$, then obviously $gcd(a, b) = b$ (we assume here that b is positive). So we have[4]

$$gcd(a, b) = b = 0 \cdot a + 1 \cdot b,$$

and the theorem is true in this case.

Let us therefore assume that a and b are both positive integers. If the theorem were to be false, then it must be false for some pair of integers a, $b \in \mathbb{N}$. Assume that $a \leq b$, and that this pair is the smallest possible counterexample to the theorem, in the sense that the theorem is true for all pairs of integers $a' \leq b'$, with $b' < b$.

But we can immediately rule out the possibility that $a = b$, since in that case we would have $gcd(a, b) = b$, and again we would have the solution

$$gcd(a, b) = b = 0 \cdot a + 1 \cdot b.$$

Thus the pair a, b would not be a counterexample to the theorem. Therefore we must have $a < b$

So let $c = b - a$. Then $c \in \mathbb{N}$ and the theorem must be true for the smaller pair c, a. Thus there exist x', $y' \in \mathbb{Z}$ with

$$gcd(a, c) = x'a + y'c = x'a + y'(b - a) = (x' - y')a + y'b.$$

But what is $gcd(a, c) = gcd(a, b - a)$? Obviously, any common divisor of a and b is also a common divisor of a and $b - a$. Also any common divisor of a and $b - a$ must be a common divisor of both a and b. Therefore $gcd(a, c) = gcd(a, b)$, and so we have

$$gcd(a, b) = (x' - y')a + y'b,$$

which contradicts the assumption that the pair a, b is a counterexample to the theorem. It follows that there can be no counterexample, and the theorem must always be true. □

[4] From now on I will use the more usual notation $a \cdot b$, or even just ab, for multiplication, rather than the notation $a \times b$, which I have been using up till now.

Solving the equation $ax \equiv b \bmod n$

So let a, $b \in \mathbb{Z}$ be given, together with a natural number $n \in \mathbb{N}$. The question is, does there exist some $x \in \mathbb{Z}$ with $ax \equiv b \bmod n$? That is to say, does n divide the number $ax - b$? Or put another way, does there exist some $y \in \mathbb{Z}$ with

$$ax - b = yn \ ?$$

That is the same as

$$b = xa + (-y)n.$$

Therefore, we see that the equation $ax \equiv b \bmod n$ can only have a solution if every common divisor of a and n is also a divisor of b. That is, we must have $gcd(a, n)$ being a divisor of b.

On the other hand, assume that $gcd(a, n)$ does, in fact, divide b. Say $b = z \cdot gcd(a, n)$. Then, according to the previous theorem, there must exist u, $v \in \mathbb{Z}$ with

$$gcd(a, n) = ua + vn.$$

Therefore, we have

$$b = z \cdot gcd(a, n) = z(ua + vn) = (zu)a + (zv)n = xa + (-y)n,$$

when we take $x = zu$ and $y = -zv$.

To summarize:

Theorem 1.3. *The equation $ax \equiv b \bmod n$ has a solution if and only if $gcd(a, n)$ is a divisor of b. If $b = z \cdot gcd(a, n)$ then a solution is $x = zu$, where $gcd(a, n) = ua + vn$.*

1.5 The system $\mathbb{Z}/p\mathbb{Z}$, when p is a prime number

The prime numbers are $2, 3, 5, 7, 11, 13, 17, 19, 23, \ldots$. A prime number $p \in \mathbb{N}$ is such that it has no divisors in $\mathbb{N}$ other than itself and 1. Or put another way, for all $1 \leq a < p$ we have $gcd(a, p) = 1$. Therefore, according to the previous theorem, for all $[a] \in \mathbb{Z}/p\mathbb{Z}$ with $[a] \neq [0]$ there must exist some $[b] \in \mathbb{Z}/p\mathbb{Z}$ with $[a][b] = [1]$. That is to say,

$$ab \equiv 1 \bmod p$$

so that in the modular arithmetic modulo p, we have that $\frac{1}{a}$ is b. Therefore it is always possible to divide numbers by a. In fact, dividing by a is simply the same as multiplying by b.

On the other hand, if n is *not* a prime number, then there exists some a with $1 < a < n$ and $gcd(a, n) > 1$. In this case, according to the theorem, there can be no solution to the equation

$$ax \equiv 1 \bmod n.$$

Therefore it is impossible to divide numbers by a in modular arithmetic modulo n when n is not a prime number and $gcd(a, n) > 1$.

1.6 Mathematical induction

An example

The formula

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

is true for all $n \in \mathbb{N}$. How do we know that this is true??

Well, first of all, we know that it is true in the simple case $n = 1$. For here we just have

$$\sum_{k=1}^{1} \frac{1}{k(k+1)} = \frac{1}{1(1+1)} = \frac{1}{1+1}.$$

But then we know it's true for $n = 2$ as well, since

$$\begin{aligned} \sum_{k=1}^{2} \frac{1}{k(k+1)} &= \frac{1}{2(2+1)} + \sum_{k=1}^{1} \frac{1}{k(k+1)} \\ &= \frac{1}{2(2+1)} + \frac{1}{1+1} \\ &= \frac{2}{2+1}. \end{aligned}$$

Note that the second equation follows, since we already know that the formula is true for the case $n = 1$.

More generally, assume that we know that the formula is true for the case n, for some particular $n \in \mathbb{N}$. Then, exactly as before, we can write

$$\begin{aligned} \sum_{k=1}^{n+1} \frac{1}{k(k+1)} &= \frac{1}{(n+1)((n+1)+1)} + \sum_{k=1}^{n} \frac{1}{k(k+1)} \\ &= \frac{1}{(n+1)((n+1)+1)} + \frac{n}{n+1} \\ &= \frac{(n+1)}{(n+1)+1}. \end{aligned}$$

Therefore, the proof that the formula is true progresses stepwise through the numbers $1, 2, 3, \ldots$, and so we conclude that the formula is true for *all* $n \in \mathbb{N}$.

This is the principle of *mathematical induction* (or *vollständige Induktion* in German). Let $P(n)$ be some statement which depends on the number n, for arbitrary $n \in \mathbb{N}$. Then $P(n)$ is true for all $n \in \mathbb{N}$ if:

- First of all, the special case $P(1)$ can be proved, and
- then it can be proved that if $P(n)$ is true for some *arbitrarily given* $n \in \mathbb{N}$, then also $P(n+1)$ must be true.

We will be using mathematical induction very often here in these lectures! It is one of the most basic principles of mathematics.

1.7 The binomial theorem: using mathematical induction

The binomial theorem is concerned with what happens when the expression $(a+b)^n$ is multiplied out. For example, we have

$$\begin{array}{ll} 1: & a+b \\ 2: & a^2+2a+b^2 \\ 3: & a^3+3a^2b+3ab^2+b^3 \\ 4: & a^4+4a^3b+6a^2b^2+4ab^3+b^4 \\ 5: & a^5+5a^4b+10a^3b^2+10a^2b^3+5ab^4+b^5 \\ & \textit{etc.} \end{array}$$

Gradually we see a pattern emerging, namely Pascal's triangle:

$$\begin{array}{ccccccccccc} & & & & & 1 & & & & & \\ & & & & 1 & & 1 & & & & \\ & & & 1 & & 2 & & 1 & & & \\ & & 1 & & 3 & & 3 & & 1 & & \\ & 1 & & 4 & & 6 & & 4 & & 1 & \\ 1 & & 5 & & 10 & & 10 & & 5 & & 1 \\ 1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \\ 1 & & 7 & & 21 & & 35 & & 35 & & 21 & & 7 & & 1 \end{array}$$

and so on...

Writing out the expression $(a+b)^n$ as a sum, one uses the *binomial coefficients*, $\binom{n}{k}$. Thus one writes

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k}b^k.$$

So looking at Pascal's triangle, we see that $\binom{2}{0} = 1$, $\binom{7}{4} = 35$, and so forth. The binomial theorem is the formula that says that for all $n \in \mathbb{N}$ and $0 \leq k \leq n$, we have

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

But for the moment, let us simply *define* the number $\binom{n}{k}$ to be $\frac{n!}{k!(n-k)!}$, and then see if these are truly the binomial coefficients.

The expression $n!$ is called "n-factorial". For $n \in \mathbb{N}$ it is defined to be

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 3 \cdot 2 \cdot 1.$$

That is, just the product of all the numbers from 1 up to n. In the special case that $n = 0$, we define

$$0! = 1.$$

So let's see how this works out in the case $\binom{7}{4}$. We have

$$\binom{7}{4} = \frac{7!}{4!(7-4)!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(4 \cdot 3 \cdot 2 \cdot 1) \cdot (3 \cdot 2 \cdot 1)} = 35,$$

in agreement with Pascal's triangle.

But how do we prove it in general?

Theorem 1.4. *As in Pascal's triangle, we have*

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k},$$

that is

$$\frac{(n+1)!}{k!((n+1)-k)!} = \frac{n!}{(k-1)!(n-(k-1))!} + \frac{n!}{k!(n-k)!},$$

for all $n \in \mathbb{N}$ *and* $1 \leq k \leq n$.

Proof.

$$\begin{aligned}
\frac{n!}{(k-1)!(n-(k-1))!} + \frac{n!}{k!(n-k)!} &= \frac{k \cdot n!}{k!(n-k+1))!} + \frac{(n-k+1) \cdot n!}{k!(n-k+1)!} \\
&= \frac{k \cdot n!}{k!(n-k+1))!} + \frac{(n+1) \cdot n! - k \cdot n!}{k!(n-k+1)!} \\
&= \frac{(n+1) \cdot n!}{k!(n-k+1))!} \\
&= \frac{(n+1)!}{k!((n+1)-k)!}
\end{aligned}$$

□

Theorem 1.5. *For all* $n \in \mathbb{N}$ *and* $0 \leq k < n$, *we have*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k,$$

with

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Proof. Induction on n. For the case $n = 1$, the theorem is trivially true. Therefore we assume that the theorem is true in the case n, and so our task is to prove that

under this assumption, the theorem must also be true in the case $n+1$. We have:

$$\begin{aligned}
(a+b)^{n+1} &= (a+b)\cdot(a+b)^n \\
&= (a+b)\left(\sum_{k=0}^{n}\binom{n}{k}a^{n-k}b^k\right) \\
&= a\cdot\left(\sum_{k=0}^{n}\binom{n}{k}a^{n-k}b^k\right) + b\cdot\left(\sum_{k=0}^{n}\binom{n}{k}a^{n-k}b^k\right) \\
&= \sum_{k=0}^{n}\binom{n}{k}a^{n-k+1}b^k + \sum_{k=0}^{n}\binom{n}{k}a^{n-k}b^{k+1} \\
&= \sum_{k=0}^{n}\binom{n}{k}a^{n-k+1}b^k + \sum_{k=1}^{n+1}\binom{n}{k-1}a^{n-(k-1)}b^k \\
&= \binom{n}{0}a^{n+1} + \sum_{k=1}^{n}\left(\binom{n}{k}+\binom{n}{k-1}\right)a^{(n+1)-k}b^k + \binom{n}{n}b^{n+1} \\
&= \sum_{k=0}^{n+1}\binom{n+1}{k}a^{(n+1)-k}b^k
\end{aligned}$$

Here we have:

- the first equation is trivial,
- the second equation is the inductive hypothesis,
- the third and fourth equations are trivial,
- the fifth equation involves substituting $k-1$ for k in the second term,
- the sixth equation is trivial, and
- the seventh equation uses the theorem which we have just proved and, also the fact that $\binom{n}{0} = \binom{n}{n} = 1$, for all $n \in \mathbb{N}$.

□

1.8 The basic structures of algebra: groups, fields

Now that we have gotten the binomial theorem out of the way, let us return to thinking about numbers. We have $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q}$. The set of natural numbers $\mathbb{N}$ has addition and multiplication, but not subtraction and division.[5] The set of integers $\mathbb{Z}$ has addition, subtraction and multiplication, but division fails. However, in the

[5]Subtraction fails in $\mathbb{N}$: for example $1-2=-1$, but -1 is not an element of $\mathbb{N}$. Also division obviously fails: for example $1/2$ is also not an element of $\mathbb{N}$.

set of rational numbers $\mathbb{Q}$, all of these four basic operations can be carried out. (Of course, we exclude the special number zero when thinking about division.)

Furthermore, in the arithmetical system $\mathbb{Z}/n\mathbb{Z}$ we have addition, subtraction and multiplication. If (and only if) n is a prime number, then we also have division.

Arithmetical systems in which these four operations can be sensibly carried out are called *fields*. (In German, *Körper*.) In order to define the concept of a field, it is best to start by defining what we mean in mathematics when we speak of a *group*. But in order to do that, we should first say what is meant when we speak of a *function*, or *mapping*.

Definition. *Let X and Y be non-empty sets. A function $f : X \to Y$ is a rule which assigns to each element $x \in X$ a* unique *element $f(x) \in Y$.*

Examples

- For example, $f : \mathbb{N} \to \mathbb{N}$ with $f(n) = n^2$ is a function.
- But $f(n) = -n$ is not a function from $\mathbb{N}$ to $\mathbb{N}$, since $-n \notin \mathbb{N}$, for all $n \in \mathbb{N}$.
- On the other hand, $f(n) = -n$ is a function from $\mathbb{N}$ to $\mathbb{Z}$. That is, $f : \mathbb{N} \to \mathbb{Z}$.

Definition. *A* group *is a set G, together with a mapping $f : G \times G \to G$ satisfying the following three conditions:*

- *$f((f(a,b),c)) = f((a,f(b,c)))$, for all a, b and c in G.*
- *There exists an element $e \in G$ with $f((e,g)) = f((g,e)) = g$, for all $g \in G$.*
- *For all $g \in G$ there exists a element, usually denoted by $g^{-1} \in G$, such that $f((g^{-1},g)) = f((g,g^{-1})) = e$.*

Actually, this mapping $f : G \times G \to G$ is usually thought of as being an abstract kind of "multiplication". Therefore, we usually write ab or $a \cdot b$, rather than this cumbersome $f((a,b))$. With this notation, the group axioms become

- *$(ab)c = a(bc)$, for all a, b and c in G (The Associative Law).*
- *There exists a special element (the "unit element") $e \in G$, with $eg = ge = g$, for all $g \in G$ (The existence of the unit, or "neutral" element).*
- *For all $g \in G$ there exists an inverse $g^{-1} \in G$ with $g^{-1}g = gg^{-1} = e$. (The existence of inverses).*

If, in addition to this, the Commutative Law holds:

- *$ab = ba$, for all a and b in G,*

then the group G is called an "Abelian group".

Remark. *When thinking about numbers, you might think that it is entirely natural that all groups are Abelian groups. However this is certainly* not true*! Many commonly used groups are definitely not Abelian. For example the matrix groups — which a computer uses to calculate 3-dimensional graphics — are non-Abelian groups.*

But now we can define the idea of a *field*.

Definition. *A field is a set F, together with two operations, which are called "addition" and "multiplication". They are mappings*

$$+ : F \times F \to F$$

$$\cdot : F \times F \to F$$

satisfying the following conditions (or "axioms").

- *F is an Abelian group with respect to addition. The neutral element of F under addition is called "zero", denoted by the symbol 0. For each element $a \in F$, its inverse under addition is denoted by $-a$. Thus, for each a, we have $a + (-a) = 0$.*
- *Let $F \setminus \{0\}$ denote the set of elements of F which are not the zero element. That is, we remove 0 from F. Then $F \setminus \{0\}$ is an Abelian group with respect to multiplication. The neutral element of multiplication is called "one", denoted by the symbol 1. For each $a \in F$ with $a \neq 0$, the inverse is denoted by a^{-1}. Thus $a \cdot a^{-1} = 1$.*
- *The "Distributive Law" holds: For all a, b and c in F we have both*

$$\begin{aligned} a(b+c) &= ab + ac, \quad \text{and} \\ (a+b)c &= ac + bc. \end{aligned}$$

Examples

1. The set of rational numbers $\mathbb{Q}$, together with the usual addition and multiplication operations, is a field.
2. The set of integers $\mathbb{Z}$ is *not* a field, since $\mathbb{Z} \setminus \{0\}$ is not a group with respect to multiplication.
3. The sets $\mathbb{Z}/n\mathbb{Z}$, together with the addition and multiplication operations we have described, are fields if n is a prime number. However, if n is not prime, then $\mathbb{Z}/n\mathbb{Z}$ is not a field.

Remark. *A set R, having an addition and a multiplication operation which satisfies all the axioms for a field* except *that the elements of $R \setminus \{0\}$ do not necessarily have inverses under multiplication, is called a "ring". Thus $\mathbb{Z}/n\mathbb{Z}$, when n is not a prime is a ring, but not a field. Another standard example of a ring is the set of all polynomials $\mathbb{Q}[x]$ in one variable x, with coefficients in the field $\mathbb{Q}$. Finally of course, $\mathbb{Z}$ itself is a ring.*

Some simple consequences of the definition are the following.

Theorem 1.6. *Let F be a field. Then the following statements are true for all a and b in F.*

1. *Both $-a$ and a^{-1} (for $a \neq 0$) are unique.*
2. $a \cdot 0 = 0 \cdot a = 0$,
3. $a \cdot (-b) = -(a \cdot b) = (-a) \cdot b$,
4. $-(-a) = a$,
5. $(a^{-1})^{-1} = a$, *if* $a \neq 0$,
6. $(-1) \cdot a = -a$,
7. $(-a)(-b) = ab$,
8. $ab = 0 \Rightarrow a = 0$ *or* $b = 0$.

Proof. This involves a few simple exercises in fiddling with the definition.

1. If $a + a' = 0$ and $a + a'' = 0$ then $a' + (a + a'') = a' + 0$. Therefore
$$a'' = 0 + a'' = (a' + a) + a'' = a' + (a + a'') = a' + 0 = a'.$$
The fact that a^{-1} is unique is proved similarly.

2. Since $0 + 0 = 0$, we have $a(0 + 0) = a \cdot 0 + a \cdot 0 = a \cdot 0$. Then
$$\begin{aligned} 0 &= a \cdot 0 + (-(a \cdot 0)) \\ &= (a \cdot 0 + a \cdot 0) + (-(a \cdot 0)) \\ &= a \cdot 0 + (a \cdot 0 + (-(a \cdot 0))) \\ &= a \cdot 0 + 0 \\ &= a \cdot 0. \end{aligned}$$
The fact that $0 \cdot a = 0$ is proved similarly.

3. $0 = a \cdot 0 = a(b + (-b)) = ab + a(-b)$. Therefore we must have $-ab = a(-b)$. The other cases are similar.

4. $-a + (-(-a)) = 0$. But also $-a + a = 0$, and from (1) we know that additive inverses are unique. Therefore $a = -(-a)$.

5. $(a^{-1})^{-1} = a$ is similar.

6. We have
$$0 = 0 \cdot a = a(1 + (-1)) = 1 \cdot a + (-1) \cdot a = a + (-1)a.$$
Therefore $(-1)a = -a$.

7.
$$0 = 0 \cdot (-1) = (1 + (-1))(-1) = -1 + (-1)(-1).$$
Therefore
$$1 + 0 = 1 = 1 + (-1) + (-1)(-1) = (-1)(-1).$$
Then
$$(-a)(-b) = ((-1)a)((-1)b) = ((-1)(-1))ab = 1 \cdot ab = ab.$$

8. If $a \neq 0$ then
$$b = 1 \cdot b = (a^{-1}a)b = a^{-1}(ab) = a^{-1} \cdot 0 = 0.$$

□

1.9 How numbers are represented

Before proceeding with the usual definitions of analysis, it might be useful to have a quick look at a different way of representing numbers.

In the usual decimal notation we have for example:
$$2009 = 2 \cdot 10^3 + 0 \cdot 10^2 + 0 \cdot 10^1 + 9 \cdot 10^0,$$
or
$$\frac{22}{7} = 3.142\cdots = 3 \cdot 10^0 + 1 \cdot 10^{-1} + 4 \cdot 10^{-2} + 2 \cdot 10^{-3} + \cdots,$$
or
$$\sqrt{2} = 1.414\cdots = 1 \cdot 10^0 + 4 \cdot 10^{-1} + 1 \cdot 10^{-2} + 4 \cdot 10^{-3} + \cdots.$$

Continued fractions (Kettenbruchzahlen)

Here we have again simply the integer 2009 as its own continued fraction expression. In fact each integer $n \in \mathbb{Z}$ is simply itself in the continued fraction representation.

But then we have
$$\frac{22}{7} = 3 + \frac{1}{7},$$
and
$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cdots}}}$$

In fact, in general, using the Euclidean algorithm, we see that any *rational* number can be represented as a *finite* continued fraction

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots \cfrac{1}{a_k}}}}$$

where $a_0 \in \mathbb{Z}$ and $a_i \in \mathbb{N}$, for $i = 1, \ldots, k$. On the other hand, if a number is irrational then its continued fraction representation must be infinite.

Chapter 2

Analysis 1

2.1 Injections, Surjections, Bijections

The subject of mathematical analysis has much to do with *functions*, or *mappings*.[1] We have already seen that a function is a rule f, which assigns to each element $x \in X$ of a set X, a unique element $f(x) \in Y$ of a set Y. One writes

$$f : X \to Y.$$

Given such a function f from X to Y, one says that X is the *domain* of f. Furthermore, the set $\{f(x) : x \in X\} \subseteq Y$ is the *range* of f. One writes $f(X)$ for the range of X. Thus,

$$f(X) = \{f(x) : x \in X\}.$$

Given any element $y \in Y$, one writes $f^{-1}(y)$ to denote the subset of X consisting of all the elements which are mapped onto y. That is,

$$f^{-1}(y) = \{x \in X : f(x) = y\}.$$

Of course, if f is *not* a surjection, then $f^{-1}(y)$ must be the empty set, for some of the elements of Y.

Definition. *Let X and Y be sets, and let $f : X \to Y$ be a function. Then we say that:*

- *f is an* injection *if, given any two* different *elements x_1, $x_2 \in X$ with $x_1 \neq x_2$, we must have $f(x_1) \neq f(x_2)$. Or put another way, the only way we can have $f(x_1) = f(x_2)$ is when $x_1 = x_2$.*
- *f is a* surjection *if, for all $y \in Y$, there exists some $x \in X$ with $f(x) = y$. That is, if $f : X \to Y$ is a surjection, then we must have $f(X) = Y$.*
- *f is a* bijection *if it is both an injection, and also a surjection.*

[1] That is, "Funktionen" and "Abbildungen" in German. The words function and mapping both mean the same thing in mathematics. Perhaps some people would say that a mapping $f : X \to Y$ is a *function* if the set Y is some sort of system of "numbers", otherwise it is a mapping. But we certainly needn't make this distinction here.

Examples

Consider the following functions $f : \mathbb{Z} \to \mathbb{Z}$:

- $f(a) = 2a$, for all $a \in \mathbb{Z}$. This is an injection, but it is not a surjection since only even numbers are of the form $2a$, for $a \in \mathbb{Z}$. For example, the number -3 is in $\mathbb{Z}$, yet there exists *no* integer a with $2a = -3$.
- $f(a) = \begin{cases} a/2, & \text{if } a \text{ is even,} \\ (a+1)/2, & \text{if } a \text{ is odd,} \end{cases}$

  is a surjection, but it is *not* an injection. For example, $f(0) = 0 = f(-1)$.
- $f(a) = -a$, for all $a \in \mathbb{Z}$, is a bijection.

Theorem 2.1. *Let $f : X \to Y$ be an injection. Then there exists a surjection $g : Y \to X$. Conversely, if there exists a* surjection *$f : X \to Y$, then there exists an* injection *$g : Y \to X$.*

Proof. Assume that there exists an injection $f : X \to Y$. A surjection $g : Y \to X$ can be constructed in the following way. First choose some particular element $x_0 \in X$. Then a surjection $g : Y \to X$ is given by the rule

$$g(y) = \begin{cases} x, & \text{where } f(x) = y \text{ if } y \in f(X), \\ x_0, & \text{if } y \notin f(X), \end{cases}$$

for all $y \in Y$.

Going the other way, assume that there exists a *surjection* $f : X \to Y$. Then an injection $g : Y \to X$ can be constructed in the following way. Since f is a surjection, we know that the set $f^{-1}(y) \subset X$ is not empty, for each $y \in Y$. Therefore, for each $y \in Y$, choose some particular element $x_y \in f^{-1}(y)$. Then the injection $g : Y \to X$ is given by the rule $g(y) = x_y$, for all $y \in Y$.

Remark: This procedure of choosing elements from a collection of sets is only valid if we use the "axiom of choice" in the theory of sets. This is certainly the usual kind of mathematics which almost all mathematicians pursue. However it is perfectly possible to develop an alternative theory of mathematics in which the axiom of choice is not true. In this alternative mathematics, this proof would *not* be valid. □

Furthermore, we have the following theorem about bijections.

Theorem 2.2 (Schröder-Bernstein)**.** *Let X and Y be sets. Assume that there exists an injection $f : X \to Y$, and also there exists a surjection $g : X \to Y$. Then there exists a bijection $h : X \to Y$.*

Proof. An exercise. □

2.2 Constructing the set of real numbers $\mathbb{R}$

2.2.1 Dedekind cuts

The simplest method for defining real numbers is to use the technique of *Dedekind cuts*.

Definition. *A Dedekind cut of the rational numbers $\mathbb{Q}$ is a pair of non-empty subsets A, $B \subset \mathbb{Q}$, such that if $a \in A$ and $x < a$, then $x \in A$ as well. Furthermore, if $b \in B$ and $y > b$, then $y \in B$ as well. Also $A \cup B = \mathbb{Q}$ and $A \cap B = \emptyset$. Finally, we require that the subset A has* no *greatest element.*

Then the set of real numbers $\mathbb{R}$ can be *defined* to be the set of Dedekind cuts of the rational numbers. One may think of each real number as the "point" between the "upper" set B and the "lower" set A. If the given real number happens to be a rational number, then it is the smallest number in the set B.

For example, it is well known that the number $\sqrt{2}$ is irrational.[2]

Theorem 2.3. *There exists no rational number $\frac{a}{b}$ with $\left(\frac{a}{b}\right)^2 = 2$.*

Proof. Assume to the contrary that there does indeed exist such a rational number $\frac{a}{b}$. Perhaps there exist many such rational square roots of 2. If so, choose the *smallest* one, $\frac{a}{b}$, in the sense that if $\frac{a'}{b'}$ is also a square root of 2, then we must have $b \leq b'$.

Now, since $\frac{a}{b}$ is a square root of 2, we must have

$$\left(\frac{a}{b}\right)^2 = 2.$$

Therefore,

$$a^2 = 2b^2.$$

But this can only be true if a is an even number. So let us write $a = 2c$, with $c \in \mathbb{Z}$. Then we have

$$a^2 = 4c^2 = 2b^2.$$

Or

$$b^2 = 2c^2.$$

Therefore b is also an even number, say $b = 2d$. But in this case we must have $\frac{c}{d} = \frac{a}{b}$, so $\frac{c}{d}$ is also a square root of 2. But this is impossible, since $d < b$ and we have assumed that $\frac{a}{b}$ was a smallest possible square root of 2. □

Given any rational number $q \in \mathbb{Q}$, we have q^2 being also a rational number. So we can make a Dedekind cut by taking the pair (A, B), with B being all the positive rational numbers b with $b^2 > 2$. Then A is the rest of the rational numbers.

[2] We have seen that this must be true, owing to the fact that the continued fraction representation of $\sqrt{2}$ is infinite.

That is, A is the set of rational numbers *less than* $\sqrt{2}$, and B is the set of rational numbers *greater than* $\sqrt{2}$. So this Dedekind cut defines the real number $\sqrt{2}$.

Of course the rational numbers themselves can also be represented in terms of Dedekind cuts. For example the number 2 is simply the Dedekind cut (A, B), with $A = \{q \in \mathbb{Q} : q < 2\}$ and $B = \{q \in \mathbb{Q} : q \geq 2\}$. So here, the number 2 is the smallest number in the set B.

The reason Dedekind brought in this definition in the 19th century is that with it, it is possible to define the real numbers *without*, having to use the axiom of choice.

2.2.2 Decimal expansions

Written as a decimal number, we have

$$\frac{1}{3} = 0.333333333333333\ldots.$$

Also

$$\sqrt{2} = 1.414213562373095\ldots.$$

Another well-known irrational number is

$$\pi = 3.141592653589793\ldots.$$

As we know, a rational number has a *repeating* decimal expansion. On the other hand, irrational numbers do not repeat when written out as decimal expansions.

One might say that, for example, the number

$$0.999999999999999999\ldots$$

is the same as the number

$$1.000000000000000000\ldots,$$

which, of course, is really just the number one. But if we exclude decimal expansions which end in a never-ending sequence of 9s, then the decimal expansion for each real number is *unique*. Therefore, an alternative way to define the real numbers is to say that they are nothing more than the set of all possible decimal expansions which do not end with an infinite sequence of 9s.

2.2.3 Convergent sequences

But the most usual method of defining the real numbers is as equivalence classes of convergent sequences. We need the idea of convergent sequences in any case, so let us take the set of real numbers $\mathbb{R}$ as given (using either of the previous definitions), and consider the theory of sequences, either in $\mathbb{Q}$ or in $\mathbb{R}$ itself.[3]

[3] Again — and this is the last time I will mention this fact — the theory of convergent sequences requires the axiom of choice.

2.3 Convergent sequences

A *sequence* is simply an infinite list of numbers. For example, the sequence

$$1, 2, 3, 4, 5, 6, 7, \ldots$$

is certainly easy to think about, but obviously it doesn't *converge*. The numbers in the sequence get larger and larger, increasing beyond all possible finite bounds. Another example is the sequence

$$1, -1, 1, -1, 1, -1, 1, -1, \ldots$$

This sequence remains bounded, just jumping back and forth between the two numbers 1 and -1. But it never converges to anything; it always keeps jumping back and forth.

An example of a convergent sequence is

$$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \ldots$$

This sequence obviously converges down to zero.

In general, when thinking about abstract sequences of numbers, we write

$$a_1, a_2, a_3, \ldots$$

So a_1 is the first number in the sequence. a_2 is the second number, and so forth. A shorter notation, for representing the whole sequence is

$$(a_n)_{n \in \mathbb{N}}.$$

But when thinking about the concept of "convergence", it is clear that we also need an idea of the *distance* between two numbers.

Definition. *Given a real (or rational) number x, the* absolute value *of x is given by*

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

So one can think of $|x|$ as being either zero, if x is zero, otherwise $|x|$ is the distance of x from zero. More generally, given two numbers a and b, the distance between them is $|a - b|$.

It is a simple matter to verify that the *triangle inequality* always holds. That is, for all x, $y \in \mathbb{R}$, we always have

$$|x + y| \leq |x| + |y|.$$

Definition. *The sequence $(a_n)_{n\in\mathbb{N}}$ converges to the number a if, for all positive numbers $\epsilon > 0$, there exists some sufficiently large natural number $N_\epsilon \in \mathbb{N}$, such that $|a - a_n| < \epsilon$, for all $n \geq N_\epsilon$. In this case, we write*

$$\lim_{n\to\infty} a_n = a.$$

If the sequence does not converge, then one says that it diverges.

This definition is rather abstract. But, for example, it doesn't really tell us what is happening with the simple sequence $1, -1, 1, -1, 1, -1, \ldots$ Although this sequence does not converge — according to our definition — still, in a way it "really" converges to the two different points 1 and -1.

2.3.1 Bounded sets

Given the set of all real numbers $\mathbb{R}$, let us consider some arbitrarily given subset $A \subset \mathbb{R}$.

Definition. *We will say that $A \subset \mathbb{R}$ is* bounded above, *if there exists some $K \in \mathbb{R}$, such that $a \leq K$, for all $a \in A$. The number K is called an* upper bound *for A. Similarly, A is* bounded below *if there exists some $L \in \mathbb{R}$ with $a \geq L$, for all $a \in A$. Then L is a* lower bound *for A. If A is bounded both above and below, then we say that A is* bounded. *In this case, clearly there exists some $M \geq 0$ with $|a| \leq M$, for all $a \in A$.*

If $A \neq \emptyset$, and if A is bounded above, then the smallest upper bound is called the least upper bound, *written lub(A). Similarly, glb(A) is the greatest lower bound. The least upper bound is also called the* Supremum, *that is,* sup(A). *The greatest lower bound is called the* Infimum, *written* inf(A).

Examples

- Let $[0, 1] = \{x \in \mathbb{R} : 0 \leq x \leq 1\}$. Then $[0, 1]$ is bounded, and the least upper bound is 1; the greatest lower bound is 0.
- This time, take $[0, 1) = \{x \in \mathbb{R} : 0 \leq x < 1\}$. This is of course also bounded, and the least upper bound is again 1, even though 1 is not contained in the subset $[0, 1)$.
- $\mathbb{N} \subset \mathbb{R}$ is bounded below (with greatest lower bound 1), but it is not bounded above.
- $\mathbb{Z} \subset \mathbb{R}$ is not bounded below, and also not bounded above.

2.3.2 Subsequences

Definition. *Let $i : \mathbb{N} \to \mathbb{N}$ be a mapping such that for all n, $m \in \mathbb{N}$ with $m < n$, we have $i(m) < i(n)$. Then given a sequence $(a_n)_{n\in\mathbb{N}}$, a* subsequence, *with respect to the mapping i, is the sequence $(a_{i(n)})_{n\in\mathbb{N}}$.*

For example, let's look again at the sequence $((-1)^n)_{n\in\mathbb{N}}$. Then take the mapping $i : \mathbb{N} \to \mathbb{N}$ with $i(n) = 2n$. In this case, we have the subsequence

$$((-1)^{i(n)})_{n\in\mathbb{N}} = ((-1)^{2n})_{n\in\mathbb{N}} = \left(\left((-1)^2\right)^n\right)_{n\in\mathbb{N}} = (1^n)_{n\in\mathbb{N}} = (1)_{n\in\mathbb{N}}.$$

But this is just the trivially convergent constant sequence of 1s, which obviously converges to 1.

So we see that in this example, the sequence really consists of two convergent subsequences, one of them converges to the number 1, and the other converges to the number -1.

On the other hand, the sequence $(n)_{n\in\mathbb{N}}$ has *no* convergent subsequences. All subsequences simply diverge to "infinity". The problem is that it just keeps getting bigger, increasing beyond all bounds. To avoid this, we have the following definition.

Definition. *The sequence $(a_n)_{n\in\mathbb{N}}$ is called bounded if the set $\{a_n : n \in A\}$ is bounded in $\mathbb{R}$. (Similarly, we say the sequence is bounded above, or below, if those conditions apply to this set.)*

We also have an interesting refinement of this definition.

Definition. *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence. Then*

$$\lim_{n\to\infty} \sup a_n = \lim_{n\to\infty} \sup\{a_m : m \geq n\},$$

assuming this limit exists. Similarly

$$\lim_{n\to\infty} \inf a_n = \lim_{n\to\infty} \inf\{a_m : m \geq n\},$$

if it exists. One says "limit superior" and "limit inferior".

Theorem 2.4 (Bolzano-Weierstraß). *Let $(a_n)_{n\in\mathbb{N}}$ be a bounded sequence in $\mathbb{R}$. Then there exists a convergent subsequence, converging to a number in $\mathbb{R}$.*

Proof. Since the sequence is bounded, there must exist two real numbers $x < y$, such that

$$x \leq a_n \leq y,$$

for all $n \in \mathbb{N}$. Let $z = (x + y)/2$. That is, z is the point half way between x and y. So now the original interval from x to y has been split into two equal subintervals, namely the lower one from x to z, and the upper one from z to y. Since our sequence contains infinitely many elements, it must be that there are

infinitely many in one of these two subintervals. For example, let's say there are infinitely many elements of the sequence in the lower subinterval. In this case, we set $x_1 = x$ and $y_1 = z$. If only finitely many elements of the sequence are in the lower subinterval, then there must be infinitely many in the upper subinterval. In this case, we set $x_1 = z$ and $y_1 = y$.

Then the interval from x_1 to y_1 is divided in half as before, and a subinterval x_2 to y_2 is chosen which contains infinitely many elements of the sequence. And so on. By this method, we construct two new sequences, $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$, and we have

$$x \leq x_1 \leq x_2 \leq x_3 \leq x_4 \leq \cdots \leq y_4 \leq y_3 \leq y_2 \leq y_1 \leq y$$

We have

$$y_n - x_n = \frac{y - x}{2^n}.$$

Therefore the two sequences approach each other more and more nearly as n gets larger.

Now take (A, B) to be the following Dedekind cut of the rational numbers $\mathbb{Q}$.

$$B = \{q \in \mathbb{Q} : q \geq x_n, \forall n\}.$$

Then set $A = \mathbb{Q} \setminus B$. Let us say that $a \in \mathbb{R}$ is the real number which is given by the Dedekind cut (A, B). Then clearly there is a subsequence $(a_{i(n)})_{n\in\mathbb{N}}$ with

$$\lim_{n\to\infty} a_{i(n)} = a.$$

□

Definition. *The sequence* $(a_n)_{n\in\mathbb{N}}$ *is called* monotonically increasing *if* $a_n \leq a_{n+1}$*, for all* n*; it is* monotonically decreasing *if* $a_n \geq a_{n+1}$*, for all* n*; finally, one simply says that it is* monotonic *if it is either monotonically increasing, or monotonically decreasing.*

It is a simple exercise to show that theorem 2.4 implies that the following theorem is also true.

Theorem 2.5. *Every bounded, monotonic sequence in* $\mathbb{R}$ *converges.*

Conversely, we have that

Theorem 2.6. *Every convergent sequence is bounded.*

Proof. This is really rather obvious. Let the sequence $(a_n)_{n\in N}$ converge to the point $a \in \mathbb{R}$. Choose $\epsilon = 1$. Then there exists some $N(1) \in \mathbb{N}$ with $|a - a_n| < 1$, for all $n \geq N(1)$. We have the numbers $|a_1|, |a_2|, \ldots, |a_{N(1)}|$. Let M be either the largest of these numbers, or else $|a| + 1$, whichever is larger. Then we must have $|a_n| \leq M$, for all $n \in \mathbb{N}$. Thus the sequence is bounded below by $-M$, and above by M. □

2.3.3 Cauchy sequences

Definition. *A sequence* $(a_n)_{n\in\mathbb{N}}$ *is called a* Cauchy sequence *if for all* $\epsilon > 0$, *there exists a number* $N(\epsilon) \in \mathbb{N}$ *such that* $|a_n - a_m| < \epsilon$, *for all* m, $n \geq N(\epsilon)$.

It is again an exercise to show that:

Theorem 2.7. *Every convergent sequence is a Cauchy sequence.*

The alternative, and more usual way to define the real numbers is as equivalence classes of Cauchy sequences of rational numbers. The equivalence relation is the following.

Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be two Cauchy sequences, with a_n and $b_n \in \mathbb{Q}$, for all n. Then we will say that the they are equivalent to one another if — and only if — for all $\epsilon > 0$, there exists some $N(\epsilon) \in \mathbb{N}$, with $|a_n - b_n| < \epsilon$, for all $n \geq N(\epsilon)$. The fact that this is, in fact, an equivalence relation is also left as an exercise. Then $\mathbb{R}$ is defined to be the set of equivalence classes in the set of Cauchy sequences in $\mathbb{Q}$.

But not all Cauchy sequences converge!!

If we always think about the set of real numbers $\mathbb{R}$, then of course every Cauchy sequence converges. As we have seen, this is simply a way of *defining* the set of real numbers!

But if we think about other sets which are not simply all of $\mathbb{R}$, then it is definitely *not true* that all Cauchy sequences converge. For example, let us consider the set

$$(0, 1] = \{x \in \mathbb{R} : 0 < x \leq 1\}.$$

Within this set, the sequence $(1/n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Considered in $\mathbb{R}$, it converges to the number 0. But considered within $(0, 1]$ alone, it *doesn't converge*, since 0 is not an element of $(0, 1]$.

Similarly, if we consider the set of rational numbers $\mathbb{Q}$, then there are many Cauchy sequences which converge to irrational numbers, when considered in $\mathbb{R}$. Yet those irrational numbers do not belong to $\mathbb{Q}$. Therefore they *do not converge* in $\mathbb{Q}$.

On the other hand, all Cauchy sequences *do* converge in $\mathbb{R}$.

To begin with, it is a simple matter to show that all Cauchy sequences are bounded. Therefore $\lim_{n\to\mathbb{N}} \sup a_n$ exists. Let B be the set of all rational numbers greater than or equal to $\lim_{n\to\mathbb{N}} \sup a_n$, and let $A = \mathbb{Q} \setminus B$. Then the pair (A, B) is a Dedekind cut of $\mathbb{Q}$, representing the real number $a \in \mathbb{R}$ say, and we must have the Cauchy sequence $(a_n)_{n\in\mathbb{N}}$ converging to a.

To see this, let $\epsilon > 0$ be chosen. The problem is to show that there exists some $N(\epsilon) \in \mathbb{N}$, such that $|a - a_n| < \epsilon$, for all $n \geq N(\epsilon)$.

Let us start by choosing some rational number $q \in A$ with $|a - q| < \epsilon/6$. Then there must exist some other rational number $p \in B$, with $p > a$ and $|p - q| < \epsilon/3$.

Looking at the definition of $\lim_{n\to\mathbb{N}} \sup a_n$, we see that there exists some $N_1 \in \mathbb{N}$ such that $a_n < p$, for all $n \geq N_1$.

Since the sequence $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, there exists a number $N_2 \in \mathbb{N}$ such that for all $n, m \geq N_2$, we have $|a_n - a_m| < \epsilon/3$. Furthermore, looking again at the definition of $\lim_{n\to\mathbb{N}} \sup a_n$, we see that there exist arbitrarily large numbers m, with $q < a_m$. Then setting $N(\epsilon) = \max\{N_1, N_2\}$, and taking $m \geq N$ with $q < a_m$, we have for all $n \geq N(\epsilon)$

$$\begin{aligned} |a - a_n| &= |(a - q) + (q - a_m) + (a_m - a_n)| \\ &\leq |a - q| + |q - a_m| + |a_m - a_n| \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\ &= \epsilon. \end{aligned}$$

Therefore we have the theorem:

Theorem 2.8. *All Cauchy sequences converge in* $\mathbb{R}$.

2.3.4 Sums, products, and quotients of convergent sequences

Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be two convergent sequences in $\mathbb{R}$ with

$$\lim_{n\to\infty} a_n = a \quad \text{and} \quad \lim_{n\to\infty} b_n = b.$$

Then the sequence $(a_n + b_n)_{n\in\mathbb{N}}$ also converges, and

$$\lim_{n\to\infty} (a_n + b_n) = a + b.$$

To see this, let $\epsilon > 0$ be given, and let $N_a(\epsilon), N_b(\epsilon) \in \mathbb{N}$ with $|a - a_n| < \epsilon/2$ and $|b - b_m| < \epsilon/2$, for all $n \geq N_a(\epsilon)$ and $m \geq N_b(\epsilon)$. Then take $N(\epsilon) = \max\{N_a(\epsilon), N_b(\epsilon)\}$, that is, the larger of the two numbers. For any $k \geq N(\epsilon)$ we then have

$$\begin{aligned} |(a + b) - (a_k - b_k)| &= |(a - a_k) + (b - b_k)| \\ &\leq |(a - a_k)| + |(b - b_k)| \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

Here, we have used the triangle inequality for the absolute value function. Obviously, the difference of two sequences also converges to the difference of their limit points.

As for multiplication, again take the convergent sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ as before. We have $\lim_{n\to\infty} a_n = a$ and $\lim_{n\to\infty} b_n = b$. Now let $M_a > 0$ be such that $|a|$ and $|a_n| \leq M_a$, for all $n \in \mathbb{N}$. Also let $M_b > 0$ be such that $|b|$ and $|b_m| \leq M_b$, for all $m \in \mathbb{N}$. (These numbers must exist, since convergent sequences are bounded.) Then, given $\epsilon > 0$, choose $N_a(\epsilon)$ such that for all $n \geq N_a(\epsilon)$ we have

$$|a - a_n| < \frac{\epsilon}{2M_b}.$$

Similarly, $N_b(\epsilon)$ is chosen such that for all $m \geq N_b(\epsilon)$ we have

$$|b - b_n| < \frac{\epsilon}{2M_a}.$$

Then take $N(\epsilon) = \max\{N_a(\epsilon), N_b(\epsilon)\}$. So again, For any $k \geq N(\epsilon)$ we have

$$\begin{aligned} |a \cdot b - a_k \cdot b_k| &= |a \cdot b - a \cdot b_k + a \cdot b_k - a_k b_k| \\ &\leq |a \cdot b - a \cdot b_k| + |a \cdot b_k - a_k b_k| \\ &= |a||b - b_k| + |b_k||a - a_k| \\ &< |a|\frac{\epsilon}{2M_a} + |b_k|\frac{\epsilon}{2M_b} \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

Finally, assume that $(a_n)_{n\in\mathbb{N}}$ is a convergent sequence such that the limit a is not zero. Then the sequence $(1/a_n)_{n\in\mathbb{N}}$ (at most finitely many elements of the sequence can be zero, and so we disregard these zero elements) converges to $1/a$. In order to see this, let $M > 0$ be a lower bound of the sequence of absolute values $(|a_n|)_{n\in\mathbb{N}}$, together with $|a|$. Given $\epsilon > 0$, this time choose $N(\epsilon) \in \mathbb{N}$ to be so large that for all $n \geq N(\epsilon)$, we have $|a - a_n| < \epsilon M^2$. Then

$$\begin{aligned} \left|\frac{1}{a} - \frac{1}{a_n}\right| &= \left|\frac{a - a_n}{aa_n}\right| \\ &= \frac{1}{|aa_n|}|a - a_n| \\ &< \frac{\epsilon M^2}{|a||a_n|} \leq \epsilon. \end{aligned}$$

Then, in order to divide a convergent sequence by a convergent sequence which does not converge to zero, we first take the convergent sequence of the inverses, then multiply with that.

In summary, we have

Theorem 2.9. *Convergent series can be added, subtracted, multiplied and divided (as long as they do not converge to zero), to obtain new convergent sequences which converge to the sum, difference, product, and quotient of the limits of the given sequences.*

2.4 Convergent series

Given a sequence $(a_n)_{n\in\mathbb{N}}$, we can imagine trying to find the sum of all the numbers in the sequence. Thus writing

$$\sum_{n=1}^{\infty} a_n,$$

we have the *series* given by the sequence $(a_n)_{n\in\mathbb{N}}$. Obviously, many series do *not* converge. For example the series

$$\sum_{n=1}^{\infty} n = 1+2+3+4+5+6+7+\cdots$$

does not converge. Also the series

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

does not converge. Why is this?

Definition. *Given the series $\sum_{n=1}^{\infty} a_n$, the n-th* partial sum *(for each $n \in \mathbb{N}$) is the finite sum*

$$S_n = \sum_{k=1}^{n} a_n.$$

The series $\sum_{n=1}^{\infty} a_n$ converges, if the sequence of its partial sums $(S_n)_{n\in\mathbb{N}}$ converges. If the series does not converge, then one says that it diverges.

So what are the partial sums for the series $\sum_{n=1}^{\infty}(-1)^n$? Clearly, we have

$$S_n = \begin{cases} -1, & \text{if } n \text{ is odd,} \\ 0, & \text{if } n \text{ is even.} \end{cases}$$

Therefore, the partial sums jump back and forth between -1 and 0, never converging.

A delicate case: the series $\sum_{n=1}^{\infty} 1/n$

But what about the series

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\cdots$$

Obviously the partial sums get larger and larger: $S_{n+1} > S_n$, for all $n \in \mathbb{N}$. But the *growth* of the sequence of partial sums keeps slowing down. So one might think that this series could converge. But does it?

In fact, it actually *diverges*. We can see this by looking at the sum split into blocks of ever increasing length.

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1+\frac{1}{2}+\underbrace{\frac{1}{3}+\frac{1}{4}}_{>1/2}+\underbrace{\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}}_{>1/2}+\cdots$$

That is to say, for each $n \in \mathbb{N}$, we have

$$\sum_{k=2^{n-1}+1}^{2^n} \frac{1}{k} > \sum_{k=2^{n-1}+1}^{2^n} \frac{1}{2^n} = \frac{1}{2},$$

so we have an infinite series of blocks, each greater than $1/2$. Therefore it must diverge.

The geometric series

This is the series

$$\sum_{n=0}^{\infty} a^n,$$

for various possible numbers $a \in \mathbb{R}$. (Note that it is sometimes convenient to take the sum from 0 to infinity, rather than from 1 to infinity. Also note that by convention, we always define $a^0 = 1$, even in the case that $a = 0$.)

Theorem 2.10. *For all real numbers a with $|a| < 1$, the sequence $(a^n)_{n\in\mathbb{N}}$ converges to zero. For $|a| \geq 1$, the sequence diverges.*

Proof. Without loss of generality, we may assume that $a > 0$. If $a < 1$ then the sequence $(a^n)_{n\in\mathbb{N}}$ is a *strictly decreasing* sequence. That is, $a^{n+1} < a^n$, for all $n \in \mathbb{N}$. This follows, since $a^{n+1} = a \cdot a^n$, and $0 < a < 1$.

So the sequence $(a^n)_{n\in\mathbb{N}}$ gets smaller and smaller, as n gets bigger. And of course, it starts with a, so it is confined to the interval between 0 and a. We can define a Dedekind cut (A, B) of $\mathbb{Q}$ as follows.

$$A^* = \{x \in \mathbb{Q} : x < a^n, \forall n \in \mathbb{N}\},$$

and $B^* = \mathbb{Q} \setminus A^*$ (the set difference). Finally, if A^* has a greatest element, say x_0, then take $A = A^* \setminus \{x_0\}$ and $B = B^* \cup \{x_0\}$. Otherwise simply take $A = A^*$ and $B = B^*$. The pair (A, B) is then a Dedekind cut.

So let ξ be the real number represented by this Dedekind cut. Then we must have $0 \leq \xi < 1$. If $\xi = 0$ then the sequence converges to zero, and we are finished. Otherwise, we must have $\xi > 0$. Now since $0 < a < 1$, it must be that the number $1/a$ is greater than 1. Thus

$$\xi < \xi \cdot \frac{1}{a}.$$

But from the definition of ξ, there must be some $m \in \mathbb{N}$ with

$$\xi < a^m < \xi \cdot \frac{1}{a}.$$

However, then we have

$$a^{m+1} = a \cdot a^m < a \cdot \xi \cdot \frac{1}{a} = \xi,$$

and this contradicts the definition of ξ. Therefore the idea that we might have $\xi > 0$ simply leads to a contradiction. The only conclusion is that $\xi = 0$, and so the sequence converges.

If $a > 1$, then, using what we have just proved, we see that the sequence $\left(\frac{1}{a^n}\right)_{n\in\mathbb{N}}$ converges to zero. Clearly, this implies that $(a^n)_{n\in\mathbb{N}}$ diverges (or, in this case, "converges to infinity"). □

Theorem 2.11. *The geometric series converges for $|a| < 1$, and it diverges for $|a| \geq 1$.*

Proof. We have

$$(a-1)\left(\sum_{k=0}^{n} a^k\right) = a^{n+1} - 1.$$

Therefore, if $a \neq 1$, we have

$$\sum_{k=0}^{n} a^k = \frac{a^{n+1}-1}{a-1},$$

for all $n \in \mathbb{N}$.

For $|a| < 1$, we know that the sequence $(a^n)_{n\in\mathbb{N}}$ converges to zero. Thus $\sum_{n=1}^{\infty} a^n$ is a convergent series for $0 < a < 1$, and we have

$$\sum_{n=0}^{\infty} a^n = \frac{-1}{a-1} = \frac{1}{1-a}.$$

If $|a| > 1$, then the series diverges since $\sum_{k=1}^{n} a^k = \frac{a^{n+1}-1}{a-1}$, and the sequence $(a^n)_{n\in\mathbb{N}}$ diverges. □

2.5 The standard tests for convergence of a series

2.5.1 The Leibniz test

Theorem 2.12. *Let $(a_n)_{n\in\mathbb{N}}$ be a decreasing sequence of numbers, that is, $a_{n+1} \leq a_n$, for all n, such that the sequence converges, with $\lim_{n\to\infty} a_n = 0$. Then the* alternating *series*

$$\sum_{n=1}^{\infty} (-1)^n a_n$$

converges.

Proof. Consider the partial sums S_n for this series. If $a_1 \neq 0$, then $S_1 = (-1)a_1$ is a negative number. But then $S_3 = -a_1 + (a_2 - a_3)$, and we see that we must have $S_1 \leq S_3$ since $a_2 \geq a_3$, and therefore $a_2 - a_3$ is a positive number or zero. More generally, if n is an odd number, say $n = 2m+1$, then we must have $S_{n+2} \geq S_n$. This follows, since

$$\begin{aligned} S_{n+2} &= S_n + (-1)^{n+1}a_{n+1} + (-1)^{n+2}a_{n+2} \\ &= S_n + (-1)^{(2m+1)+1}a_{n+1} + (-1)^{(2m+1)+2}a_{n+2} \\ &= S_n + (a_{n+1} - a_{n+2}), \end{aligned}$$

and $a_{n+1} - a_{n+2} \geq 0$. Therefore the sequence of odd partial sums is an increasing sequence.

$$S_1 \leq S_3 \leq S_5 \leq S_7 \leq \cdots$$

On the other hand, we have that the sequence of even partial sums is a *decreasing* sequence.

$$S_2 \geq S_4 \geq S_6 \geq S_8 \geq \cdots$$

This is proved analogously to the situation with the odd partial sums. Furthermore, it is easy to see that

$$S_{2m} \geq S_{2m+1},$$

and

$$S_{2m+1} \leq S_{2m+2},$$

for all $m \in \mathbb{N}$. Therefore the even partial sums are always greater than, or equal to, the odd partial sums. On the other hand, the distance between adjacent partial sums is $|S_{n+1} - S_n| = |a_{n+1}|$, and we know that $\lim_{n\to\infty} |a_n| = 0$. Thus the even and the odd sums must converge from above and below to some common limit point, which is then the limit of the series. □

An example

We have already seen that the series $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges. But according to Leibniz test, the alternating series

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

must converge. In fact, if we write $T = \sum_{n=1}^{\infty} \frac{(-1)^n}{n}$, then we know from the proof of theorem 2.12 that T must lie somewhere between the first and the second partial sums. That is

$$S_1 = -1 < T < -\frac{1}{2} = -1 + \frac{1}{2} = S_2.$$

In other words, the sum of the whole series is a negative number lying somewhere between -1 and $-\frac{1}{2}$.

Reordering the terms in the series

While all of what has been said above is true, there is a strange twist to the story which makes one realize that it is important to be careful.

To begin with, note that we have the following.

$$\begin{array}{lcl}
\frac{1}{4} & < & \frac{1}{2} + \frac{1}{4} \\
\frac{1}{4} & < & \frac{1}{6} + \frac{1}{8} \\
\frac{1}{4} & < & \frac{1}{10} + \frac{1}{12} + \frac{1}{14} + \frac{1}{16} \\
\frac{1}{4} & < & \frac{1}{18} + \frac{1}{20} + \frac{1}{22} + \frac{1}{24} + \frac{1}{26} + \frac{1}{28} + \frac{1}{30} + \frac{1}{32} \\
& \textit{etc.} &
\end{array}$$

Therefore, if we rearrange the terms in the sum, we get

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{(-1)^n}{n} \overset{???}{=}\ & -1+\left(\frac{1}{2}+\frac{1}{4}\right)\\ & -\frac{1}{3}+\left(\frac{1}{6}+\frac{1}{8}\right)\\ & -\frac{1}{5}+\left(\frac{1}{10}+\frac{1}{12}+\frac{1}{14}+\frac{1}{16}\right)\\ & -\frac{1}{7}+\left(\frac{1}{18}+\frac{1}{20}+\frac{1}{22}+\frac{1}{24}+\frac{1}{26}+\frac{1}{28}+\frac{1}{30}+\frac{1}{32}\right)\\ & -\frac{1}{9}+\quad \textit{etc.}\end{aligned}$$

Obviously the sum is getting bigger and bigger. It suddenly doesn't converge! The problem is that our original sum is convergent, but not *absolutely* convergent. It is only *conditionally* convergent. Conditionally convergent series can be made to converge to practically *anything* — or else they can be made to diverge — if we allow ourselves to rearrange the order of the terms in the sum in any way we want.

But let's look at the other convergence tests, before coming back to this problem.

2.5.2 The comparison test

Theorem 2.13. *Let*

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

be a series which is known *to be convergent, where* $c_n \geq 0$*, for all* n*. Furthermore, let*

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

be some other series, where $0 \leq a_n \leq c_n$*, for all* n*. Then the series* $\sum_{n=1}^{\infty} a_n$ *is convergent, and the limit of the series is no greater than the limit of the series* $\sum_{n=1}^{\infty} c_n$*.*

Proof. This is obvious. Let S_n be the n-th partial sum of the series $\sum_{n=1}^{\infty} a_n$, and let

$$\sum_{n=1}^{\infty} c_n = C,$$

say. Then we have that the sequence of partial sums $(S_n)_{n\in\mathbb{N}}$ is monotonically increasing, and it is bounded below by zero, and above by C. Thus it must converge to a number between zero and C. □

2.5.3 Absolute convergence

Definition. *The series*

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

is called absolutely convergent *if the series consisting of the absolute values of the individual terms*

$$\sum_{n=1}^{\infty} |a_n|$$

converges.

Theorem 2.14. *Each series which is absolutely convergent is also convergent.*

Proof. Assume that the series $\sum_{n=1}^{\infty} |a_n|$ converges, where $a_n \in \mathbb{R}$ for all n. Let

$$\sum_{n=1}^{\infty} |a_n| = C,$$

say, and let S_n^* be the partial sums of this series. Since $|a_n| \geq 0$ for all n, it must be that the sequence $(S_n^*)_{n\in\mathbb{N}}$ is monotonically increasing. Therefore, for each $\epsilon > 0$, there exists some $N(\epsilon) \in \mathbb{N}$ such that $|C - S_n^*| < \epsilon$, for all $n \geq N(\epsilon)$. But then, in particular, we must have $|S_n^* - S_m^*| < \epsilon$, for all n, $m \geq N(\epsilon)$. But (assuming that $m \leq n$), we have

$$|S_n^* - S_m^*| = \sum_{k=m+1}^{n} |a_k| < \epsilon.$$

So now we can show that the sequence of partial sums S_n for the original series $\sum_{n=1}^{\infty} a_n$ is a Cauchy sequence. For all n, $m \geq N(\epsilon)$ (and again, we assume without loss of generality that $m \leq n$) we have

$$\begin{aligned} |S_n - S_m| &= \left| \sum_{k=m+1}^{n} a_k \right| \\ &\leq \sum_{k=m+1}^{n} |a_k| \\ &< \epsilon. \end{aligned}$$

The first inequality here is simply the triangle inequality for the absolute-value function, and the second inequality is $|S_n^* - S_m^*| < \epsilon$, which we have already found. □

Corollary (Majorantenkriterium). *Let $(c_n)_{n\in\mathbb{N}}$ be sequence with $c_n \geq 0$, for all n, such that*

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

converges. Then given another sequence $(a_n)_{n\in\mathbb{N}}$*, with* $|a_n| \leq c_n$ *for all* n*, we must have the series*

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

also converging.

Theorem 2.15. *Let* $\sum_{n=1}^{\infty} a_n$ *be an absolutely convergent series, and let* $\sum_{n=1}^{\infty} b_n$ *be the same series, but with the terms possibly rearranged in some way. Then* $\sum_{n=1}^{\infty} b_n$ *is also absolutely convergent, and we have*

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} b_n.$$

But first we prove the following lemma.

Lemma. *Let* $\sum_{n=1}^{\infty} c_n$ *be a convergent series with* $c_n \geq 0$ *for all* n*. If* $\sum_{n=1}^{\infty} d_n$ *is the same series, but perhaps with the terms rearranged in some other order, then we still have* $\sum_{n=1}^{\infty} d_n$ *being convergent, and*

$$\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{\infty} d_n.$$

Proof. In both cases, the sequence of partial sums is monotonically increasing. Given the partial sum $\sum_{n=1}^{N_1} c_n$, for some $N_1 \in \mathbb{N}$, then we can find some $N_2 \geq N_1$ which is sufficiently large that all the numbers $c_1, \ldots, c_{N_1}$ appear in the list $d_1, \ldots, d_{N_2}$. Therefore we must have

$$\sum_{n=1}^{N_1} c_n \leq \sum_{n=1}^{N_2} d_n.$$

But we can just as easily show that for all $N_3 \in \mathbb{N}$, there exists some $N_4 \geq N_3$ with

$$\sum_{n=1}^{N_4} c_n \geq \sum_{n=1}^{N_3} d_n.$$

Therefore we must have the limits of the sequences of partial sums being equal. □

Proof. (Of theorem 2.15)

Let

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} a_n^+ - \sum_{n=1}^{\infty} a_n^-,$$

where

$$a_n^+ = \begin{cases} a_n, & \text{if } a_n \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

$$a_n^- = \begin{cases} -a_n, & \text{if } a_n \leq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Similarly,

$$\sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty} b_n^+ - \sum_{n=1}^{\infty} b_n^-.$$

But, according to the lemma, we must have

$$\sum_{n=1}^{\infty} a_n^+ = \sum_{n=1}^{\infty} b_n^+$$

and

$$\sum_{n=1}^{\infty} a_n^- = \sum_{n=1}^{\infty} b_n^-.$$

□

2.5.4 The quotient test

Theorem 2.16. *Assume that the series $\sum_{n=1}^{\infty} a_n$ is such that there exists some real number $\xi \in \mathbb{R}$ with $0 \leq \xi < 1$, such that*

$$\left| \frac{a_{n+1}}{a_n} \right| \leq \xi,$$

for all $n \in \mathbb{N}$. Then the series is absolutely convergent, hence also convergent.

Proof. We have already seen that the geometric series

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

converges. So if we simply multiply each term by the number $|a_1|$, we see that also the series

$$\sum_{n=1}^{\infty} |a_1| \xi^n$$

converges. In fact it converges to the number

$$|a_1| \left(\sum_{n=1}^{\infty} \xi^n \right).$$

Now since $|a_2/a_1| \leq \xi$, we must have $|a_2| \leq |a_1|\xi$. Also, since $|a_3/a_2| \leq \xi$, we must have $|a_3| \leq |a_2|\xi$. That is, $|a_3| \leq |a_2|\xi \leq |a_1|\xi^2$. Similarly, we have $|a_4| \leq |a_1|\xi^3$, and in general, for each n, we have

$$|a_n| \leq |a_1|\xi^{n-1}.$$

Therefore, using the comparison test, we see that the series

$$\sum_{n=1}^{\infty} |a_n|$$

converges. □

Corollary. *Let $N \in \mathbb{N}$ be given, and we assume that the series $\sum_{n=1}^{\infty} a_n$ is such that there exists some real number $\xi \in \mathbb{R}$ with $0 \leq \xi < 1$, such that*

$$\left| \frac{a_{n+1}}{a_n} \right| \leq \xi,$$

for all $n \geq N$. Then the series is absolutely convergent, hence also convergent.

Proof. This follows, since the argument in the proof of the previous theorem can be applied to the numbers greater than or equal to N. So the series

$$\sum_{n=N}^{\infty} a_n$$

is absolutely convergent. However we can then simply add in the finitely many terms

$$a_1 + a_2 + \cdots + a_{N-1},$$

and this cannot change the fact that the whole infinite series is absolutely convergent. □

Example: the exponential series is convergent everywhere

Rather than always taking the sum in a series from 1 to ∞, it is often convenient to sum from 0 to ∞. In particular, for each $x \in \mathbb{R}$ we have the famous *exponential series*

$$\sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

Using the quotient test, it is easy to see that the exponential series is absolutely convergent, for all $x \in \mathbb{R}$.

For let some arbitrary $x \in \mathbb{R}$ be given. Now if we happen to have $x = 0$, then the exponential series is obviously absolutely convergent. Therefore we assume that $x \neq 0$. Then let $N \in \mathbb{N}$ be the smallest integer with $N \geq |x|$. We have

$$\left| \frac{\frac{x^{n+1}}{(n+1)!}}{\frac{x^n}{n!}} \right| = \left| \frac{x}{n+1} \right| \leq \left| \frac{x}{N+1} \right| < 1,$$

for all $n \geq N$, and it follows that the exponential series must be absolutely convergent in this case as well.

2.6 Continuous functions

Let $A \subset \mathbb{R}$ be some interval. For example we might have $A = [a, b]$, for $a < b$. That is the *closed* interval from a to b. The open interval from a to b is $(a, b) = \{x \in \mathbb{R} : a < x < b\}$. Then we have the half closed, and half open intervals $[a, b)$

and $(a, b]$. We can also consider the whole of $\mathbb{R}$ to be the interval $(-\infty, \infty)$. That is also an open interval. For most of the time, we will consider functions

$$f : A \to \mathbb{R}$$

from some *open* interval $A \subset \mathbb{R}$ into $\mathbb{R}$.

Definition. *The function $f : A \to \mathbb{R}$ is* continuous *in the point $x_0 \in A$ if for all $\epsilon > 0$, there exists some $\delta > 0$ such that $|f(x) - f(x_0)| < \epsilon$, for all $x \in A$ with $|x - x_0| < \delta$. If the function f is continuous in x_0 for* all *$x_0 \in A$, then one simply says that f is* continuous.

Examples

For these examples, we consider in each case a function $f : \mathbb{R} \to \mathbb{R}$. That is, our open interval is $A = \mathbb{R}$. We will specify f by specifying what $f(x)$ is, for each $x \in \mathbb{R}$.

- If there exists some constant number $c \in \mathbb{R}$, such that $f(x) = c$, for all $x \in \mathbb{R}$, then f is a *constant function*. Obviously, f is then continuous.
- If $f(x) = x$ for all x, then f is continuous. For let $x_0 \in \mathbb{R}$ be some arbitrary real number. Let $\epsilon > 0$ be given. Then choose $\delta = \epsilon$. With this choice, if we have $x \in \mathbb{R}$ with $|x - x_0| < \delta = \epsilon$, then we must have $|f(x) - f(x_0)| = |x - x_0| < \delta = \epsilon$. Therefore f is continuous in x_0, and since x_0 was arbitrary, f is continuous everywhere.
- If $f(x) = x^n$, for some $n \in \mathbb{N}$ larger than one, then f is also continuous. This is not quite as trivial to prove, so we will put off the proof till later.
- This time let
$$f(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases}$$
Then f is continuous for all $x_0 \neq 0$, but f is *not* continuous at the point 0.

An alternative way to describe continuity

Theorem 2.17. *The function $f : A \to \mathbb{R}$ is continuous in the point $x_0 \in A$ if and only if for all convergent sequences $(a_n)_{n\in\mathbb{N}}$ with $a_n \in A$ for all n, and $\lim_{n\to\infty} a_n = x_0$, we have that $(f(a_n))_{n\in\mathbb{N}}$ is a convergent sequence with $lim_{n\to\infty} f(a_n) = f(x_0)$.*

Proof. Assume first that f is continuous at $x_0 \in A$. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence with $a_n \in A$ for all n, and $\lim_{n\to\infty} a_n = x_0$. That means that for all $\delta > 0$, there exists some $N(\delta) \in \mathbb{N}$ with $|x_0 - a_n| < \delta$ for all $n \geq N(\delta)$. Now let $\epsilon > 0$ be given. Since f is assumed to be continuous at x_0, there must exist some $\delta > 0$ with $|f(x) - f(x_0)| < \epsilon$, for all $x \in A$ with $|x - x_0| < \delta$. Therefore, given our $N(\delta)$,

we must have $|x_0 - a_n| < \delta$ for all $n \geq N(\delta)$. That means that for all $n \geq N(\delta)$ we have $|f(x) - f(x_0)| < \epsilon$. Therefore $\lim_{n\to\infty} f(a_n) = f(x_0)$.

Now assume that $\lim_{n\to\infty} f(a_n) = f(x_0)$ for all convergent sequences $(a_n)_{n\in\mathbb{N}}$ in A with $\lim_{n\to\infty} a_n = x_0$. In order to obtain a contradiction, assume furthermore that f is *not* continuous at x_0. That would mean that there must exist some $\epsilon_0 > 0$, such that for all $\delta > 0$ some $u_\delta \in A$ must exist with $|x_0 - u_\delta| < \delta$, yet $|f(x_0) - f(u_\delta)| \geq \epsilon_0$. In particular, we can progressively take $\delta = 1/n$, for $n = 1, 2, 3, \ldots$.

That is, we take the sequence $(a_n)_{n\in\mathbb{N}}$ with $a_n = u_{\frac{1}{n}}$, for all n. Then we have $\lim_{n\to\infty} a_n = x_0$, yet $|f(x_0) - f(a_n)| \geq \epsilon_0$, for all n. Therefore the series $(f(a_n))_{n\in\mathbb{N}}$ cannot possibly converge to $f(x_0)$. This contradicts our assumption. □

2.6.1 Sums, products, and quotients of continuous functions are continuous

Theorem 2.18. *Assume that $f, g : A \to \mathbb{R}$ are two continuous functions from A to $\mathbb{R}$. Then $f + g$ is also continuous. Here, $f + g$ is the function whose value at each $x \in A$ is simply $(f + g)(x) = f(x) + g(x)$.*

Proof. Let $x_0 \in A$ be given. The problem then is to show that the function $f + g$ is continuous at x_0. For this we will use theorem 2.17. Let $(a_n)_{n\in\mathbb{N}}$ be some convergent sequence in A with $\lim_{n\to\infty} a_n = x_0$. Then, since f is continuous at x_0, we have $\lim_{n\to\infty} f(a_n) = f(x_0)$. Similarly, we have $\lim_{n\to\infty} g(a_n) = g(x_0)$. But then, according to theorem 2.9, the series

$$(f(a_n) + g(a_n))_{n\in\mathbb{N}}$$

converges to $f(x_0) + g(x_0) = (f + g)(x_0)$. Therefore, again according to theorem 2.17, the function $f + g$ must be continuous at x_0. □

Of course, this also means that the difference of two continuous functions $f - g$ is also continuous.

Theorem 2.19. *The functions f and g are given as before. Then also their product $f \cdot g$ is continuous. Here, the product is simply the function whose value at $x \in A$ is given by $(f \cdot g)(x) = f(x) \cdot g(x)$, for all such x.*

Proof. The same as for theorem 2.18 □

Similarly we have

Theorem 2.20. *The functions f and g are given as before, where we assume that $g(x) \neq 0$, for all $x \in A$. Then the quotient f/g is continuous, where the quotient is the function whose value at $x \in A$ is given by $(f/g)(x) = f(x)/g(x)$, for all $x \in A$.*

Theorem 2.21. *Assume that $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$, and we have two functions $f : A \to \mathbb{R}$ and $g : B \to \mathbb{R}$ such that $f(A) \subset B$. We can consider the function $f \circ g : A \to \mathbb{R}$, where $f \circ g(x) = g(f(x))$, for all $x \in A$. Then if f is continuous at $x_0 \in A$, and g is continuous at $f(x_0)$, it follows that $f \circ g$ is continuous at x_0.*

Proof. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence in A, converging to x_0. Then, since f is continuous at x_0, the sequence $(f(a_n))_{n\in\mathbb{N}}$ must converge to $f(x_0)$ in B. But then since g is continuous at $f(x_0)$, the sequence $(g(f(a_n)))_{n\in\mathbb{N}}$ must converge to $g(f(x_0)) = f \circ g(x_0)$. □

All polynomials are continuous

This is now obvious. Let

$$f(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n$$

be some polynomial. Then, as we have seen, the constant function c_0 is continuous. Also the identity function $x \to x$ is continuous. Therefore the product $c_1 x$ gives a continuous function. Also the product $x \cdot x = x^2$ is a product of two continuous functions, therefore continuous. So $c_1 x^2$ is continuous. And so forth. Finally the polynomial is seen to be just a sum of continuous functions, therefore itself continuous.

2.7 The exponential function

We have already seen that the series

$$\sum_{n=0}^{\infty} \frac{x^n}{n!}$$

converges for all $x \in \mathbb{R}$. This gives the exponential function

Definition. *The exponential function* $\exp(x)$*, often written* e^x*, is defined for real numbers $x \in \mathbb{R}$ to be* $\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$*. The defining series here is called the* exponential series*.*

Obviously, by looking at the exponential series, we see that $\exp(0) = 1$. But what is $\exp(x)$ for other values of x? Let us take another look at the exponential series and then think about the following points.

- As already seen, we have $\exp(0) = 1$.
- For $x > 0$ we must have $\exp(x) > 0$ since all terms in the exponential series are positive.

- In fact, if we have two positive numbers $0 < x < y$, then we must have $1 < \exp(x) < \exp(y)$. This follows, since we must have $x^n < y^n$, for all n; therefore the exponential series for y dominates the exponential series for x. Therefore, for non-negative real numbers, we see that the exponential function is a strictly monotonically increasing function.
- But for negative numbers $x < 0$, the situation remains unclear.

Theorem 2.22. *For all x and $y \in \mathbb{R}$ we have* $\exp(x+y) = \exp(x) \cdot \exp(y)$.

Proof.

$$\begin{aligned}
\exp(x+y) &= \sum_{n=0}^{\infty} \frac{(x+y)^n}{n!} \\
&= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k \\
&= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{n} \frac{n!}{(n-k)!k!} x^{n-k} y^k \\
&= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{x^{n-k} y^k}{(n-k)!k!} \\
&= \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} \frac{x^{n-k}}{(n-k)!} \cdot \frac{y^k}{k!} \right) \\
&= \left(\sum_{n=0}^{\infty} \frac{x^n}{n!} \right) \cdot \left(\sum_{n=0}^{\infty} \frac{y^n}{n!} \right) \\
&= \exp(x) \cdot \exp(y)
\end{aligned}$$

Note here that:

- The second equation is our Binomial theorem (theorem 1.4).
- The sixth equation is our Exercise 7.1.
- The other equations are nothing but the definitions of the various things, and simple arithmetic operations.

□

Consequences of this "functional equation" for the exponential function

- Let $x < 0$ be a negative number. Then we know that $-x$ is a positive number, and thus $\exp(-x) > 0$. But also

  $$\exp(x)\exp(-x) = \exp(x-x) = \exp(0) = 1.$$

  Therefore it follows that

  $$\exp(x) = \frac{1}{\exp(-x)} > 0$$

  and we see that $\exp(x) > 0$ for *all* real numbers $x \in \mathbb{R}$.

- In fact, if $x < y < 0$ then we have

$$\exp(y) - \exp(x) = \frac{1}{\exp(-y)} - \frac{1}{\exp(-x)} = \frac{\exp(-x) - \exp(-y)}{\exp(-y)\exp(-x)} > 0,$$

since the exponential function is strictly monotonically increasing, and $-x > -y$.

Therefore, the exponential function is strictly monotonically increasing for *all* $\mathbb{R}$.

- Let $x \in \mathbb{R}$ be a real number with $0 < x < 1$. Then the sequence

$$\left(\frac{x^n}{n!}\right)_{n\in\mathbb{N}}$$

is a strictly decreasing sequence of positive real numbers. Therefore, looking at the proof of Leibniz test (theorem 2.12), we see that the exponential series for $-x$ must converge to some real number between $1 - x$ and $1 - x + x^2/2$. That is, we must have

$$1 - x < \exp(-x) < 1 - x + \frac{x^2}{2} < 1,$$

or

$$0 < 1 - \exp(-x) < x.$$

In particular, given any real number $y \in (-1, 0)$, then we must have

$$|\exp(y) - \exp(0)| < |y|.$$

- On the other hand, if x is a positive number with $x \in (0, 1/2)$, then we must have

$$|\exp(x) - \exp(0)| = \left|\frac{1}{\exp(-x)} - 1\right| < \left|\frac{1}{1-x} - 1\right| = \left|\frac{x}{1-x}\right| < \left|\frac{x}{\frac{1}{2}}\right| = 2|x|.$$

- Putting these two things together, we have that if $|x| < 1/2$, that is $|x - 0| < 1/2$, then $|\exp(x) - \exp(0)| < 2|x|$. Therefore, the exponential function must be continuous at the point $0 \in \mathbb{R}$.

- Finally, take any other element $y \in \mathbb{R}$. Let $(y_n)_{n\in\mathbb{N}}$ be some convergent sequence, with $\lim_{n\to\infty} y_n = y$. Then if we take $z_n = y_n - y$ for all n, we must have that $(z_n)_{n\in\mathbb{N}}$ is a convergent sequence, with

$$\lim_{n\to\infty} z_n = 0.$$

Therefore, since the exponential function is continuous at 0, we must have $(\exp(z_n))_{n\in\mathbb{N}}$ being a convergent sequence, with

$$\lim_{n\to\infty} \exp(z_n) = \exp(0) = 1.$$

But

$$\begin{aligned} 1 = \lim_{n\to\infty} \exp(z_n) &= \lim_{n\to\infty} \exp(y_n - y) \\ &= \lim_{n\to\infty} \exp(y_n) \cdot \exp(-y) \\ &= \lim_{n\to\infty} \frac{\exp(y_n)}{\exp(y)} \\ &= \frac{1}{\exp(y)} \lim_{n\to\infty} \exp(y_n), \end{aligned}$$

since $\exp(y)$ is constant (that is, independent of the number n). Therefore, in the end we have

$$\lim_{n\to\infty} \exp(y_n) = \exp(y),$$

and it follows that the exponential function is also continuous at y.

So to summarize all of this, we have shown that:

Theorem 2.23. *The exponential function is strictly monotonically increasing, positive, continuous, with* $\exp(-x) = \frac{1}{\exp(x)}$*, for all* $x \in \mathbb{R}$*. Therefore, also* $\exp(0) = 1$*.*

2.8 Some general theorems concerning continuous functions

So now that we have seen the standard examples of continuous functions — namely the polynomials and the exponential function[4] — it is time to look at some of the theorems which show us why the idea of continuity is so important.

Theorem 2.24. *Let* $[a, b] \subset \mathbb{R}$ *be a closed interval, and let* $f : [a, b] \to \mathbb{R}$ *be continuous. Then* f *is bounded (that is, the set* $f([a, b]) = \{f(x) : x \in [a, b]\}$ *is bounded), and in fact, there exists both an* $x_* \in [a, b]$ *such that* $f(x_*) = \sup\{f([a, b])\}$*, and also there exists* $y_* \in [a, b]$ *such that* $f(y_*) = \inf\{f([a, b])\}$*.*

Proof. If f were not bounded, then it is either unbounded above, or below. Let us assume that it is unbounded above, so that for every $n \in \mathbb{N}$, there exists some $x_n \in [a, b]$, such that $f(x_n) > n$. Therefore, $(f(x_n))_{n\in\mathbb{N}}$ is a sequence in $\mathbb{R}$ which can have no convergent subsequences. On the other hand, $(x_n)_{n\in\mathbb{N}}$ is a bounded sequence in $\mathbb{R}$, therefore it contains a convergent subsequence (theorem 2.4). So let $(x_{i(n)})_{n\in\mathbb{N}}$ be such a convergent subsequence, with

$$\lim_{n\to\infty} x_{i(n)} = x_* \in [a, b],$$

[4]The other "standard functions" like sin, cos, ln, and so forth, are simply defined in terms of the exponential function. So, at least in principle, we now have all of them.

say. Then, since f is continuous at x_*, we must have that the subsequence $(f(x_i(n)))_{n\in\mathbb{N}}$ is also convergent, with

$$\lim_{n\to\infty} f(x_{i(n)}) = f(x_*).$$

This is a contradiction, and so we must conclude that f is bounded.

Next, let us consider the number $\sup\{f([a,b])\}$. Since it is the *least* upper bound, it must be that for each $n \in \mathbb{N}$, we can choose some $x_n \in [a,b]$ with

$$|\sup\{f([a,b])\} - f(x_n)| < \frac{1}{n}.$$

Therefore, not only does the sequence $(f(x_n))_{n\in\mathbb{N}}$ converge to $\sup\{f([a,b])\}$, in fact, *every* subsequence must also converge to $\sup\{f([a,b])\}$. But, considered in $[a,b]$, we have that $(x_n)_{n\in\mathbb{N}}$ is a bounded sequence; therefore there is a convergent subsequence $(x_{i(n)})_{n\in\mathbb{N}}$, with

$$\lim_{n\to\infty} x_{i(n)} = x_* \in [a,b],$$

say. Then since f is continuous at x_*, we must have

$$f(x_*) = \lim_{n\to\infty} f(x_{i(n)}) = \sup\{f([a,b])\}.$$

The proof with regard to $\inf\{f([a,b])\}$ is analogous. □

But be careful! Here is *almost* a counterexample.

The function $f : (0,1) \to \mathbb{R}$, with

$$f(x) = \frac{1}{x}$$

is obviously continuous everywhere in $(0,1)$. Yet it is unbounded! Why can't we apply our theorem 2.24 here? The answer is that we can indeed construct a sequence $(x_n)_{n\in\mathbb{N}}$ such that the sequence $(f(x_n))_{n\in\mathbb{N}}$ increases without bound. But in this case, we will simply have

$$\lim_{n\to\infty} x_n = 0,$$

but $0 \notin (0,1)$, therefore *the sequence does not converge* when considered as a sequence taken within the set $(0,1)$.

Theorem 2.25 (Intermediate value theorem, or "Zwischenwertsatz")**.** *Let $f : [a,b] \to \mathbb{R}$ be a continuous function, such that $f(a)f(b) < 0$. (That is, both $f(a) \neq 0$ and $f(b) \neq 0$, and furthermore one is positive and the other is negative.) Then there exists some point $x \in (a,b)$, such that $f(x) = 0$.*

Proof. Let $x_1 = (b-a)/2$ be the half-way point between a and b. If $f(x_1) = 0$, then we have a solution. Otherwise, $f(x_1) \neq 0$, and so either $f(a)$ and $f(x_1)$ have opposite signs, or else $f(x_1)$ and $f(b)$ have opposite signs. In any case, the original interval $[a, b]$ can be divided into two smaller sub-intervals $[a, x_1]$ and $[x_1, b]$, both of which are only half as big as the original interval. Choose the sub-interval which is such that the endpoints have opposite signs under f. Then subdivide that subinterval in half. And so on.

In the end, either we end up with a solution, or else, by taking say the upper endpoint of each sub-interval, we obtain a convergent sequence $(y_n)_{n\in\mathbb{N}}$. Let $\lim_{n\to\infty} y_n = y$. Then there are both positive, as well as negative values of f arbitrarily near to $f(y)$. Since f is continuous, we must then have $f(y) = 0$. □

Definition. *Let $W \subset \mathbb{R}$ be some subset of $\mathbb{R}$. The function $f : W \to \mathbb{R}$ is called* uniformly continuous *if for all $\epsilon > 0$, there exists some $\delta > 0$ such that for all x, $y \in W$ with $|x - y| < \delta$ we have $|f(x) - f(y)| < \epsilon$.*

Theorem 2.26. *Let $a < b$ in $\mathbb{R}$, and let $f : [a, b] \to \mathbb{R}$ be continuous. Then f is uniformly continuous.*

Proof. Assume that f is not uniformly continuous. That would mean that there exists some $\epsilon_0 > 0$ such that for all $\delta > 0$, two points p_δ, $q_\delta \in [a, b]$ must exist, with the property that $|p_\delta - q_\delta| < \delta$, and yet $|f(p_\delta) - f(q_\delta)| \geq \epsilon_0$. In particular, for each $n \in \mathbb{N}$, we can find x_n, $y_n \in [a, b]$ with

$$|x_n - y_n| < \frac{1}{n},$$

yet

$$|f(x_n) - f(y_n)| \geq \epsilon_0.$$

But, as we know (theorem 2.4), there must be a convergent subsequence $(x_{i(n)})_{n\in\mathbb{N}}$, with say

$$\lim_{n\to\infty} x_{i(n)} = t \in [a, b].$$

But then the corresponding subsequence $(y_{i(n)})_{n\in\mathbb{N}}$ must also converge, and we have

$$\lim_{n\to\infty} x_{i(n)} = \lim_{n\to\infty} y_{i(n)} = t.$$

Since f is continuous, we must also have

$$\lim_{n\to\infty} f(x_{i(n)}) = \lim_{n\to\infty} f(y_{i(n)}) = f(t).$$

But this is a contradiction, since $|f(x_{i(n)}) - f(y_{i(n)})| \geq \epsilon_0$, for all n. □

Remark. *Again, the important property of closed, bounded intervals like $[a, b]$ is that they are compact. Thus the more general formulation of theorem 2.26 would be:*

Let $K \subset \mathbb{R}$ be compact and let $f : K \to \mathbb{R}$ be continuous; then f is uniformly continuous.

2.9 Differentiability

In this section, it is convenient to consider functions $f : U \to \mathbb{R}$, where U is an *open* subset of $\mathbb{R}$. In particular, we will take $U = (a, b)$, with $a < b$, or else simply $U = \mathbb{R}$.

Definition. *The function $f : U \to \mathbb{R}$ is* differentiable *at the point $x_0 \in U$ if there exists some number $f'(x_0) \in \mathbb{R}$, such that for all $\epsilon > 0$, a $\delta > 0$ exists with*

$$\left| \frac{f(x) - f(x_0)}{x - x_0} - f'(x_0) \right| < \epsilon,$$

for all $x \in U$ with $x \neq x_0$ and $|x - x_0| < \delta$.

Another way of saying the same thing is to say that

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} = f'(x_0).$$

But when writing this, we must always be careful to say that we do not allow h to be zero (after all, you can't divide by zero!), and also we must ensure that the point $x_0 + h$ is always an element of U.

That is, the function f is differentiable at x_0 if for any convergent sequence $(u_n)_{n \in \mathbb{N}}$, with $u_n \in U$, $\lim_{n \to \infty} u_n = x_0$, but $u_n \neq x_0$ for all n, we have

$$\lim_{n \to \infty} \frac{f(u_n) - f(x_0)}{u_n - x_0} = f'(x_0).$$

Theorem 2.27. *If $f : U \to \mathbb{R}$ is differentiable at the point $x_0 \in U$, then f is also continuous at x_0.*

Proof. Obvious! □

We also have the following theorem, which you have undoubtedly seen at school.

Theorem 2.28. *Let $f, g : U \to \mathbb{R}$ be differentiable at the point $x_0 \in U$. Then*

- *$(f + g) : U \to \mathbb{R}$ is differentiable at x_0, and we have $(f + g)'(x_0) = f'(x_0) + g'(x_0)$.*
- *$(f \cdot g) : U \to \mathbb{R}$ is differentiable at x_0, and we have*

$$(f \cdot g)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0).$$

- *If $g(x_0) \neq 0$ then $(f/g)' : U \to \mathbb{R}$ is differentiable at x_0, and we have*

$$\left(\frac{f}{g} \right)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{(g(x_0))^2}.$$

- *Assuming g is differentiable at $f(x_0)$, with $g : f(U) \to \mathbb{R}$, then $(f \circ g) : U \to \mathbb{R}$ is differentiable at x_0, and we have $(f \circ g)'(x_0) = f'(x_0)g'(f(x_0))$.*

Proof. A simple exercise, using the results for convergent sequences which we have already studied. But perhaps it might be worthwhile to look at the proof for the chain rule.

Given the function $f \circ g$, that is, $(f \circ g)(x) = g(f(x))$, let us define

$$h(y) = \begin{cases} \frac{g(y)-g(f(x_0))}{y-f(x_0)}, & \text{if } y \neq f(x_0), \\ g'(f(x_0)), & \text{if } y = f(x_0). \end{cases}$$

Since g is differentiable at $f(x_0)$, we have

$$\lim_{y \to f(x_0)} h(y) = g'(f(x_0)).$$

(That is, given any sequence $(y_n)_{n\in\mathbb{N}}$ of points in $f(U)$ with $\lim_{n\to\infty} y_n = f(x_0)$, then we must have $\lim_{n\to\infty} h(y_n) = g'(f(x_0))$.)

Therefore, we have

$$g(y) - g(f(x_0)) = h(y)(y - f(x_0)),$$

for all $y \in U$, and so

$$\begin{aligned} (f \circ g)'(x_0) &= \lim_{x\to x_0} \frac{(f \circ g)(x) - (f \circ g)(x_0)}{x - x_0} \\ &= \lim_{x\to x_0} \frac{g(f(x)) - g(f(x_0))}{x - x_0} \\ &= \lim_{x\to x_0} \frac{h(f(x))(f(x) - f(x_0)}{x - x_0} \\ &= \left(\lim_{x\to x_0} h(f(x))\right) \left(\lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0}\right) \\ &= g'(f(x_0))f'(x_0). \end{aligned}$$

□

Theorem 2.29. *Let $f : (a, b) \to \mathbb{R}$ be a strictly monotonic, continuous function with $f((a, b)) = (c, d)$, say, such that the mapping $f : (a, b) \to (c, d)$ is a bijection whose inverse is the mapping $\phi : (c, d) \to (a, b)$. Assume that f is differentiable at the point $x_0 \in (a, b)$, such that $f'(x_0) \neq 0$. Then ϕ is differentiable at the point $f(x_0)$, and we have*

$$\phi'(f(x_0)) = \frac{1}{f'(x_0)}.$$

Proof. Let $(y_n)_{n\in\mathbb{N}}$ be any convergent sequence in (c, d), with $\lim_{n\to\infty} y_n = f(x_0)$, such that $y_n \neq f(x_0)$, for all n. Then, taking $z_n = \phi(y_n)$ for each $n \in \mathbb{N}$ (that

means that $f(z_n) = y_n$), we have that $\lim_{n\to\infty} z_n = x_0$, since ϕ is continuous. Therefore

$$\begin{aligned}
\phi'(f(x_0)) &= \lim_{n\to\infty} \frac{\phi(y_n) - \phi(f(x_0))}{y_n - f(x_0)} \\
&= \lim_{n\to\infty} \frac{z_n - x_0}{f(z_n) - f(x_0)} \\
&= \lim_{n\to\infty} \frac{1}{\frac{f(z_n)-f(x_0)}{z_n - x_0}} \\
&= \frac{1}{f'(x_0)}.
\end{aligned}$$

□

2.10 Taking another look at the exponential function

Theorem 2.30. *Let $(u_n)_{n\in\mathbb{N}}$ be a convergent sequence of real numbers, with $u_n \neq 0$, for all n, and furthermore, $\lim_{n\to\infty} u_n = 0$. Then we have*

$$\lim_{n\to\infty} \frac{\exp(u_n) - 1}{u_n} = 1.$$

In order to prove this theorem, we first prove the following

Lemma. *For all $x \in \mathbb{R}$ with $|x| \leq 1$ we have $|\exp(x) - (1 + x)| < |x|^2$.*

Proof. We have

$$\begin{aligned}
|\exp(x) - (1 + x)| &= \left| \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \right| \\
&\leq |x|^2 \left(\frac{1}{2!} + \frac{|x|}{3!} + \frac{|x|^2}{4!} + \cdots \right) \\
&\leq |x|^2 \left(\frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots \right) \\
&= \frac{|x|^2}{2} \left(1 + \frac{2}{3!} + \frac{2}{4!} + \cdots \right) \\
&< \frac{|x|^2}{2} \left(\sum_{n=0}^{\infty} \left(\frac{1}{2} \right)^n \right) \\
&= \frac{|x|^2}{2} \left(\frac{1}{1 - \frac{1}{2}} \right) \\
&= |x|^2.
\end{aligned}$$

□

Proof. (of theorem 2.30)

$$\left|\frac{\exp(u_n)-1}{u_n}-1\right| = \left|\frac{\exp(u_n)-(1+u_n)}{u_n}\right| < |u_n|,$$

for $|u_n| < 1$. And this converges to zero as the sequence converges to zero. □

Theorem 2.31. *The exponential function is everywhere differentiable, with* $\exp'(x) = \exp(x)$*, for all* $x \in \mathbb{R}$.

Proof. Let $(u_n)_{n\in\mathbb{N}}$ be a convergent sequence of real numbers, with $u_n \neq 0$, for all n, and furthermore, $\lim_{n\to\infty} u_n = 0$. Then we have

$$\begin{aligned}
\exp'(x) &= \lim_{n\to\infty} \frac{\exp(x+u_n)-\exp(x)}{u_n} \\
&= \exp(x) \lim_{n\to\infty} \frac{\exp(u_n)-\exp(0)}{u_n} \\
&= \exp(x) \lim_{n\to\infty} \frac{\exp(u_n)-1}{u_n} \\
&= \exp(x)\cdot 1 \\
&= \exp(x).
\end{aligned}$$

□

2.11 The logarithm function

Definition. *From the properties of the exponential function (continuous, strictly monotonic, positive, etc.), we see that the mapping* $\exp : \mathbb{R} \to (0,\infty)$ *is a bijection. The inverse mapping from* $(0,\infty)$ *back to* $\mathbb{R}$ *is called the* logarithm, *denoted by*

$$\ln : (0,\infty) \to \mathbb{R}.$$

Remark. *This is the* natural logarithm. *The logarithm to the base 10, sometimes written* $\log_{10}$*, which you might encounter in practical computer applications, plays no role in mathematics. How do we convert natural logarithms into logarithms to the base 10? The answer: by means of the formula*

$$\log_{10}(x) = \frac{\ln(x)}{\ln(10)}.$$

Since we know that $\exp(x+y) = \exp(x)\cdot\exp(y)$, for all x, $y \in \mathbb{R}$, it follows that

$$x + y = \ln(\exp(x+y)) = \ln(\exp(x)\cdot\exp(y)).$$

Now let $a = \exp(x)$, and $b = \exp(y)$. Then we have $x = \ln(a)$ and $y = \ln(b)$.

All of this gives the functional equation for the logarithm function:

$$\ln(a\cdot b) = \ln(a) + \ln(b),$$

for all a, $b > 0$.

Identifying the exponential function with powers and roots: the number e

But thinking about this leads to the more general question: given x, $y \in \mathbb{R}$, what is x^y. After all, every pocket calculator these days has a button marked "x^y".

Well, to begin with, given a, then we all know that $a^2 = a \cdot a$. More generally, given m, $n \in \mathbb{N}$, we write $a^{m+n} = a^m \cdot a^n$. This is beginning to look like the functional equation for the exponential function!

Following this additive business, if $a \geq 0$, then the square root of a is the number which, when multiplied with itself gives $a = a^1$. Therefore, it is natural to write $\sqrt{a} = a^{1/2}$. Also $\frac{1}{a^n} = a^{-n}$, for $n \in \mathbb{N}$. And in general, following this plan, we have the rule

$$a^{\frac{p}{q}} = \left(\sqrt[q]{a}\right)^p,$$

for all $a \geq 0$, $p \in \mathbb{Z}$ and $q \in \mathbb{N}$.

But looking at the functional equations for both the exponential and the logarithm functions, we see that for $a \geq 0$ we have

$$a^n = \exp(\ln(a^n)) = \exp(n \cdot \ln(a)),$$

for $n \in \mathbb{N}$. But then also

$$\frac{1}{a^n} = a^{-n} = \exp(\ln(a^{-n})) = \exp(-n \cdot \ln(a)),$$

since $a^n \cdot \frac{1}{a^n} = 1 = \exp(0)$. Similarly,

$$a^{\frac{1}{n}} = \exp(\frac{1}{n} \cdot \ln(a)).$$

Therefore, by extension we have

$$a^{\frac{p}{q}} = \exp(\frac{p}{q} \cdot \ln(a)),$$

for all rational numbers p/q. Finally, since exp and ln are continuous, we must have

$$a^b = \exp(b \cdot \ln(a)),$$

for all $b \in \mathbb{R}$.

At this stage, mathematicians become interested in the special number $\exp(1)$, which we call "e", for short. It is an important mathematical constant, similar to that other special number π. People have worked out that

$$e \approx 2.718281828459045.$$

Now, given any $n \in \mathbb{N}$, we have

$$n = \ln(\exp(n)) = \ln(\underbrace{\exp(1) \cdots \exp(1)}_{n \text{ times}}) = \ln(e^n).$$

Therefore

$$\exp(n) = \exp(\ln(e^n)) = e^n,$$

and so on. Following our reasoning from before, we conclude that

$$\exp(x) = e^x,$$

for all $x \in \mathbb{R}$. Thus, in general we have

$$a^b = e^{b \cdot \ln(a)},$$

for all $a \geq 0$ and $b \in \mathbb{R}$.

Theorem 2.32. *For all $x \in (0, \infty)$, we have*

$$\ln'(x) = \frac{1}{x}.$$

Proof. We have $\exp(\ln(x)) = x$, for all $x \in (0, \infty)$. Therefore

$$1 = \exp(\ln(x))' = \ln'(x) \cdot \exp'(\ln(x)) = \ln'(x) \cdot \exp(\ln(x)) = \ln'(x) \cdot x.$$

□

2.12 The mean value theorem

Theorem 2.33 (Rolle)**.** *Let $a < b$ in $\mathbb{R}$, and let $f : [a, b] \to \mathbb{R}$ be continuous in $[a, b]$ and differentiable everywhere in (a, b). Assume furthermore, that $f(a) = f(b)$. Then there exists some point $\xi \in (a, b)$, such that $f'(\xi) = 0$.*

Proof. If f is the constant function, $f(x) = f(a)$, for all $x \in [a, b]$, then obviously $f(\xi) = 0$, for all $\xi \in (a, b)$. On the other hand, if f is not constant, then either

1. there exists $y \in (a, b)$ with $f(y) > f(a)$, or else
2. there exists $z \in (a, b)$ with $f(z) < f(a)$.

Assume that we have case (1.). (Case (2.) is similar.) Then, according to theorem 2.24, there exists some $\xi \in (a, b)$ with $f(\xi) \geq f(x)$, for all $x \in [a, b]$. For each $n \in \mathbb{N}$, let $u_n = \xi - \frac{\xi - a}{n+1}$. Then $(u_n)_{n \in \mathbb{N}}$ is a convergent sequence in (a, b) with $\lim_{n \to \infty} u_n = \xi$. Thus we must have

$$f'(\xi) = \lim_{n \to \mathbb{N}} \frac{f(u_n) - f(\xi)}{u_n - \xi}.$$

However $f(u_n) - f(\xi) \leq 0$, since $f(\xi)$ is the largest possible value. Also $u_n - \xi < 0$ for all n. Thus we must have $f'(\xi) \leq 0$.

On the other hand, let $v_n = \xi + \frac{b-\xi}{n+1}$, for all n. Then $(v_n)_{n\in\mathbb{N}}$ is also a convergent sequence in (a,b) with $\lim_{n\to\infty} u_n = \xi$. Thus we must have

$$f'(\xi) = \lim_{n\to\mathbb{N}} \frac{f(v_n) - f(\xi)}{v_n - \xi}.$$

However $f(v_n) - f(\xi) \leq 0$, since $f(\xi)$ is the largest possible value, and also $u_n - \xi > 0$ for all n. Thus we must have $f'(\xi) \geq 0$.

Combining these two conclusions, we see that the only possibility is that $f'(\xi) = 0$. □

Theorem 2.34 (Mean value theorem (or "Mittelwertsatz")). *Let $a < b$ in $\mathbb{R}$, and let $f : [a,b] \to \mathbb{R}$ be continuous in $[a,b]$ and differentiable everywhere in (a,b). Then there exists some point $\xi \in (a,b)$ with*

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}.$$

Proof. Let the new function $F : [a,b] \to \mathbb{R}$ be defined by

$$F(x) = f(x) - \frac{f(b) - f(a)}{b - a}(x - a).$$

Obviously the function F fulfills the conditions of Rolle's theorem (2.33). So let $\xi \in (a,b)$ with $F'(\xi) = 0$. Then we have

$$F'(\xi) = 0 = f'(\xi) - \frac{f(b) - f(a)}{b - a}.$$

□

2.13 Complex numbers

The equation $x^2 + 1 = 0$ has no solution within the system of real numbers $\mathbb{R}$. To solve this "problem", mathematicians have invented an imaginary number, called i (for "i"maginary), which is supposed to solve the equation. So we could imagine that we have

$$i = \sqrt{-1}.$$

But then, since $(-1)^2 = 1$, it would seem to make sense to agree that also

$$(-i)^2 = ((-1)\cdot i)^2 = (-1)^2 \cdot i^2 = 1 \cdot -1 = -1.$$

More generally, given any $x \in \mathbb{R}$, we can imagine that ix is also a number, such that $(ix)^2 = -x^2$.

In order to combine these imaginary numbers with the "real" numbers of our normal existence, we just add the two kinds of numbers together. This results in the field of *complex numbers*, denoted by $\mathbb{C}$. That is,

$$\mathbb{C} = \{a + ib : a, b \in \mathbb{R}\}.$$

Addition in $\mathbb{C}$ is given by

$$(a+ib)+(c+id)=(a+c)+i(b+d).$$

The rule for multiplication uses the fact that we have agreed to make $i^2=-1$. Therefore,

$$(a+ib)\cdot(c+id)=(ac-bd)+i(ad+bc).$$

It is a simple exercise to verify that with these rules for addition and multiplication, $\mathbb{C}$ is a field. The zero element is $0+i0$ and the one is $1+i0$. In particular, if $a+ib$ is not zero, then the inverse under multiplication is

$$(a+ib)^{-1}=\frac{a-ib}{a^2+b^2}.$$

Using the ideas of linear algebra, we see that $\mathbb{C}$ is a 2-dimensional vector space over $\mathbb{R}$. Therefore it is natural to picture the numbers in $\mathbb{C}$ on the 2-dimensional Euclidean plane, the horizontal axis representing $\mathbb{R}$, the real numbers, and the vertical axis representing the imaginary numbers $i\mathbb{R}$. Thus we have $\mathbb{R}\subset\mathbb{C}$ when real numbers $x\in\mathbb{R}$ are identified with their real counterparts $x+i0\in\mathbb{C}$.

We have seen how important it is to think about the distance between two numbers. Therefore, in $\mathbb{C}$, we define the distance between pairs of complex numbers to be the usual Euclidean distance. That is, given $a+bi$ and $c+id$ in $\mathbb{C}$, then the distance between them is

$$|(a+ib)-(c+id)|=\sqrt{(a-c)^2+(b-d)^2}.$$

So let $z\in\mathbb{C}$ be some complex number. That is, there are two real numbers, a and b, with $z=a+ib$. We sometimes write $re(z)$ to represent the *real part* of z. That is, $re(z)=a$. Also the *imaginary part* of z is $im(z)=b$. The *complex conjugate* $\overline{z}$ to z is the complex number

$$\overline{z}=a-ib.$$

Then we have

$$z\overline{z}=(a+ib)(a-ib)=a^2+b^2=|z|^2.$$

Here, $|z|$ denotes the distance between z and the zero of $\mathbb{C}$, namely $0+i0$. It is the *absolute value* of z, and for real numbers it corresponds to the absolute value function which we have already seen in $\mathbb{R}$.

We have

- $|z|=0\Leftrightarrow z=0$,
- $|\overline{z}|=|z|$, and
- $|w\cdot z|=|w|\cdot|z|$, for all w, $z\in\mathbb{C}$.

Also, the combinations of addition and multiplication with complex conjugates are

- $\overline{w+z} = \overline{w} + \overline{z}$ and
- $\overline{w \cdot z} = \overline{w} \cdot \overline{z}$.

Therefore, if we have a polynomial with real coefficients

$$P(z) = a_0 + a_1 z + \cdots + a_n z^n,$$

where $a_j \in \mathbb{R}$, for $j = 0, \ldots, n$, then the complex conjugate is $\overline{P(z)} = P(\overline{z})$.

Given two complex numbers w, $z \in \mathbb{C}$, we have

$$|w + z| \leq |w| + |z|.$$

In order to see this, begin by observing that for all complex numbers $u \in \mathbb{C}$, we have both

$$re(u) \leq |u| \quad \text{and} \quad im(u) \leq |u|.$$

In particular, we have

$$re(w\overline{z}) \leq |w\overline{z}| = |w| \cdot |\overline{z}| = |w| \cdot |z|.$$

Therefore

$$\begin{aligned} |w+z|^2 &= (w+z)\overline{(w+z)} \\ &= (w+z)(\overline{w}+\overline{z}) \\ &= w\overline{w} + w\overline{z} + z\overline{w} + z\overline{z} \\ &= w\overline{w} + w\overline{z} + \overline{w\overline{z}} + z\overline{z} \\ &= |w|^2 + 2re(w\overline{z}) + |z|^2 \\ &\leq |w|^2 + 2|w| \cdot |z| + |z|^2 \\ &= (|w| + |z|)^2 \end{aligned}$$

And so the triangle inequality $|w + z| \leq |w| + |z|$ holds.[5]

It is now a simple exercise to verify that for arbitrary triples of complex numbers u, v, $w \in \mathbb{C}$, we have

$$|u - w| \leq |u - v| + |v - w|.$$

All of our ideas concerning convergent sequences and series of real numbers can be taken over directly into the realm of complex numbers. The proofs are the same, again using absolute values to measure distances. In particular, we see that a sequence $(z_n)_{n\in\mathbb{N}}$, with $z_n = x_n + iy_n$, for each n, converges if and only if both the sequences of real numbers $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ converge. Thus if $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty} y_n = y$, then

$$\lim_{n\to\infty} z_n = \lim_{n\to\infty} (x_n + iy_n) = \lim_{n\to\infty} x_n + i \lim_{n\to\infty} y_n = x + iy.$$

[5] Of course the fact that the absolute value of a complex number corresponds with the norm of the vector representing that number in $\mathbb{R}^2$ means that the triangle inequality in $\mathbb{C}$ is nothing more than the triangle inequality in $\mathbb{R}^2$, considered as a normed vector space.

Specifically, given an $\epsilon > 0$, there exists an $N_r \in \mathbb{N}$ and an $N_i \in \mathbb{N}$ such that for all $n_r \geq N_r$ we have $|x_n - x| < \epsilon/2$ and for all $n_i \geq N_i$ we have $|y_n - y| < \epsilon/2$. Now take $N = \max\{N_r, N_i\}$. Then for all $n \geq N$ we have

$$|z_n - z| = |(x_n - x) + i(y_n - y)| \leq |x_n - x| + |y_n - y| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

In particular, we have that:

- Every bounded sequence in $\mathbb{C}$ contains a convergent subsequence.[6]
- Every Cauchy sequence in $\mathbb{C}$ converges.
- The idea of absolutely convergent series of complex numbers is defined analogously to the real case.
- Absolutely convergent series are convergent, and furthermore, the limit of an absolutely convergent series does not depend on its ordering.
- Both the comparison test and the quotient test are valid in $\mathbb{C}$.
- A function $f : \mathbb{C} \to \mathbb{C}$ is continuous at a point $z_0 \in \mathbb{C}$ if for all $\epsilon > 0$, a $\delta > 0$ exists with $|f(z) - f(z_0)| < \epsilon$ for all $z \in \mathbb{C}$ with $|z - z_0| < \delta$. Equally, f is continuous at z_0 if for every convergent sequence $(z_n)_{n\in\mathbb{N}}$ with $z_n \longrightarrow z_0$, we have $f(z_n) \longrightarrow f(z_0)$.

If we have a convergent power series $\sum_{n=0}^{\infty} a_n z^n$, then the complex conjugate is

$$\overline{\sum_{n=0}^{\infty} a_n z^n} = \sum_{n=0}^{\infty} a_n \overline{z}^n.$$

For complex numbers $z \in \mathbb{C}$, we have that the exponential series

$$\exp(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}$$

is also absolutely convergent, and the resulting function $\exp : \mathbb{C} \to \mathbb{C}$ is continuous. So, in particular, we have

$$\overline{\exp(z)} = \exp(\overline{z}),$$

for all $z \in \mathbb{C}$.

And, of course, the functional equation for the exponential function

$$\exp(w + z) = \exp(w)\exp(z)$$

also holds in $\mathbb{C}$.

[6] The sequence $(z_n)_n \in \mathbb{N}$ is bounded if there exists some positive real number $K > 0$ such that $|z_n| \leq K$, for all $n \in \mathbb{N}$.

2.14 The trigonometric functions: sine and cosine

Definition. *For all $x \in \mathbb{R}$ the functions* sin *and* cos *are defined by*

$$\cos(x) = re(\exp(ix)) \quad \textit{and} \quad \sin(x) = im(\exp(ix)).$$

That is, the sine and cosine functions are *defined* in terms of Euler's formula

$$e^{ix} = \cos(x) + i\sin(x).$$

Since $e^{i(-x)} = e^{-ix} = \overline{e^{ix}}$, we have

$$\cos(x) = \frac{1}{2}\left(e^{ix} + e^{-ix}\right)$$

and

$$\sin(x) = \frac{1}{2i}\left(e^{ix} - e^{-ix}\right).$$

Therefore

$$\cos(-x) = \cos(x) \quad \text{and} \quad \sin(-x) = -\sin(x).$$

Now, we have

$$\begin{aligned} |\exp(ix)| &= \sqrt{\exp(ix)\cdot\overline{\exp(ix)}} \\ &= \sqrt{\exp(ix)\cdot\exp(\overline{ix})} \\ &= \sqrt{\exp(ix)\exp(-ix)} \\ &= \sqrt{\exp(ix-ix)} \\ &= \sqrt{\exp(0)} \\ &= \sqrt{1} = 1. \end{aligned}$$

Therefore, it must be that $|\sin(x)| \leq 1$ and $|\cos(x)| \leq 1$, for all $x \in \mathbb{R}$. But also,

$$\sin^2(x) + \cos^2(x) = (re(\exp(ix)))^2 + (im(\exp(ix)))^2 = |\exp(ix)| = 1.$$

Furthermore, using the functional equation of the exponential function, we have

$$\begin{aligned} \cos(x+y) + i\sin(x+y) &= \exp(i(x+y)) \\ &= \exp(ix)\cdot\exp(iy) \\ &= (\cos(x) + i\sin(x))(\cos(y) + i\sin(y)) \\ &= (\cos(x)\cos(y) - \sin(x)\sin(y)) + i(\cos(x)\sin(y) + \sin(x)\cos(y)) \end{aligned}$$

Since the real, and the imaginary parts must be equal, we have the two equations

$$\cos(x+y) = \cos(x)\cos(y) - \sin(x)\sin(y),$$

and

$$\sin(x+y) = \cos(x)\sin(y) + \sin(x)\cos(y).$$

It is now an easy exercise to obtain the standard formulas

$$\sin(x) - \sin(y) = 2\cos\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right),$$

and

$$\cos(x) - \cos(y) = -2\sin\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right).$$

In particular let $u = \frac{x+y}{2}$ and $v = \frac{x-y}{2}$, so that $x = u+v$ and $y = u-v$. Then we have

$$\begin{aligned}
\sin(x) - \sin(y) &= \sin(u+v) - \sin(u-v) \\
&= (\sin(u)\cos(v) + \cos(u)\sin(v)) \\
&\quad -(\sin(u)\cos(-v) + \cos(u)\sin(-v)) \\
&= 2\cos(u)\sin(v) \\
&= 2\cos\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right)
\end{aligned}$$

Also

$$\begin{aligned}
\cos(x) - \cos(y) &= \cos(u+v) - \cos(u-v) \\
&= (\cos(u)\cos(v) - \sin(u)\sin(v)) \\
&\quad -(\cos(u)\cos(-v) - \sin(u)\sin(-v)) \\
&= -2\sin(u)\sin(v) \\
&= -2\sin\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right)
\end{aligned}$$

The trigonometric functions can be expressed in terms of power series as follows

Theorem 2.35.

$$\sin(x) = \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$

and

$$\cos(x) = \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots$$

Proof. This follows by looking at the exponential series, observing that $i^2 = -1$.

Namely,

$$\begin{aligned}
\exp(ix) &= \sum_{n=0}^{\infty} \frac{(ix)^n}{n!} \\
&= \sum_{n=0}^{\infty} i^n \frac{x^n}{n!} \\
&= \left(\sum_{k=0}^{\infty} i^{2k} \frac{x^{2k}}{(2k)!}\right) + \left(\sum_{k=0}^{\infty} i^{2k+1} \frac{x^{2k+1}}{(2k+1)!}\right) \\
&= \left(\sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}\right) + i\left(\sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}\right) \\
&= \cos(x) + i\sin(x).
\end{aligned}$$

□

The derivatives of the trigonometric functions are also found using the exponential function. In fact within the theory of complex analysis, we define the derivative in exactly the same way it is defined in real analysis. Namely the function $f : \mathbb{C} \to \mathbb{C}$ is differentiable at the point $z_0 \in \mathbb{C}$ if there exists a complex number $f'(z_0)$, such that

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} = f'(z_0),$$

with $z \neq z_0$. That is to say, for all convergent sequences $(z_n)_{n \in \mathbb{N}}$, with $\lim_{n\to\infty} z_n = z_0$ and $z_n \neq z_0$ for all n, we have

$$\lim_{n \to \infty} \frac{f(z_n) - f(z_0)}{z_n - z_0} = f'(z_0).$$

The proof that the exponential function is differentiable for all complex numbers $z_0 \in \mathbb{C}$ is the same as in the real case. Again, we obtain the result that

$$\exp'(z) = \exp(z),$$

for all $z \in \mathbb{C}$.

Theorem 2.36. $\sin'(x) = \cos(x)$ *and* $\cos'(x) = -\sin(x)$, *for all* $x \in \mathbb{R}$.

Proof. Using the chain rule, we have $\exp'(ix) = i\exp(x)$. That is,

$$\cos'(x) + i\sin'(x) = \exp'(ix) = i\exp(ix) = i(\cos(x) + i\sin(x)) = -\sin(x) + i\cos(x).$$

□

But we can also find a more direct proof, where the functions sin and cos are considered as being simply real functions $\mathbb{R} \to \mathbb{R}$.

Lemma.

$$\lim_{\substack{x\to 0\\ x\neq 0}} \frac{\sin(x)}{x} = 1$$

Proof.

$$\begin{aligned} |\sin(x) - x| &= \left| -\frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \right| \\ &= \left| \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} - \cdots \right| \\ &< \frac{|x|^3}{3!} \quad \text{when} \quad |x| < 1. \end{aligned}$$

Therefore

$$\left| \frac{\sin(x) - x}{x} \right| = \left| \frac{\sin(x)}{x} - 1 \right| < \frac{|x|^2}{3!}.$$

□

Then, for $h \neq 0$, we have

$$\begin{aligned} \frac{\sin(x+h) - \sin(x)}{h} &= \frac{2\cos\left(\frac{2x+h}{2}\right)\sin\left(\frac{h}{2}\right)}{h} \\ &= \cos\left(\frac{2x+h}{2}\right) \cdot \frac{\sin\left(\frac{h}{2}\right)}{\frac{h}{2}}. \end{aligned}$$

Therefore in the limit $h \to 0$ we obtain

$$\sin'(x) = \lim_{\substack{h\to 0\\ h\neq 0}} \left(\cos\left(\frac{2x+h}{2}\right) \cdot \frac{\sin\left(\frac{h}{2}\right)}{\frac{h}{2}} \right) = \cos(x).$$

Similarly, we have

$$\begin{aligned} \frac{\cos(x+h) - \cos(x)}{h} &= \frac{-2\sin\left(\frac{2x+h}{2}\right)\sin\left(\frac{h}{2}\right)}{h} \\ &= -\sin\left(\frac{2x+h}{2}\right) \cdot \frac{\sin\left(\frac{h}{2}\right)}{\frac{h}{2}}, \end{aligned}$$

leading to $\cos'(x) = -\sin(x)$.

2.15 The number π

Theorem 2.37. *The function* cos *has exactly one single zero in the open interval* $(0, 2)$*. That is, there exists a unique* $x_0 \in (0, 2)$ *with* $\cos(x_0) = 0$*.*

The proof of this theorem starts by looking at the power series expression for the cosine function, namely $\cos(x) = \sum_{n=0}^{\infty}(-1)^n \frac{x^{2n}}{(2n)!}$. Obviously we have $\cos(0) = 1$. But then

$$\begin{aligned}\cos(2) &= 1 - \frac{2^2}{2!} + \frac{2^4}{4!} - \frac{2^6}{6!} + \cdots \\ &= 1 - \frac{4}{2} + \frac{16}{24} - \frac{64}{720} + \cdots \\ &= 1 - 2 + \frac{2}{3} - \frac{4}{45} + \cdots\end{aligned}$$

Thinking about Leibniz convergence test for series, we see that it must be that $\cos(2) < 0$. Then theorem 2.25 shows that there must be a zero somewhere between 0 and 2. On the other hand, the power series expression for the sine function shows that $\sin(x) > 0$, for all $x \in (0, 2)$. Then given $0 < x < y < 2$, we must have

$$\cos(y) - \cos(x) = -2\sin\left(\frac{y+x}{2}\right)\sin\left(\frac{y-x}{2}\right) < 0.$$

Therefore, the cosine function must be strictly monotonically decreasing between 0 and 2.

Definition. *The number π is defined to be $\pi = 2x_0$, where x_0 is the unique zero of* cos *in the open interval* $(0, 2)$.

Theorem 2.38. *We have*

- $\cos(\pi) = -1$, $\cos(3\pi/2) = 0$ *and* $\cos(2\pi) = 1$,
- $\sin(\pi/2) = 1$, $\sin(\pi) = 0$, $\sin(3\pi/2) = -1$ *and* $\sin(2\pi) = 0$,
- $\cos(x + 2\pi) = \cos(x)$, $\sin(x + 2\pi) = \sin(x)$,
- $\cos(x + \pi) = -\cos(x)$, $\sin(x + \pi) = -\sin(x)$,
- $\cos(\pi/2 - x) = \sin(x)$, *and* $\sin(\pi/2 - x) = \cos(x)$

for all $x \in \mathbb{R}$.

The proof involves lots of little exercises which you can look up in the standard textbooks on analysis. For example, since we know that $\cos(\pi/2) = 0$, $\sin(\pi/2) > 0$, and $\cos^2(\pi/2) + \sin^2(\pi/2) = 1$, it must follow that $\sin(\pi/2) = 1$. But then

$$\cos(\pi) = \cos\left(\frac{\pi}{2} + \frac{\pi}{2}\right) = \cos^2\left(\frac{\pi}{2}\right) - \sin^2\left(\frac{\pi}{2}\right) = 0 - 1 = -1.$$

The other points in this theorem can be similarly proved.

Using these ideas, people have found various formulas for the number π. One particularly interesting formula (which is related to the famous *Riemann zeta function* in number theory) is the following

$$\frac{\pi^2}{6} = \sum_{n=1}^{\infty} \frac{1}{n^2}.$$

2.16 The geometry of the complex numbers

Given a complex number $z = x + iy \in \mathbb{C}$, with x, $y \in \mathbb{R}$, we can say that z is the point $(x, y) \in \mathbb{R}^2$, where $\mathbb{R}^2$ is the Euclidean plane. Then, given another complex number $w = u + iv$, we have that the sum $z + w$ is the point $(x + u, y + v) \in \mathbb{R}^2$. This is just the normal vector addition operation of linear algebra.

But things become more interesting when we multiply two complex numbers together. For this, another representation, using *polar coordinates*, is more appropriate. Taking $z = x + iy$, and using the trigonometric functions, we see that there is a unique $r \in \mathbb{R}$ with $r \geq 0$, and (if $r > 0$) a unique $\theta \in [0, 2\pi)$, such that $x = r\cos(\theta)$ and $y = r\sin(\theta)$. That is

$$z = r\cos(\theta) + ir\sin(\theta).$$

Similarly, there exist $s \geq 0$ and $\phi \in [0, 2\pi)$, such that

$$w = s\cos(\phi) + is\sin(\phi).$$

Then we have

$$\begin{aligned} z \cdot z &= (r\cos(\theta) + ir\sin(\theta)) \cdot (s\cos(\phi) + is\sin(\phi)) \\ &= rs((\cos(\theta)\cos(\phi) - \sin(\theta)\sin(\phi)) + i(\cos(\theta)\sin(\phi) + \sin(\theta)\cos(\phi))) \\ &= rs(\cos(\theta + \phi) + i\sin(\theta + \phi)) \end{aligned}$$

Another way to say the same thing is to write $z = re^{i\theta}$ and $w = se^{i\phi}$. Then

$$zw = re^{i\theta} \cdot se^{i\phi} = rs \cdot e^{i(\theta+\phi)}.$$

When writing $z = re^{i\theta}$, we can think of the complex number z as being the two-dimensional vector with length r, and with angle θ to the x-axis. Then we see that multiplying two complex numbers z and w gives as the result the vector with length the product of the lengths of z and w, and the angle to the x-axis is the *sum* of the angles of z and w.

In particular, multiplying z by $e^{i\theta}$ simply results in the vector z having its length remain unchanged (since $|e^{i\theta}| = 1$), but its angle is increased by θ. Also, one sees that if we take increasing values of $x \in \mathbb{R}$, then the complex number e^{ix} just winds around the unit circle of the complex plane, in direct proportion to x.

2.17 The Riemann integral

Definition. *Let $a < b$ in $\mathbb{R}$. A* partition *of the interval $[a, b]$ is a finite sequence of numbers $t_0, \ldots, t_n$, such that $t_0 = a$, $t_n = b$, and $t_{k-1} < t_k$ for $k = 1, \ldots, n$. Therefore, we can imagine that the partition splits the interval into n* subintervals

$$[a, b] = [t_0, t_1] \cup [t_1, t_2] \cup \cdots \cup [t_{n-1}, t_n].$$

The fineness *of the partition is the length of the longest subinterval, namely*

$$\max_{k=1,\dots,n} t_k - t_{k-1}.$$

Definition. *Let $f : [a,b] \to \mathbb{R}$ be a function, and let $P = \{[t_0,t_1],\dots,[t_{n-1},t_n]\}$ be a partition of $[a,b]$. A Riemann sum for f with respect to P is a sum of the form*

$$S = \sum_{k=1}^{n} f(x_k)(t_k - t_{k-1}),$$

where $t_{k-1} \leq x_k \leq t_k$, for each k.

Definition. *Let $f : [a,b] \to \mathbb{R}$ be a function. We say that f is Riemann integrable if there exists a real number, denoted by $\int_a^b f(x)dx$, such that for all $\epsilon > 0$, a $\delta > 0$ exists, such that for all Riemann sums S over partitions with fineness less than δ, we have*

$$\left| S - \int_a^b f(x)dx \right| < \epsilon.$$

2.17.1 Step functions

The usual way to think about integrals is to consider *step functions*. Again, take the interval $[a,b]$, and a partition $a = t_0 < t_1 < \cdots < t_{n-1} < t_n = b$. Next, choose n real numbers, $c_1, \cdots, c_n$. Then the step function corresponding to these choices would be the function $f : [a,b] \to \mathbb{R}$ given by

$$f(x) = c_k \Leftrightarrow x \in (t_{k-1}, t_k).$$

The values of $f(t_k)$ can be arbitrarily chosen. Obviously every step function is Riemann integrable (in fact, this follows from our theorem 2.39), and the integral is simply

$$\sum_{k=1}^{n} c_k(t_k - t_{k-1}).$$

Furthermore, just as obviously, most step functions are not continuous — they make a "jump" between adjacent intervals of the partition. So let us denote by $S([a,b],\mathbb{R})$ the set of all step functions from $[a,b]$ to $\mathbb{R}$.

Now, given two step functions $g, h \in S([a,b],\mathbb{R})$ with $g \leq h$, that is $g(x) \leq h(x)$, for all $x \in [a,b]$ then we must have

$$\int_a^b g(x)dx \leq \int_a^b h(x)dx.$$

2.17.2 Integrals defined using step functions

So this leads to another way of thinking about integrals. For let $f : [a,b] \to \mathbb{R}$ be a function such that there exist two step functions $g, h \in S([a,b], \mathbb{R})$ with $g \leq f \leq h$. Then, *assuming* that f is, indeed, Riemann integrable, it would follow that we must have

$$\int_a^b g(x)dx \leq \int_a^b f(x)dx \leq \int_a^b h(x)dx.$$

Definition. *Let $f : [a,b] \to \mathbb{R}$ be a function such that there exist two step functions $g, h \in S([a,b], \mathbb{R})$ with $g \leq f \leq h$. The* upper integral *of f, denoted by $\int^* f$, is given by*

$$\int^* f = \inf \left\{ \int_a^b h(x)dx : f \leq h, \textit{where } h \in S([a,b], \mathbb{R}) \right\}.$$

Similarly, the lower integral $\int_* f$ *is*

$$\int_* f = \sup \left\{ \int_a^b g(x)dx : g \leq f, \textit{where } g \in S([a,b], \mathbb{R}) \right\}.$$

Theorem 2.39. *The bounded function $f : [a,b] \to \mathbb{R}$ is Riemann integrable if and only if $\int_* f = \int^* f$. In this case, we have $\int_a^b f(x)dx = \int_* f$.*

Proof.

- "$\Rightarrow$": Let $\epsilon > 0$ be given. The problem then is to show that $\int^* f - \int_* f < \epsilon$. Since f is Riemann integrable, there must exist some $\delta > 0$ which is sufficiently small that

  $$\left| \sum_{k=1}^n f(\xi_k)(t_k - t_{k-1}) - \int_a^b f(x)dx \right| < \frac{\epsilon}{2},$$

  for every partition whose fineness is less than δ. Given such a partition, for each k, let

  $$\begin{aligned} u_k &= \inf\{f(x) : x \in [t_{k-1}, t_k]\} \\ v_k &= \sup\{f(x) : x \in [t_{k-1}, t_k]\} \end{aligned}$$

  Then we have

  $$\begin{aligned} S_u &= \sum_{k=1}^n u_k(t_k - t_{k-1}) \leq \int_a^b f(x)dx, \quad \text{and} \\ S_v &= \sum_{k=1}^n v_k(t_k - t_{k-1}) \geq \int_a^b f(x)dx. \end{aligned}$$

  However,

  $$S_v \geq \int^* f \geq \int_a^b f(x)dx \geq \int_* f \geq S_u,$$

and

$$S_v - S_u \leq \left(S_v - \int_a^b f(x)dx\right) + \left(\int_a^b f(x)dx - S_u\right) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

- "$\Leftarrow$": Again, let $\epsilon > 0$ be given. Since $\int^* f = \int_* f$, there must exist two step functions g, $h \in S([a,b], \mathbb{R})$ with $g \leq f \leq h$ and

$$\int_a^b h(x)dx - \int_a^b g(x)dx < \frac{\epsilon}{2}.$$

By possibly subdividing the partitions defining g and h we may assume that both are defined along a *single* partition of $[a,b]$, namely

$$a = x_0 < x_1 < \cdots < x_m = b.$$

Since f lies between the two step functions g and h, which are both bounded, it follows that f is also bounded. So let

$$M = \sup\{|f(x)| : x \in [a,b]\}.$$

Then choose

$$\delta = \frac{\epsilon}{8Mm}.$$

The problem now is to show that the Riemann sum with respect to any partition of $[a,b]$ of fineness less than δ is within ϵ of $\int^* f = \int_* * f$. So let

$$a = t_0 < t_1 < \cdots t_n = b$$

be a partition whose fineness is less than δ, and let $\xi_k \in [t_{k-1}, t_k]$, for each k. We define the new function $F : [a,b] \to \mathbb{R}$ by the rule

$$F(x) = \begin{cases} 0, & \text{if } x \in \{t_0, \ldots, t_n\}, \\ f(\xi_k), & \text{if } x \in (t_{k-1}, t_k). \end{cases}$$

Then F is Riemann integrable, and we have

$$\int_a^b F(x)dx = \sum_{k=1}^n f(\xi_k)(t_k - t_{k-1}).$$

A further function $s : [a,b] \to \mathbb{R}$ is now defined as follows.

$$s(x) = \begin{cases} 0, & \text{if } x \in [t_{k-1}, t_k], \text{ where } [t_{k-1}, t_k] \cap \{x_0, \ldots, x_m\} = \emptyset, \\ 2M, & \text{otherwise.} \end{cases}$$

Then we have $g - s \leq F \leq h + s$, and furthermore, both $g - s$ and $h + s$ are step functions. But we can only have $s(\xi_k) \neq 0$ for at most $2m$ of the numbers ξ_k. Therefore

$$\int_a^b s(x)dx = \sum_{k=1}^n s(\xi_k)(t_k - t_{k-1}) \leq 2m \cdot 2M \cdot \frac{\epsilon}{8Mm} = \frac{\epsilon}{2}.$$

This means that

$$\begin{aligned}\int_a^b g(x)dx - \frac{\epsilon}{2} &< \int_a^b (g(x) - s(x))dx \\ &\leq \int_a^b F(x)dx \\ &\leq \int_a^b (h(x) + s(x))dx \\ &< \int_a^b h(x)dx + \frac{\epsilon}{2}.\end{aligned}$$

But we also have

$$\int_a^b h(x)dx - \frac{\epsilon}{2} < \int^* f = \int_* f < \int_a^b g(x)dx + \frac{\epsilon}{2}.$$

It is now a simple exercise to show that we have

$$\left|\int_a^b f(x)dx - \int_a^b F(x)dx\right| = \left|\int_a^b f(x)dx - \sum_{k=1}^n f(\xi_k)(t_k - t_{k-1})\right| < \epsilon,$$

where the number $\int_a^b f(x)dx$ is taken to be equal to the upper and lower integrals

$$\int^* f = \int_* f.$$

□

2.17.3 Simple consequences of the definition

By thinking about integrals defined in terms of step functions, we immediately see that the following theorem is true.

Theorem 2.40. *Let f, $g : [a, b] \to \mathbb{R}$ be integrable functions, and let $\lambda \in \mathbb{R}$ be some constant. Then we have:*

1. *The function $f + g$ is also integrable, and*

$$\int_a^b (f + g)(x)dx = \int_a^b f(x)dx + \int_a^b g(x)dx.$$

2. *λf is integrable, with*

$$\int_a^b \lambda f(x)dx = \lambda \int_a^b f(x)dx.$$

3. *If $f \geq g$ then*

$$\int_a^b f(x)dx \geq \int_a^b g(x)dx.$$

4. *The functions* $\max\{f,g\}$ *and* $\min\{f,g\}$*, given by* $\max\{f,g\}(x) = \max\{f(x), g(x)\}$ *and* $\min\{f,g\}(x) = \min\{f(x), g(x)\}$ *are both integrable.*

5. *If* f_+ *is given by* $f_+(x) = \max\{0, f(x)\}$ *then* f_+ *is integrable. The function* f_- *can be similarly defined to be* $f_-(x) = \min\{0, f(x)\}$ *, and we have*

$$\int_a^b f_+(x)dx + \int_a^b f_-(x)dx = \int_a^b f(x)dx.$$

Also $|f|$*, given by* $|f|(x) = |f(x)|$ *for all* x *is integrable, and we have*

$$\left|\int_a^b f(x)dx\right| \leq \int_a^b |f(x)|dx.$$

6. *The function* fg *is integrable.*[7]

2.17.4 Integrals of continuous functions

Theorem 2.41. *Let* $[a,b] \subset \mathbb{R}$ *be a closed interval, and let* $f : [a,b] \to \mathbb{R}$ *be some continuous function. Then the integral*

$$\int_a^b f(x)dx$$

exists.[8]

Proof. Since the interval is closed, the function is uniformly continuous (theorem 2.26). The problem is to show that $\int^* f = \int_* f$, or in other words, to show that for all $\epsilon > 0$, we have $\int^* f - \int_* f \leq \epsilon$.

So let some $\epsilon > 0$ be given. Since f is uniformly continuous, there exists some $\delta > 0$ such that we have

$$|f(u) - f(v)| < \frac{\epsilon}{2(b-a)},$$

for all u, $v \in [a,b]$ with $|u-v| < \delta$. Next choose $n \in \mathbb{N}$ to be sufficiently large that $n\delta > b - a$ and we define two step functions g and h from $[a,b]$ to $\mathbb{R}$ as follows.

$$g(x) = f\left(a + \frac{m(b-a)}{n}\right) + \frac{\epsilon}{2(b-a)}$$

and

$$h(x) = f\left(a + \frac{m(b-a)}{n}\right) - \frac{\epsilon}{2(b-a)},$$

[7] But, of course, we do not always have $\int fg = \int f \cdot \int g$. For example, $\int_{-1}^{+1} x dx = 0$, yet $\int_{-1}^{+1} x^2 dx = \frac{2}{3}$.

[8] If f is only defined on an open interval (a,b) then the integral may not exist, even if f is continuous. For example, $\lim_{\epsilon \to 0} \int_\epsilon^1 \frac{1}{x} dx = \infty$.

when

$$x \in \left[a + \frac{m(b-a)}{n},\, a + \frac{(m+1)(b-a)}{n} \right),$$

for each $m \in \{0, \ldots, n-1\}$, and finally $g(b) = h(b) = f(b)$.

Then we have $g(x) \geq f(x) \geq h(x)$ for all $x \in [a,b]$, and furthermore

$$g(x) - h(x) \leq \frac{\epsilon}{b-a}.$$

Therefore we must have

$$\int^* f - \int_* f \leq \int_a^b (g(x) - h(x))dx \leq \frac{\epsilon}{b-a} \cdot (b-a) = \epsilon.$$

□

We also have the following simple analogue of the intermediate value theorem for continuous functions.

Theorem 2.42 (Mean value theorem for integrals). *Let f, $g : [a,b] \to \mathbb{R}$ be continuous functions with $g(x) \geq 0$, for all $x \in [a,b]$. Then there exists some $\xi \in [a,b]$ with*

$$\int_a^b f(x)g(x)dx = f(\xi) \int_a^b g(x)dx.$$

Proof. Let $m = \inf\{f(x) : x \in [a,b]\}$ and $M = \sup\{f(x) : x \in [a,b]\}$. Then $mg(x) \leq f(x)g(x) \leq Mg(x)$, for all $x \in [a,b]$. Therefore

$$m \int_a^b g(x)dx \leq \int_a^b f(x)g(x)dx \leq M \int_a^b g(x)dx,$$

and if we write

$$\int_a^b f(x)g(x)dx = \mu \int_a^b g(x)dx,$$

for some $\mu \in \mathbb{R}$, we must have $m \leq \mu \leq M$. But then, according to the intermediate value theorem (theorem 2.25), there must exist some $\xi \in [a,b]$, with $f(\xi) = \mu$. □

2.18 The fundamental theorem of calculus

Theorem 2.43. *Let $[a,b] \subset \mathbb{R}$ be a closed interval, and let $f : [a,b] \to \mathbb{R}$ be some continuous function. Then the function $F : [a,b] \to \mathbb{R}$, given by*

$$F(x) = \int_a^x f(t)dt$$

is differentiable in (a,b), and we have $F'(x) = f(x)$, for all $x \in (a,b)$.

Proof. Theorem 2.41 shows that the function F does exist. So let $x \in (a,b)$ be given, and we first examine

$$\lim_{h\to 0} \frac{F(x+h)-F(x)}{h} = \lim_{h\to 0} \frac{1}{h}\left(\int_a^{a+x} f(t)dt - \int_a^{a+x+h} f(t)dt\right) = \lim_{h\to 0} \frac{1}{h}\int_x^{x+h} f(t)dt,$$

where $h > 0$. According to theorem 2.42 (and taking the function g to be $g(x) = 1$, for all x), there exists some $\xi_h \in [x, x+h]$ with

$$\int_x^{x+h} f(t)dt = hf(\xi_h).$$

Since f is continuous at x, we have $\lim_{h\to 0} f(\xi_h) = f(x)$, therefore

$$\lim_{h\to 0} \frac{F(x+h)-F(x)}{h} = \lim_{h\to 0} \frac{1}{h}\int_x^{x+h} f(t)dt = \lim_{h\to 0} \frac{1}{h} hf(\xi_h) = f(x).$$

If $h < 0$ the argument is analogous. First of all, for $a < b$ we *define* the integral $\int_b^a f(x)dx$ to be

$$\int_b^a f(x)dx = -\int_a^b f(x)dx.$$

Then one need only observe that

$$F(x+h) - F(x) = -\int_{x+h}^{x} f(x)dx = +\int_x^{x+h} f(x)dx.$$

□

2.18.1 Anti-derivatives, or "Stammfunktionen"

Definition. *Let $f : (a,b) \to \mathbb{R}$ be a continuous function. A differentiable function $G : (a,b) \to \mathbb{R}$, such that $G'(x) = f(x)$, for all $x \in (a,b)$ is called an anti-derivative (Stammfunktion, in German) to f.*

Theorem 2.44. *Given a continuous function f, then any two anti-derivatives to f differ by at most a constant.*

Proof. Let G_1 and G_2 be anti-derivatives to f. Then we have $G_1' = f = G_2'$, which is to say, $G_1' - G_2' = (G_1 - G_2)' = 0$. But then the mean value theorem (theorem 2.34) shows that we must have $G_1 - G_2$ being constant, say $G_1(x) - G_2(x) = C$, for some constant $C \in \mathbb{R}$. □

But we have seen that the integral $\int_a^x f(t)dt$ is an anti-derivative. Therefore, all possible anti-derivatives are of the form

$$\int_a^x f(t)dt + C,$$

for various constants $C \in \mathbb{R}$.

In fact, we can be more specific.

Theorem 2.45. *Let $f : [a, b] \to \mathbb{R}$ be continuous, and let G be some anti-derivative to f. Then we have*

$$\int_a^b f(x)dx = G(b) - G(a).$$

Proof. In order to see this, we need only look at our original anti-derivative $F(x) = \int_a^x f(t)dt$. Therefore, we have $F(a) = 0$ and $F(b) = \int_a^b f(t)dt$. But if $F(x) - G(x) = C$, for all x, then we must have in particular

$$F(a) - G(a) = C = F(b) - G(b),$$

or

$$G(b) - G(a) = F(b) - F(a) = \int_a^b f(x)dx.$$

□

Note that people often use the notation

$$\int_a^b f(x)dx = G(x)\Big|_a^b$$

2.18.2 Another look at the fundamental theorem

Given that

$$f(x) = F'(x) = \frac{d}{dx}\left(\int_a^x f(t)dt\right),$$

then one can think of the differential operator $\frac{d}{dx}$, and the integral operator $\int$, as being inverses of one another, in some sense. We have seen that the combination $\frac{d}{dx}\int$, when applied to a continuous function f, simply gives us f back again. How about the reversed combination $\int \frac{d}{dx}$?

For this, we need to have a differentiable function f, defined on an open interval containing the interval $[a, b]$. Then, the assertion is:

Theorem 2.46. *Let $f : (c, d) \to \mathbb{R}$ be a differentiable function, and let $[a, b] \subset (c, d)$. Then we have*

$$\int_a^b f'(x)dx = f(b) - f(a).$$

Proof. This is obvious! We need only observe that f is an anti-derivative to f'. □

2.18.3 Partial integration

This is a trivial consequence of what we have done up till now. Let (c, d) be an open interval with $[a, b] \subset (c, d)$, and let f, $g : (c, d) \to \mathbb{R}$ be two differentiable functions.

Then, according to the product rule, we have $(fg)'(x) = f'(x)g(x) + f(x)g'(x)$, for all $x \in (c, d)$. Therefore it follows that

$$\int_a^b (fg)'(x) = f(x)g(x)\Big|_a^b = \int_a^b f'(x)g(x)dx + \int_a^b f(x)g'(x)dx.$$

Often, one writes this equation as

$$\int_a^b f'(x)g(x)dx = f(x)g(x)\Big|_a^b - \int_a^b f(x)g'(x)dx.$$

2.18.4 The substitution rule

Another trivial consequence. Let $f : [a, b] \to \mathbb{R}$ be continuous and $g : [c, d] \to \mathbb{R}$ be differentiable, with $g([c, d]) \subset [a, b]$. (In order to have differentiability at the endpoints, we assume that the functions are defined in open intervals containing the given closed intervals $[a, b]$ and $[c, d]$.) Since f is continuous, it is integrable; thus there exists some anti-derivative F, with $F' = f$. Then according to the chain rule of differentiation, we have

$$(F \circ g)'(x) = g'(x)F'(g(x)) = g'(x)f(g(x)).$$

Integrating both sides of the equation gives the substitution rule:

$$\int_c^d f(g(x))g'(x)dx = (F \circ g)(x)\Big|_c^d = F(g(d)) - F(g(c)) = \int_{g(c)}^{g(d)} f(x)dx.$$

In particular, let the real-valued function f be defined and Riemann integrable on an appropriate interval of $\mathbb{R}$. Then we have the following simple consequences of the substitution rule.

$$\begin{aligned}
\int_a^b f(t + c)dt &= \int_{a+c}^{b+c} f(x)dx \\
\int_a^b f(ct)dt &= \frac{1}{c}\int_{ac}^{bc} f(x)dx, \quad c \neq 0 \\
\int_a^b tf(t^2)dt &= \frac{1}{2}\int_{a^2}^{b^2} f(x)dx
\end{aligned}$$

2.19 Various examples

2.19.1 x^m for $m \in \mathbb{Z}$

Since $(x^n)' = nx^{n-1}$, for $n \in \mathbb{N}$, it follows that $\frac{1}{n+1}x^{n+1}$ is the anti-derivative to x^n. Therefore

$$\int_a^b x^n = \frac{1}{n+1}x^{n+1}\Big|_a^b$$

We have seen that $\ln'(x) = \frac{1}{x}$ when $x > 0$. But also, using the chain rule, if $x < 0$ we have that

$$\ln'(-x) = -\frac{1}{-x} = \frac{1}{x}$$

Thus $\ln'(|x|) = \frac{1}{x}$ for all x not equal to zero. It follows that for $a < b$ with $0 \notin [a, b]$ we have

$$\int_a^b \frac{dx}{x} = \ln(|x|)\Big|_a^b$$

Then, since $x^{-n}x^n = 1$, we differentiate both sides of the equation, using the product rule, to obtain that $(x^{-n})' = -nx^{-n-1}$, for all $n \in \mathbb{N}$. Therefore, $\frac{1}{-n+1}x^{-n+1}$ is the anti-derivative to x^{-n}, for $n \geq 2$. Thus if $0 \notin [a, b]$ and $m \in \mathbb{Z}$ with $m \leq -2$, we have

$$\int_a^b x^m = \frac{1}{m+1}x^{m+1}\Big|_a^b$$

So much for the integrals of monomials — and thus polynomials.

2.19.2 The exponential, trigonometric, and logarithm functions

We have $\exp'(x) = \exp(x)$, $\sin'(x) = \cos(x)$ and $\cos'(x) = -\sin(x)$. Therefore

$$\begin{aligned}
\int_a^b \exp(x)dx &= \exp(x)\Big|_a^b, \\
\int_a^b \cos(x)dx &= \sin(x)\Big|_a^b, \quad \text{and} \\
\int_a^b \sin(x)dx &= -\cos(x)\Big|_a^b
\end{aligned}$$

Also for $x > 0$ we have $(x(\ln(x) - 1))' = \ln(x)$. Therefore for $0 < a < b$

$$\int_a^b \ln(x)dx = x(\ln(x) - 1)\Big|_a^b$$

There are books containing hundreds of mathematical formulas — and in particular the integrals of all sorts of functions which might come up in practice.[9] But the more modern seeker after formulas involving integrals will undoubtedly make use of one of the various computer algebra programs which are widely available these days.[10]

Despite this, let us look at a couple further integrals. For example, the tangent function is defined to be

$$\tan(x) = \frac{\sin(x)}{\cos(x)}.$$

Since $\cos(x) = 0$ when x is $\pi/2$ or $3\pi/2$, we must exclude these numbers (and also $\pi/2 + 2m\pi$ and $3\pi/2 + 2m\pi$, for all $m \in \mathbb{Z}$) when looking at the tangent function.

In any case, one immediately sees that for an interval $[a, b]$ which avoids such "bad" numbers, we have

$$\int_a^b \tan(x)dx = -\ln(|\cos(x)|)\Big|_a^b$$

2.19.3 The hyperbolic functions

The hyperbolic sine and hyperbolic cosine functions are defined to be

$$\sinh(x) = \frac{e^x - e^{-x}}{2} \quad \text{and} \quad \cosh(x) = \frac{e^x + e^{-x}}{2}.$$

Looking at the defining power series, one can say that

$$\sinh(x) = -i\sin(ix) \quad \text{and} \quad \cosh(x) = \cos(ix).$$

In any case, using the fact that the exponential is its own anti-derivative, we immediately obtain

$$\int_a^b \sinh(x)dx = \cosh(x)\Big|_a^b \quad \text{and} \quad \int_a^b \cosh(x)dx = \sinh(x)\Big|_a^b.$$

2.19.4 The inverse trigonometric functions

We have that the sine function has the value -1 at $-\pi/2$ and $+1$ at $+\pi/2$. Between those two numbers, the sine function is strictly monotonically increasing. The inverse function is called "arcsine". Therefore for $-\pi/2 < \theta < \pi/2$ and $x = \sin(\theta)$, we have $\arcsin(x) = \theta$. In particular,

$$\begin{aligned} \arcsin(\sin(\theta)) &= \theta, \text{ and} \\ \sin(\arcsin(x)) &= x. \end{aligned}$$

[9] For example, "Formeln der Mathematik", by Dr.-Ing. Dipl.-Math. G. Arnold, herausgegeben von Prof. Dr.-Ing. H. Netz. (I have this one at home!)

[10] The open-source program "Maxima" is part of the Ubuntu distribution, and is thus freely available.

Remembering theorem 2.29, we have

$$\arcsin'(x) = \arcsin'(\sin(\theta)) = \frac{1}{\sin'(\theta)} = \frac{1}{\cos(\theta)}.$$

But

$$\cos^2(\theta) + \sin^2(\theta) = \cos^2(\theta) + x^2 = 1.$$

That is,

$$\arcsin'(x) = \frac{1}{\sqrt{1-x^2}}.$$

Therefore, if $-1 < a < b < 1$ we have

$$\int_a^b \frac{1}{\sqrt{1-x^2}} dx = \arcsin(x)\Big|_a^b.$$

Analogously, one finds that

$$\arccos'(x) = -\frac{1}{\sqrt{1-x^2}}$$

and

$$\arctan'(x) = -\frac{1}{x^2+1}.$$

In particular, in the case of arctan, we have the mapping

$$\arctan : \mathbb{R} \to (-\pi/2, \pi/2),$$

and so for any $a < b$ we have

$$\int_a^b \frac{1}{x^2+1} dx = \arctan(x)\Big|_a^b.$$

Of course the list could be extended almost indefinitely. For example our "standard textbook", namely *Analysis 1*, by Otto Forster, contains many interesting examples of what can be done with such functions. For example he proves the Wallis' product formula:

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdots = \prod_{n=1}^{\infty} \frac{4n^2}{4n^2-1},$$

making use of the integrals of various standard functions.

2.19.5 The area of a unit circle is π

For $-1 < x < 1$, let $\theta = \arcsin(x)$. Then for $-1 < a < b < 1$, we have

$$\int_a^b \sqrt{1-x^2}dx = \int_{\arcsin(a)}^{\arcsin(b)} \sqrt{1-\sin^2(\theta)}\sin'(\theta)d\theta = \int_{\arcsin(a)}^{\arcsin(b)} \cos^2(\theta)d\theta.$$

But then

$$\cos^2(x) = \left(\frac{e^{ix}+e^{-ix}}{2}\right)^2 = \frac{1}{4}\left(e^{2ix}+e^{-2ix}+2\right) = \frac{1}{2}(\cos(2x)+1),$$

so that

$$\int_a^b \sqrt{1-x^2}dx = \left(\frac{\sin(2\theta)}{4}+\frac{\theta}{2}\right)\Bigg|_{\arcsin(a)}^{\arcsin(b)}.$$

But

$$\sin(2\theta) = 2\sin(\theta)\cos(\theta) = 2\sin(\theta)\sqrt{1-\sin^2(\theta)} = 2x\sqrt{1-x^2}.$$

Thus

$$\int_a^b \sqrt{1-x^2}dx = \left(\frac{2x\sqrt{1-x^2}}{4}+\frac{\arcsin(x)}{2}\right)\Bigg|_a^b.$$

Taking the limit as $a \to -1$ and $b \to 1$, we have

$$\frac{2x\sqrt{1-x^2}}{4}\Bigg|_a^b \to 0,$$

and

$$\frac{\arcsin(x)}{2}\Bigg|_a^b \to \frac{\pi}{2}.$$

We conclude that

$$\int_{-1}^{+1} \sqrt{1-x^2}dx = \frac{\pi}{2},$$

which represents the area of the half-circle.

2.20 Uniformly convergent sequences of functions

The exponential function is defined in terms of the exponential series. But looking at the partial sums, we see a sequence of functions

$$f_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!}.$$

Each f_n is simply a polynomial, and for each $x \in \mathbb{R}$, the sequence of numbers $(f_n(x))_{n\in\mathbb{N}}$ converges to $\exp(x)$. That is to say, we have *point-wise convergence* of the sequence of functions

$$f_n \to \exp.$$

Nevertheless, this convergence is not *uniform*.[11]

Definition. *Let $U \subset \mathbb{R}$ and $f_n : U \to \mathbb{R}$ be a function for each $n \in \mathbb{N}$, giving a sequence $(f_n(x))_{n\in\mathbb{N}}$ of functions. The sequence will be called* uniformly convergent *if there exists some function $f : U \to \mathbb{R}$ such that for all $\epsilon > 0$ there exists an $N \in \mathbb{N}$, such that if $m \geq N$ then $|f_m(x) - f(x)| < \epsilon$ for all $x \in U$.*

Another way to look at this situation is to imagine that the set of all functions from U to $\mathbb{R}$ is an abstract "space". Within this space it is possible to establish a sensible idea of the "distance" between pairs of functions. For this we use the *supremum norm*.

Definition. *Let X be a set, and let $f : X \to \mathbb{R}$ be a bounded function on X. Then the* supremum norm *of f is*

$$\|f\|_X = \sup\{|f(x)| : x \in X\}.$$

It is obvious that the supremum norm satisfies the properties of a norm function, namely:

- $\|f\|_X \geq 0$,
- $\|\lambda f\|_X = |\lambda| \cdot \|f\|_X$, for all $\lambda \in \mathbb{R}$, and
- $\|f + g\|_X \leq \|f\|_X + \|g\|_X$, for any further bounded function $g : X \to \mathbb{R}$.

Then the *distance between* any two such bounded functions f, $g : X \to \mathbb{R}$ is given by

$$d(f, g) = \|f - g\|_X.$$

With this definition, we see that a uniformly convergent sequence of bounded functions is simply a convergent sequence with respect to the supremum norm. That is, for all $\epsilon > 0$ there exists some $N \in \mathbb{N}$ with $\|f_m - f\|_X < \epsilon$, for all $m \geq N$.

Theorem 2.47. *Let $U \subset \mathbb{R}$ be an interval, and let $f_n : U \to \mathbb{R}$ be a continuous function, for all $n \in \mathbb{N}$. If the sequence $(f_n(x))_{n\in\mathbb{N}}$ is uniformly convergent, with $f_n \to f$, then the function $f : U \to \mathbb{R}$ is also continuous.*

Proof. Let $x \in U$ and $\epsilon > 0$ be given. Choose $N \in \mathbb{N}$ to be so large that $|f_m(y) - f(y)| < \epsilon/3$, for all $m \geq N$ and $y \in U$. Since the function f_N is continuous,

[11] On the other hand, the series defining the exponential function is *locally uniformly convergent*. That is, for every $x \in \mathbb{R}$ there exists an $r > 0$ such that $f_n \to \exp$ uniformly in the interval $[x - r, x + r]$. (And of course the analogous statement is also true for the exponential function applied to complex numbers. Given any $z \in \mathbb{C}$, there exists an $r > 0$ such that the series $f_n \to \exp$ is uniformly convergent in the disc $\{w \in \mathbb{C} : |w - z| \leq r\}$.

there exists some $\delta > 0$ such that $|f_N(u) - f_N(x)| < \epsilon/3$ for all $u \in U$ with $|u - x| < \delta$. Then for all such u we have

$$\begin{aligned} |f(u) - f(x)| &\leq |f(u) - f_N(u)| + |f_N(u) - f_N(x)| + |f_N(x) - f(x)| \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\ &= \epsilon. \end{aligned}$$

□

Theorem 2.48. *Let $a < b$ and for each $n \in \mathbb{N}$ let $f_n : U \to \mathbb{R}$ be a Riemann integrable function. Assume that $f_n \to f$ uniformly. Then f is also Riemann integrable, and we have*

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

Proof. The fact that f is Riemann integrable is left as an exercise. To show that

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

let $\epsilon > 0$ be given, and let $N \in \mathbb{N}$ be sufficiently large that

$$\|f_m - f\|_{[a,b]} < \frac{\epsilon}{b-a}$$

for all $m \geq N$. Then we have

$$\begin{aligned} \left|\int_a^b f_m(x)dx - \int_a^b f(x)dx\right| &= \left|\int_a^b (f_m(x) - f(x)dx\right| \\ &\leq \int_a^b |f_m(x) - f(x)|dx \\ &< \frac{\epsilon}{b-a} \cdot (b-a) \\ &= \epsilon. \end{aligned}$$

□

Remark. *All of these results may be generalized to complex-valued functions. One need only take the absolute-value function in $\mathbb{C}$, rather than simply restricting it to $\mathbb{R}$.*

2.21 Taylor series; Taylor formula

2.21.1 The Taylor formula

Theorem 2.49 (Taylor's formula). *Let $f : [a, b] \to \mathbb{R}$ be an $(n+1)$-times continuously differentiable function defined on an open interval $(c, d) \supset [a, b]$. (That*

is, let $f'(x) = f^{(1)}(x)$*, and then recursively we define* $f^{(k+1)}(x) = \left(f^{(k)}(x)\right)'$*. Then the requirement is that* $f^{(n+1)}(x)$ *exists for all* x *in* $[a,b]$*, and the function* $f^{(n+1)} : [a,b] \to \mathbb{R}$ *which is so defined is continuous.) Then for any* x_0 *and* $x \in [a,b]$ *we have*

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + R_{n+1}(x),$$

where

$$R_{n+1}(x) = \frac{1}{n!}\int_{x_0}^{x}(x-t)^n f^{(n+1)}(t)dt.$$

Proof. Use induction on n. For $n = 0$, Taylor's formula is simply the fundamental theorem

$$f(x) = f(x_0) + \int_{x_0}^{x} f'(t)dt.$$

So now assume it is true for the case $n \geq 0$. In particular, we assume that the remainder term is

$$R_n(x) = \frac{1}{(n-1)!}\int_{x_0}^{x}(x-t)^{n-1} f^{(n)}(t)dt.$$

Applying partial integration, we obtain

$$\begin{aligned} R_n(x) &= \frac{1}{(n-1)!}\int_{x_0}^{x}(x-t)^{n-1} f^{(n)}(t)dt \\ &= -\int_{x_0}^{x} f^{(n)}(t)\left(\frac{(x-t)^n}{n!}\right)' dt \\ &= -f^{(n)}(t)\frac{(x-t)^n}{n!}\Bigg|_{x_0}^{x} + \int_{x_0}^{x}\frac{(x-t)^n}{n!} f^{(n+1)}(t)dt \\ &= \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \int_{x_0}^{x}\frac{(x-t)^n}{n!} f^{(n+1)}(t)dt, \end{aligned}$$

which is just the next term in the Taylor formula, with the corresponding remainder term.[12] □

We can also express the remainder term in a different way. Since $\frac{(x-t)^n}{n!}$ is always non-negative, we can use the mean value theorem for integrals to find some

[12] In the case that $x < x_0$ we use the general rule that for integrable functions $\phi : [a,b] \to \mathbb{R}$ we have $\int_a^b \phi(x)dx = -\int_b^a \phi(x)dx$, and the proof is then the same as when $x_0 < x$.

$\xi \in [x_0, x]$ with

$$\begin{aligned} R_{n+1}(x) &= \int_{x_0}^{x} \frac{(x-t)^n}{n!} f^{(n+1)}(t)dt \\ &= f^{(n+1)}(\xi) \int_{x_0}^{x} \frac{(x-t)^n}{n!} dt \\ &= -f^{(n+1)}(\xi) \frac{(x-t)^{n+1}}{(n+1)!}\Big|_{x_0}^{x} \\ &= \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}. \end{aligned}$$

Then Taylor's formula takes the simple form

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}.$$

2.21.2 The Taylor series

If f is infinitely differentiable[13] then we can consider the series

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n.$$

In fact, if you think about it, you will see that all of our standard functions are simply *defined* in terms of their Taylor series.

Back in the "old days", 200 years ago and more, mathematicians thought that the only sensible way to define the idea of a function was by means of a Taylor series. Yet, in modern mathematics, we see that there are many infinitely differentiable real functions which are different from their Taylor series. (Assuming that the series converges in the first place!)

On the other hand, things are quite different when we consider differentiable functions of complex numbers. There, all differentiable functions are always infinitely often differentiable, and furthermore, they are given by their Taylor series. The subject of complex analysis is called "Funktionentheorie" in German, paying tribute to this old-fashioned way of looking at functions.

2.21.3 The "standard" functions

At this stage of things, we take these to be: polynomials, the exponential function, the trigonometrical functions and the logarithm function.

But of these, the first three categories are simply *defined* in terms of power series, which are the Taylor series for the respective functions, taken at the point $x_0 = 0$.[14] Thus we need only look at the logarithm function.

[13] Of course this is the case with our "standard functions", namely polynomials, the exponential function, and the things which come out of that: sine, cosine, and so forth.

[14] polynomials can be thought of as being power series where all but a finite number of the terms are zero.

But since the logarithm is only defined for positive numbers, we need to look at the Taylor series at some point $x_0 > 0$. The natural idea is to take $x_0 = 1$. Then we have $\ln(1) = 0$, $\ln'(1) = 1$, and in general

$$\ln^{(n)}(1) = (-1)^{n-1}(n-1)!$$

This gives the Taylor series

$$\begin{aligned} T(x) &= 0 + (x - x_0) - \frac{(x-x_0)^2}{2!} + \frac{2!(x-x_0)^3}{3!} - \frac{3!(x-x_0)^4}{4!} + \cdots \\ &= (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \frac{(x-1)^4}{4} + \cdots \end{aligned}$$

It is often written

$$T(1+t) = t - \frac{t^2}{2} + \frac{t^3}{3} - \frac{t^4}{4} + \cdots$$

Another way to write it is

$$T(1-t) = -(t + \frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{4} + \cdots)$$

Note however that this series only converges for $|t| < 1$.

Is it true that the Taylor series for the logarithm does actually equal the logarithm function itself? That is, do we have

$$T(1+t) = \ln(1+t)$$

for all $|t| < 1$?

To answer this question, the simplest idea is to consider the sequence of functions

$$f_n(x) = \sum_{k=1}^{n} (-1)^{k-1} x^k.$$

Lemma. *For $|x| < 1$, the series $(f_n(x))_{n\in\mathbb{N}}$ is absolutely convergent. If $-1 < a < b < 1$ then the series of functions f_n, when restricted to $[a,b]$, converges uniformly.*

Proof.

$$\sum_{k=1}^{n} \left|(-1)^{k-1}x^k\right| = \sum_{k=1}^{n} |x|^k = \frac{1}{1-|x|}.$$

Therefore the convergence is absolute, and as a consequence the series of functions $(f_n(x))_{n\in\mathbb{N}}$ converges point-wise to a function f on the interval $(-1,+1)$. For

$x \in [a, b]$, we have

$$\begin{aligned} |f(x) - f_n(x)| &= \left| \sum_{k=n+1}^{\infty} (-1)^{k-1} x^k \right| \\ &\leq \sum_{k=n+1}^{\infty} |x|^k \\ &= |x|^{n+1} \sum_{k=0}^{\infty} |x|^k \\ &\leq M^{n+1} \frac{1}{1-M}, \end{aligned}$$

where $M = \max\{|a|, |b|\} < 1$. □

This result can be applied to the logarithm function. We have

$$\ln'(1+x) = \frac{1}{1+x}.$$

Therefore, using theorem 2.48, we have

$$\begin{aligned} \ln(1+x) &= \ln(1+t)\Big|_0^x \\ &= \int_0^x \left(\sum_{n=0}^{\infty} (-1)^n t^n \right) dt \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1} \\ &= \sum_{n=1}^{\infty} (-1)^{n-1} \frac{x^n}{n} \end{aligned}$$

for $|x| < 1$.

What can we do to represent $\ln(x)$ when $x \geq 1$? In this case we can find some y with $1 < y < 2$, and $y^k = x$, for some $k \in \mathbb{N}$. Then

$$\ln(x) = \ln(y^k) = k \cdot \ln(y) = k \cdot \left(\sum_{n=1}^{\infty} (-1)^{n-1} \frac{(y-1)^n}{n} \right).$$

For the special case $\ln(2)$, the Taylor series gives

$$T(2) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots$$

In order to see that, in fact, $T(2) = \ln(2)$, let us examine the remainder term in the Taylor formula. For this, we first observe that

$$\ln^{(n+1)}(1+\xi) = (-1)^n n! (1+\xi)^{-n}.$$

Then

$$R_{n+1}(2) = \frac{\ln^{(n+1)}(1+\xi)}{(n+1)!} (2-1)^{n+1} = \frac{1}{(n+1)(1+\xi)^n},$$

where $0 < \xi < 1$, and thus $R_{n+1}(2) \to 0$ as $n \to \infty$.

2.22 Improper integrals

We have only defined integrals for functions on closed intervals $[a, b]$, where $a < b$ are real numbers. This definition can be extended to include integrals of the form

$$\int_a^b f(x)dx,$$

where $b = -\infty$ and/or $a = \infty$. Namely, if

$$\lim_{R\to\infty} \int_a^R f(x)dx$$

exists, then the limit is taken to be $\int_a^\infty f(x)dx$. The other cases are defined analogously.

For example

$$\int_1^x \frac{dt}{t^2} = -t^{-1}\Big|_1^x = -\frac{1}{x} + 1.$$

Therefore

$$\int_1^\infty \frac{dt}{t^2} = 1.$$

However, the function $1/t^2 \to \infty$ as $t \to 0$. Can it be that the integral

$$\int_0^1 \frac{dt}{t^2} = \lim_{\epsilon\to 0} \int_\epsilon^1 \frac{dt}{t^2}$$

exists?

Obviously not, since

$$\int_\epsilon^1 \frac{dt}{t^2} = -1 + \frac{1}{\epsilon} \to \infty$$

as $\epsilon \to 0$.

On the other hand,

$$\int_\epsilon^1 \frac{dt}{\sqrt{t}} = 2\sqrt{t}\Big|_\epsilon^1 = 2 - 2\sqrt{\epsilon} \to 2$$

as $\epsilon \to 0$.

As a matter of fact, we have the following theorem.

Theorem 2.50. *The improper integral*

$$\int_0^\infty t^s dt$$

diverges for all $s \in \mathbb{R}$.

Proof. Otherwise, we could write

$$\int_0^\infty t^s dt = \int_0^1 t^s dt + \int_1^\infty t^s dt.$$

If $s = -1$, then the anti-derivative to t^{-1} is $\ln(t)$. And

$$\int_1^R \frac{dt}{t} = \ln(R) \to \infty$$

as $R \to \infty$.

If $s \neq -1$ then the anti-derivative to t^s is

$$\frac{t^{s+1}}{s+1}.$$

If $s < -1$ then

$$\int_\epsilon^1 t^s dt = \frac{1}{s+1}\left(1 - \epsilon^{s+1}\right) \to \infty,$$

since $(s+1) < 0$, and thus $\epsilon^{s+1} \to \infty$ as $\epsilon \to 0$.

If $s > -1$ then

$$\int_1^R t^s dt = \frac{1}{s+1}\left(R^{s+1} - 1\right) \to \infty$$

since $s + 1 > 0$, and thus $R^{s+1} \to \infty$ as $R \to \infty$. In all cases, the integrals do not converge. □

2.23 The integral comparison test for series

Theorem 2.51. *Let $a < b$ and $f : [a, b] \to \mathbb{R}$ be a monotonic function. Then the Riemann integral $\int_a^b f(x)dx$ exists.*

Proof. Assume without loss of generality that f is monotonically increasing; that is, $f(s) \leq f(t)$ for all $s \leq t$ in $[a, b]$. For each $n \in \mathbb{N}$, we can take the partition $a = x_0 < x_1 < \cdots < x_n = b$, where

$$x_k = a + k\frac{b-a}{n},$$

$k = 0, 1, \ldots, n$. Let the step functions g, $h : [a, b] \to \mathbb{R}$ be given by

$$g(x) = f(x_{k-1}), \text{ for } x_{k-1} \leq x < x_k,$$
$$h(x) = f(x_k), \text{ for } x_{k-1} \leq x < x_k,$$

and then $g(b) = h(b) = f(b)$. Therefore $g \leq f \leq h$. But

$$\int_a^b g(x)dx \leq \int_* f \leq \int^* f \leq \int_a^b h(x)dx$$

and

$$\begin{aligned}\int_a^b h(x)dx - \int_a^b g(x)dx &= \int_a^b (h(x) - g(x))dx \\ &= \sum_{k=1}^{n} \frac{b-a}{n}(f(x_k) - f(x_{k-1})) \\ &= \frac{b-a}{n}(f(b) - f(a)).\end{aligned}$$

Since both the numbers $f(b) - f(a)$ and $b - a$ are constant, and we can choose n to be arbitrarily large, it follows that we must have $\int_* f = \int^* f$, so that f is Riemann integrable. □

Theorem 2.52 (Integral comparison test). *Let $f : [1, \infty) \to \mathbb{R}$ be monotonically decreasing. Then the improper integral $\int_1^\infty f(x)dx$ exists if and only if the series $\sum_{n=1}^\infty f(n)$ converges.*

Proof. Obviously both the sum and the integral can only converge if $f(x) \geq 0$, for all $x \geq 1$. For each $n \in \mathbb{N}$ let g_n, $h_n : [1, n] \to \mathbb{R}$ be given by

$$\begin{aligned}g_n(x) &= f(k+1), \text{ for } k \leq x < k+1, \\ h_n(x) &= f(k), \text{ for } k \leq x < k+1,\end{aligned}$$

for $k = 1, \ldots, n-1$ and $g_n(n) = h_n(n) = f(n)$. Then

$$\int_1^n g_n(x)dx = \sum_{k=2}^{n} f(n)$$

and

$$\int_1^n h_n(x)dx = \sum_{k=1}^{n-1} f(n).$$

Therefore we see that $\sum_{n=1}^\infty f(n)$ converges if and only if both

$$\lim_{n\to\infty} \int_1^n g_n(x)dx$$

and

$$\lim_{n\to\infty} \int_1^n h_n(x)dx$$

exist. But $g_n \leq f \leq h_n$ on $[1, n]$, for each n. □

2.23.1 Riemann's zeta function

Since for each $x > 1$ we have

$$\int_1^\infty \frac{dt}{t^x} = \frac{1}{x-1},$$

it follows from the integral comparison test that

$$\zeta(x) = \sum_{n=1}^{\infty} n^{-x}$$

defines a function $\zeta : (1, \infty) \to \mathbb{R}$. As an exercise, we see that the series also converges for all complex numbers $z = x + iy$ with $x > 1$. Riemann proved that the zeta function satisfies a certain functional equation which allows it to be "analytically continued" throughout the whole of the complex plane except for the obvious singularity at $z = 1$. This function plays a central role in the subject of analytic number theory.

There are certain relationships with the Gamma function, which also has a functional equation, allowing an analytic continuation within the theory of complex analysis.

2.24 The Gamma function

For $x > 0$, one writes

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt.$$

Theorem 2.53. *The integral defining the Gamma function converges for all $x > 0$.*

In order to prove this, let us begin with a simple lemma.

Lemma. *For all $n \in \mathbb{N}$, we have*

$$\lim_{t \to \infty} \frac{e^t}{t^n} = \infty.$$

Proof. Since

$$e^t = 1 + t + \cdots + \frac{t^{n+1}}{(n+1)!} + \cdots,$$

it follows that

$$\frac{e^t}{t^n} > \frac{t}{(n+1)!}$$

for all $t > 0$, and thus

$$\frac{e^t}{t^n} \to \infty \text{ as } t \to \infty.$$

□

Therefore we must also have

$$\lim_{t \to \infty} \frac{t^n}{e^t} \to 0 \text{ as } t \to \infty.$$

Proof of theorem 2.53. Let $x > 0$ be given, and let $t_0 > 0$ be sufficiently large that $t^{x+1}e^{-t} < 1$, for all $t \geq t_0$. Then

$$t^{x-1}e^{-t} < t^{-2} \text{ for all } t \geq t_0.$$

Therefore

$$\int_{t_0}^{\infty} t^{x-1}e^{-t}dt < \int_{t_0}^{\infty} \frac{dt}{t^2} = \frac{1}{t_0},$$

so that the integral from t_0 to ∞ converges.[15]

On the other hand we have $e^{-t} < 1$ for all $t > 0$, therefore $t^{x-1}e^{-t} < t^{x-1}$, and

$$\begin{aligned} \int_{\epsilon}^{t_0} t^{x-1}e^{-t}dt &< \int_{\epsilon}^{t_0} t^{x-1}dt \\ &= \frac{t^x}{x}\Big|_{\epsilon}^{t_0} = \frac{t_0^x}{x} - \frac{\epsilon^x}{x} \\ &\to \frac{t_0^x}{x} \text{ as } \epsilon \to 0. \end{aligned}$$

Thus also the integral from 0 to t_0 converges. □

2.24.1 The functional equation for the Gamma function

For $0 < \epsilon < R < \infty$, partial integration gives

$$\int_{\epsilon}^{R} t^x e^{-t}dt = -t^x e^{-t}\Big|_{\epsilon}^{R} + \int_{\epsilon}^{R} xt^{x-1}e^{-t}dt.$$

However, for $x > 0$, we have $\lim_{\epsilon \to 0} \epsilon^x e^{-\epsilon} = 0$ and $\lim_{R \to 0} R^x e^{-R} = 0$.

Therefore we obtain the functional equation for the Gamma function:

$$\Gamma(x+1) = \int_0^{\infty} t^x e^{-t}dt = x\int_0^{\infty} t^{x-1}e^{-t}dt = x\Gamma(x).$$

In particular we have

$$\Gamma(1) = \int_0^{\infty} t^{1-1}e^{-t}dt = \int_0^{\infty} e^{-t}dt = -e^{-t}\Big|_0^{\infty} = 1.$$

It follows that

$$\Gamma(n+1) = n!,$$

for all $n \in \mathbb{N}$.

For $n = 0$ we have $\Gamma(0+1) = \Gamma(0) = 1$. But then the functional equation gives $\Gamma(0+1) = 1 = 0 \cdot \Gamma(0)$. This is impossible! So the Gamma function is not defined at 0. And therefore it is also not defined at all of the negative integers $-n$, for $n \in \mathbb{N}$.

[15] We have $t^{x-1}e^{-t} > 0$ for $t \geq t_0$. Thus $\int_{t_0}^{R} t^{x-1}e^{-t}dt$ increases as R increases.

On the other hand, if we take say $x = 1/2$ then we have $\Gamma(1/2)$ being some positive number. In fact, it turns out that $\Gamma(1/2) = \sqrt{\pi}$. Then, using the functional equation, we find that $\Gamma(-1/2) = -2\sqrt{\pi}$, $\Gamma(-3/2) = 4\sqrt{\pi}/3$, and so forth. In fact, apart from zero and the negative integers, the Gamma function is defined for all of $\mathbb{R}$.

2.24.2 The Gamma function in complex analysis

For $z = x + iy \in \mathbb{C}$, we can write

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

It can be shown that the integral converges when the real part of z is positive, namely $x > 0$. We again obtain the same functional equation, namely $\Gamma(z+1) = z\Gamma(z)$, and this allows us to analytically continue the definition of the Gamma function into the region of the complex plane where the numbers have real part negative or zero. It turns out that the function can be defined everywhere except at zero and the negative real integers. These are points of singularity of the function. Apart from these points, the Gamma function is everywhere differentiable. There are various interesting formulas which can be proved. For example, we have

$$\Gamma(z) = \lim_{n\to\infty} \frac{n^z n!}{z(z+1)\cdots(z+n)}.$$

In order to get a feel for how such formulas can be proved, let us again look at the situation for $\Gamma(x)$, for $x > 0$.

2.24.3 Two formulas

Theorem 2.54. *For $t \in \mathbb{R}$, $n \in \mathbb{N}$ we have*

$$\lim_{n\to\infty} \left(1 + \frac{t}{n}\right)^n = e^t.$$

Proof.

$$\begin{aligned}
\ln\left(\lim_{n\to\infty}\left(1+\frac{t}{n}\right)^n\right) &= \lim_{n\to\infty} \ln\left(1+\frac{t}{n}\right)^n \\
&= \lim_{n\to\infty} n \ln\left(1+\frac{t}{n}\right) \\
&= t \lim_{n\to\infty} \frac{n}{t} \ln\left(1+\frac{t}{n}\right) \\
&= t \lim_{n\to\infty} \frac{\ln\left(1+\frac{t}{n}\right) - \ln(1)}{\frac{t}{n}} \\
&= t \ln'(1) = t
\end{aligned}$$

The first equation follows from the continuity of the logarithm function; the fourth equation follows since $\ln(1) = 0$, and of course the fifth equation results from $\ln'(x) = 1/x = 1$, when $x = 1$.

Therefore

$$e^{\ln\left(\left(\lim_{n\to\infty}\left(1+\frac{t}{n}\right)^n\right)\right)} = \lim_{n\to\infty}\left(1+\frac{t}{n}\right)^n = e^t.$$

□

Theorem 2.55. *For $x > 0$ we have*

$$\int_0^n \left(1-\frac{t}{n}\right)^n t^{x-1}dt = \frac{n^x n!}{x(x+1)\cdots(x+n)}.$$

Proof. Progressively using partial integration, we have

$$\begin{aligned}
\int_0^n \left(1-\frac{t}{n}\right)^n t^{x-1}dt &= \frac{t^x}{x}\left(1-\frac{t}{n}\right)^n\Big|_0^n - \int_0^n \frac{n}{-n}\left(1-\frac{t}{n}\right)^{n-1}\frac{t^x}{x}dt \\
&= \frac{n}{xn}\int_0^n \left(1-\frac{t}{n}\right)^{n-1} t^x dt \\
&= \frac{n}{xn}\left(\frac{t^{x+1}}{x+1}\left(1-\frac{t}{n}\right)^{n-1}\Big|_0^n - \int_0^n \frac{n-1}{-n}\left(1-\frac{t}{n}\right)^{n-2}\frac{t^{x+1}}{x+1}dt\right) \\
&= \frac{n(n-1)}{x(x+1)n^2}\int_0^n \left(1-\frac{t}{n}\right)^{n-2} t^{x+1}dt \\
&\vdots \\
&= \frac{n(n-1)\cdots(n-(n-1))}{x(x+1)\cdots(x+(n-1))n^n}\int_0^n t^{x+(n-1)}dt \\
&= \frac{n(n-1)\cdots(n-(n-1))}{x(x+1)\cdots(x+(n-1))n^n}\left(\frac{t^{x+n}}{x+n}\Big|_0^n\right) \\
&= \frac{n^x n!}{x(x+1)\cdots(x+n)}.
\end{aligned}$$

Notice here that we always have the expressions of the form

$$\frac{t^{x+k}}{x+k}\left(1-\frac{t}{n}\right)^{n-k}\Big|_0^n$$

being zero, so that only the integral is carried through from one line to the next. □

It is now an exercise to show that the convergence given by theorem 2.54 is uniform, and that the improper integral defining the Gamma function also converges when theorem 2.55 is used. All of this gives us the formula

$$\Gamma(x) = \lim_{n\to\infty}\frac{n^x n!}{x(x+1)\cdots(x+n)},$$

for $x \in \mathbb{R}$ with $x > 0$.

2.25 Convexity

Definition. *Let $D \subset \mathbb{R}$ and $f : D \to \mathbb{R}$ be some function. An element $x_0 \in D$, together with its value $f(x_0)$ under f, is called an* isolated local minimum *of the function if there exists some $\delta > 0$ such that for all $x \in D$ with $|x - x_0| < \delta$ and $x \neq x_0$, we have $f(x) > f(x_0)$. The idea of an isolated local* maximum *is defined analogously.*

Theorem 2.56. *Let $f : (a, b) \to \mathbb{R}$ be a differentiable function which is twice differentiable at the point $x_0 \in (a, b)$, such that $f'(x_0) = 0$ and $f''(x_0) > 0$. Then $f(x_0)$ is an isolated local minimum of the function f. If $f''(x_0) < 0$ then $f(x_0)$ is an isolated local maximum.*

Proof. Since

$$f''(x_0) = \lim_{\xi \to x_0} \frac{f'(\xi) - f'(x_0)}{\xi - x_0},$$

there must exist some $\epsilon > 0$ such that for all $\xi \in (a, b)$ with $|\xi - x_0| < \epsilon$ and $\xi \neq x_0$ we have

$$\frac{f'(\xi) - f'(x_0)}{\xi - x_0} > 0.$$

But $f'(x_0) = 0$. Therefore we must have $f'(\xi) < 0$ if $\xi < x_0$ and $f'(\xi) > 0$ if $\xi > x_0$. It then follows from the mean value theorem (2.34) that f is strictly monotonically decreasing for $\xi < x_0$, and strictly monotonically increasing for $\xi > x_0$.

The result when $f''(x_0) < 0$ follows analogously. □

Definition. *Let $D \subset \mathbb{R}$ be an interval. (Thus if $a < b$ are two points of D, and x is some point with $a < x < b$, then we must also have $x \in D$ as well.) A function $f : D \to \mathbb{R}$ is called* convex *if for any two points a, $b \in D$, and for any λ with $0 \leq \lambda \leq 1$, we have*

$$f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda) f(b).$$

Theorem 2.57. *Assume that D is an open interval, and let $f : D \to \mathbb{R}$ be twice differentiable. Then f is convex $\Leftrightarrow$ $f''(x) \geq 0$ for all $x \in D$.*

Proof. "$\Rightarrow$" In order to produce a contradiction, assume that f is convex and that there exists some $x_0 \in D$ with $f''(x_0) < 0$. Let $g : D \to \mathbb{R}$ be given by

$$g(x) = f(x) - f'(x_0)(x - x_0),$$

for all $x \in D$. Then we have $g'(x_0) = 0$ and $g''(x_0) = f''(x_0) < 0$. Therefore, $g(x_0)$ is an isolated local maximum of the function g. That is, there exists some $\delta > 0$ with $g(x) < g(x_0)$ for all $x \neq x_0$ in D, with $|x - x_0| \leq \delta$. But then

$$f(x_0) = g(x_0) > \frac{1}{2}(g(x_0 - \delta) + g(x_0 + \delta)) = \frac{1}{2}(f(x_0 - \delta) + f(x_0 + \delta)),$$

contradicting the fact that f is convex.

"$\Leftarrow$" Since $f''(x) \geq 0$ for all $x \in D$, we must have the function f' monotonically increasing. So let $a < b$ be two points of D, and let $0 < \lambda < 1$ be given. Take

$$x = \lambda a + (1 - \lambda)b.$$

Then, according to the mean value theorem, there exist two further points ξ_1, ξ_2 with $a < \xi_1 < x < \xi_2 < b$, such that

$$\frac{f(x) - f(a)}{x - a} = f'(\xi_1) \leq f'(\xi_2) = \frac{f(b) - f(x)}{b - x}.$$

But $x - a = (1 - \lambda)(b - a)$ and $b - x = \lambda(b - a)$. Thus

$$\frac{f(x) - f(a)}{(1 - \lambda)(b - a)} \leq \frac{f(b) - f(x)}{\lambda(b - a)},$$

or

$$\lambda(f(x) - f(a)) \leq (1 - \lambda)(f(b) - f(x)).$$

That is,

$$f(x) = f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)f(b).$$

Since a, b and λ were chosen arbitrarily, it follows that f must be convex. □

Similarly, one proves that f is concave if, and only if, $f''(x) \leq 0$ for all $x \in D$.

Theorem 2.58. *Let p, $q > 1$ such that $1/p + 1/q = 1$. Then for all x, $y > 0$ we have*

$$x^{\frac{1}{p}} y^{\frac{1}{q}} \leq \frac{x}{p} + \frac{y}{q}.$$

Proof. Since for all $x > 0$ we have $\ln''(x) = -1/x^2 < 0$, it follows that the logarithm must be a concave function, and we have

$$\ln(\lambda x + (1 - \lambda)y) \geq \lambda \ln(x) + (1 - \lambda)\ln(y).$$

In particular, this is true for $\lambda = 1/p$, and thus $1 - \lambda = 1/q$. The result then follows by applying the exponential function to both sides and remembering that the exponential function is monotonically increasing. □

Corollary (Young's inequality). *For x and y non-negative real numbers and $p > 1$ with $1/p + 1/q = 1$, we have*

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}.$$

Definition. *Let K be either the real numbers $\mathbb{R}$, or else the complex numbers $\mathbb{C}$. For any vector $x = (x_1, x_2, \ldots, x_n) \in K^n$, and $p \geq 1$, the* p-norm *of x is*

$$\|x\|_p = \left(\sum_{k=1}^{n} |x_k|^p \right)^{\frac{1}{p}}.$$

Theorem 2.59. *Let $x = (x_1, x_2, \ldots, x_n) \in K^n$ be a vector ($K = \mathbb{R}$ or $\mathbb{C}$), and $p > 1$. Then for $q > 1$ with $1/p + 1/q = 1$ and also $y = (y_1, y_2, \ldots, y_n) \in K^n$, we have*

$$\sum_{k=1}^{n} |x_k y_k| \leq \|x\|_p \|y\|_q.$$

Proof. If $\|x\|_p = 0$ or $\|y\|_q = 0$ then the result is trivial. Otherwise, for $k = 1, \ldots, n$ let

$$u_k = \frac{|x_k|^p}{\|x\|_p^p}, \text{ and } v_k = \frac{y_k|^q}{\|y\|_q^q}.$$

We have

$$\sum_{k=1}^{n} u_k = \frac{1}{\|x\|_p^p} \left(|x_1|^p + \cdots + |x_n|^p\right) = \frac{1}{\sum_{k=1}^{n} |x_k|^p} \left(\sum_{k=1}^{n} |x_k|^p\right) = 1.$$

Similarly, $\sum_{k=1}^{n} v_k = 1$. But

$$\frac{|x_k y_k|}{\|x\|_p \|y\|_q} = u_k^{\frac{1}{p}} v_k^{\frac{1}{q}} \leq \frac{1}{p} u_k + \frac{1}{q} v_k.$$

Therefore

$$\sum_{k=1}^{n} \frac{|x_k y_k|}{\|x\|_p \|y\|_q} \leq \frac{1}{p} \sum_{k=1}^{n} u_k + \frac{1}{q} \sum_{k=1}^{n} v_k = \frac{1}{p} + \frac{1}{q} = 1.$$

And finally,

$$\sum_{k=1}^{n} |x_k y_k| \leq \|x\|_p \|y\|_q.$$

□

In particular, when $p = q = 2$ we obtain the familiar Cauchy-Schwarz inequality of linear algebra:

$$|\langle x, y \rangle| = \left|\sum_{k=1}^{n} \overline{x_k} y_k\right| \leq \sum_{k=1}^{n} |x_k y_k| \leq \left(\sum_{k=1}^{n} x_k^2\right)^{\frac{1}{2}} \left(\sum_{k=1}^{n} y_k^2\right)^{\frac{1}{2}} = \|x\| \cdot \|y\|.$$

Definition. *Let $a < b$ and $f : [a, b] \to \mathbb{R}$ be Riemann integrable. For $p \geq 1$, we define*

$$\|f\|_p = \left(\int_a^b |f(x)|^p dx\right)^{\frac{1}{p}}.$$

Theorem 2.60 (Hölder's inequality). *Let $a < b$ and f, $g : [a, b] \to \mathbb{R}$ be Riemann integrable. For $p > 1$, we have*

$$\int_a^b |f(x)g(x)| dx \leq \|f\|_p \|g\|_q,$$

where $1/p + 1/q = 1$.

Proof. For each $n \in \mathbb{N}$, take the partition $a = x_0 < x_1 < \cdots < x_n = b$ of the interval $[a, b]$ where $x_k = a + \frac{k(b-a)}{n}$, for each k, and consider the Riemann sum

$$\sum_{k=1}^{n} |f(\xi_k)g(\xi_k)|(x_k - x_{k-1}) = \frac{b-a}{n} \sum_{k=1}^{n} |f(\xi_k)g(\xi_k)|,$$

where $x_{k-1} < \xi_k < x_k$, for each k. Then according to theorem 2.59, we have

$$\begin{aligned} \frac{b-a}{n}\left(\sum_{k=1}^{n} |f(\xi_k)g(\xi_k)|\right) &\leq \frac{b-a}{n}\left(\sum_{k=1}^{n} |f(\xi_k)|^p\right)^{\frac{1}{p}}\left(\sum_{k=1}^{n} |g(\xi_k)|^q\right)^{\frac{1}{q}} \\ &= \left(\frac{b-a}{n}\right)^{\frac{1}{p}+\frac{1}{q}}\left(\sum_{k=1}^{n} |f(\xi_k)|^p\right)^{\frac{1}{p}}\left(\sum_{k=1}^{n} |g(\xi_k)|^q\right)^{\frac{1}{q}} \\ &= \left(\frac{b-a}{n}\sum_{k=1}^{n} |f(\xi_k)|^p\right)^{\frac{1}{p}}\left(\frac{b-a}{n}\sum_{k=1}^{n} |g(\xi_k)|^q\right)^{\frac{1}{q}} \end{aligned}$$

According to the definition of the Riemann integral, we have both

$$\sum_{k=1}^{n} |f(\xi_k)g(\xi_k)|(x_k - x_{k-1}) = \frac{b-a}{n} \sum_{k=1}^{n} |f(\xi_k)g(\xi_k)| \to \int_a^b |f(x)g(x)|dx$$

and also

$$\frac{b-a}{n}\sum_{k=1}^{n} |f(\xi_k)|^p \to \int_a^b |f(x)|^p dx \quad \text{and} \quad \frac{b-a}{n}\sum_{k=1}^{n} |g(\xi_k)|^q \to \int_a^b |g(x)|^q dx,$$

for $n \to \infty$. □

Theorem 2.61. *Let $p \geq 1$ and x, $y \in K^n$ (where, again, K is either $\mathbb{R}$ or $\mathbb{C}$). Then we have*

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p.$$

Proof. If $p = 1$, then this is just the triangle inequality for the absolute value function. Therefore assume $p > 1$, and let $1/p + 1/q = 1$. Define $z \in K^n$ by $z = (z_1, \ldots, z_n)$, where $z_k = |x_k + y_k|^{p-1}$ for each k. We have $\frac{1}{p} + \frac{1}{q} = 1$. That is, $\frac{q}{p} + 1 = q$, or $q(p-1) = p$. Therefore

$$z_k^q = |x_k + y_k|^{q(p-1)} = |x_k + y_k|^p,$$

and so $\|z\|_q = \|x + y\|_p^{p/q}$.

According to theorem 2.59, and using the triangle inequality, we have

$$\begin{aligned}
\|x+y\|_p^p &= \sum_{k=1}^n |x_k+y_k|^p \\
&= \sum_{k=1}^n |x_k+y_k||z_k| \\
&\leq \sum_{k=1}^n |x_k z_k| + \sum_{k=1}^n |y_k z_k| \\
&\leq (\|x\|_p + \|y\|_p)\,\|z\|_q \\
&= (\|x\|_p + \|y\|_p)\,\|x+y\|_p^{p/q} \\
&= (\|x\|_p + \|y\|_p)\,\|x+y\|_p^{p-1}.
\end{aligned}$$

The last equality here follows from the observation that $\frac{p}{q} = p-1$. Finally, dividing by $\|x+y\|_p^{p-1}$ gives the result. (Of course the theorem is trivially true if $\|x+y\|_p^{p-1} = 0$.) □

This shows that the mapping $\|\cdot\|_p : K^n \to K$ is, in fact, a norm. For we certainly have $\|x\|_p \geq 0$ for all $x \in K^n$, and $x = 0$ if and only if $\|x\|_p = 0$. Also $\|\lambda x\|_p = |\lambda|\|x\|_p$ is easy to verify. Then finally, Minkowski's inequality gives us the triangle inequality for a norm.

In analogy to the proof of theorem 2.60, we can extend theorem 2.61 to integrals.

Theorem 2.62 (Minkowski's inequality). *Again let $a < b$ and $f,\ g : [a,b] \to \mathbb{R}$ be Riemann integrable. For $p \geq 1$, we have*

$$\|f+g\|_p \leq \|f\|_p + \|g\|_p.$$

It is important to note however, that we can have $\|f\|_p = 0$ even when f is not the zero function. For example, take the function $f : [-1,+1] \to \mathbb{R}$ to be given by

$$f(x) = \begin{cases} 1, & \text{if } x = 1 \\ 0, & \text{if } x \neq 1. \end{cases}$$

Then clearly $\|f\|_p = 0$, for all $p \geq 1$, yet f is not the zero function. Thus, strictly speaking, $\|\cdot\|_p$ is not a norm on the space of integrable functions on a given interval.

If you continue on to the Analysis 3 lecture, then you will learn more about such things.

Chapter 3

Analysis 2

3.1 Metric spaces

Definition. *Let M be some arbitrary non-empty set. A mapping*

$$d : M \times M \to \mathbb{R}$$

is called a metric *on M if it satisfies the following three properties.*

- $d(x, y) = 0$ *if, and only if* $x = y$.
- $d(x, y) = d(y, x)$, *for all* x *and* y *in* M. *(*Symmetry*)*
- $d(x, z) \leq d(x, y)+d(y, z)$, *for all* x, y, *and* $z \in M$. *(*The triangle inequality*)*

We can think of this function "d" as giving us a sort of abstract "distance" function within the set M. Obviously the distance function in our usual 3-dimensional space of everyday experience is a metric.

Theorem 3.1. *Given a metric d, defined on a set M, then we have $d(x, y) \geq 0$, for all x, $y \in M$.*

Proof. This follows trivially from the fact that

$$0 = d(x, x) \leq d(x, y) + d(y, x) = 2d(x, y),$$

for all x, $y \in M$. □

Examples

- The real numbers $\mathbb{R}$, together with the metric given by $d(x, y) = |x - y|$.
- The complex numbers $\mathbb{C}$ with $d(w, z) = |w - z|$.
- Let V be any normed vector space, with norm $\|\cdot\| : V \to \mathbb{R}$. Then $d(u, v) = \|u-v\|$ is the corresponding metric on V. Therefore we see that *any* normed vector space — for example our usual $\mathbb{R}^3$ — is automatically also a metric space.

- Let $C_0([a,b],\mathbb{R})$ be the set of continuous real-valued functions defined on an interval $[a,b] \subset \mathbb{R}$. Then, as we have already seen, $C_0([a,b],\mathbb{R})$ can be considered with the norm $\|f\| = \sup\{|f(x)| : a \leq x \leq b\}$. Therefore $C_0([a,b],\mathbb{R})$ is also a metric space.

- The 2-sphere $S^2 = \{x \in \mathbb{R}^3 : \|x\| = 1\}$, together with the metric given by $d(x,y) = \|x - y\|$, where the norm here is simply the usual Euclidean norm of $\mathbb{R}^3$.

3.1.1 Open sets, closed sets

Definition. *Let (M,d) be a set, together with a metric. Given any number $\epsilon > 0$, and any $x \in M$, then the* open ball *around x with radius ϵ is the subset*

$$B(x,\epsilon) = \{y \in M : d(y,x) < \epsilon\}.$$

In general, a subset $U \subset M$ will be called open*, if for all $x \in U$, there exists some $\epsilon_x > 0$ (depending on x), such that $B(x,\epsilon_x) \subset U$. A subset $A \subset M$ will be called* closed *if the compliment $M \setminus A$ is open.*

Obviously, if $\epsilon_1 < \epsilon_2$ are two positive numbers, then $B(x,\epsilon_1) \subset B(x,\epsilon_2)$.

Examples

- Given any metric space (M,d), then the empty set $\emptyset$ is both open and closed. Also the whole set M is always both open and closed.

- Within the real numbers $\mathbb{R}$, together with the usual metric, we have that all open intervals (a,b) are open, and furthermore, all closed intervals $[a,b]$ are closed. On the other hand, intervals which are half open and half closed, such as $(a,b]$, are neither open nor closed. But for example (a,∞) is open, while $[a,\infty)$ is closed.

- If (M,d) is an arbitrary metric space, then any subset consisting of a single element $\{x\} \subset M$ is closed. To see this, it is only necessary to observe that we have $B(y,d(x,y)/2) \subset (M \setminus \{x\})$, for all $y \neq x$.

  By extension, we have that any *finite* subset of M must also be closed.

Theorem 3.2. *An arbitrary union of open subsets is open. On the other hand, we can only say that every* finite *union of closed subsets is closed.*

Proof. Let (M,d) be a metric space. Let $U_i \subset M$ be open subsets of M, indexed by some "index set" I, so that $i \in I$ for all i. This index set might be infinitely large, even of some higher order of infinity. The problem is then to show that $\bigcup_{i \in I} U_i \subset M$ is open.

But given any $x \in \bigcup_{i\in I} U_i$ then, in particular, $x \in U_i$ for some $i \in I$. But since U_i is open in M, there exists some $\epsilon_x > 0$ such that $B(x, \epsilon_x) \subset U_i \subset \bigcup_{i\in I} U_i$. Therefore, since x was arbitrarily chosen, it follows that $\bigcup_{i\in I} U_i$ is open in M.

Now let $A_1, \ldots, A_n$ be a finite collection of closed subsets of M. The problem is to show that $M \setminus \bigcup_{k=1}^n A_k$ is open. So let $x \in (M \setminus \bigcup_{k=1}^n A_k)$. However, since each A_k is closed, it follows that each $(M \setminus A_k)$ is open. Furthermore, $x \in (M \setminus A_k)$ for all $k = 1, \ldots, n$. Therefore, for each k, there is some $\epsilon_k > 0$ with $B(x, \epsilon_k) \subset (M \setminus A_k)$. Choose $\epsilon = \min\{\epsilon_k : k = 1, \ldots, n\}$. Then we have $B(x, \epsilon) \subset (M \setminus \bigcup_{k=1}^n A_k)$. Since x was arbitrary, it follows that $M \setminus \bigcup_{k=1}^n A_k$ is open, hence $\bigcup_{k=1}^n A_k$ is closed. □

Thinking about the basic relationships of set theory leads to the following corollary.

Corollary. *Arbitrary intersections of closed sets are closed. Finite intersections of open sets are open.*

Definition. *Let (M, d) be a metric space, and let $\emptyset \neq V \subset M$ be a subset. Denote by $\overline{V}$ the intersection of all closed subsets of M which contain V. The set $\overline{V}$ is called the* closure *of V in M.*

Note that since M itself is always closed, this intersection is not empty. Thus $\overline{V}$ is a closed set, and it is the smallest closed set containing V. Obviously, if V is already closed, then we have $V = \overline{V}$

Definition. *Given some subset $V \subset M$, a point $x \in M$ is said to be on the* boundary *of V if for all $\epsilon > 0$, the open ball $B(x, \epsilon)$ at x with radius ϵ contains points of V and also points of $M \setminus V$. That is*

$$V \cap B(x, \epsilon) \neq \emptyset \neq (M \setminus V) \cap B(x, \epsilon)$$

for all ϵ. The set of all boundary points of V is denoted by ∂V.

Theorem 3.3. *Given any V with $\emptyset \neq V \subset M$, then we have $\overline{V} = V \cup \partial V$.*

Proof. Let $x \in M$ with $x \notin V \cup \partial V$. Since $x \notin \partial V$, there must exist some $\epsilon > 0$ with $B(x, \epsilon) \cap V = \emptyset$. But also $B(x, \epsilon/2) \cap \partial V = \emptyset$, since for any point $y \in \partial V$, we have $B(y, \epsilon/2) \cap V \neq \emptyset$, and if $y \in B(x, \epsilon/2)$, then we would have $B(x, \epsilon) \cap V \neq \emptyset$, which is impossible. Since x was chosen arbitrarily in $M \setminus (V \cup \partial V)$, it follows that $M \setminus (V \cup \partial V)$ must be open. Thus $V \cup \partial V$ must be closed.

Now take W to be any closed subset of M with $V \subset W$. Let $y \in \partial V$ be an arbitrary point of the boundary of V. If $y \notin W$ then we would have $y \in (M \setminus W)$, which is an open set, since W is closed. But then there would be an $\epsilon > 0$ such that $B(y, \epsilon) \subset (M \setminus W) \subset (M \setminus V)$. But this is impossible, since then y would not be on the boundary of V. Therefore $y \in W$, and we must have that $V \cup \partial V$ is contained within every closed set which contains V. That is, $\overline{V} = V \cup \partial V$. □

3.1.2 Compact sets

The ideas of convergence which we have already encountered in Analysis I can be generalized to the case of arbitrary metric spaces.

Definition. *Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of elements in the metric space (M,d). Then the sequence* converges *to the element $x \in M$ if for all $\epsilon > 0$, a sufficiently large $N \in \mathbb{N}$ exists, such that $d(x_n, x) < \epsilon$, for all $n \geq N$.*

We will say that x is a cluster point *(or "Häufungspunkt") of the sequence if for every $\epsilon > 0$, there exist infinitely many elements of the sequence contained within $B(x,\epsilon)$. Thus a cluster point is the limit point of a convergent subsequence.*

Definition. *Let $K \subset M$ be some subset of a metric space (M,d). We will say that K is* sequentially compact *if for every sequence in K there exists a cluster point in K.*

Examples

- Any subset consisting of only finitely many elements must be sequentially compact.
- In $\mathbb{R}$, any open set such as (a,b) is *not* sequentially compact. For example, consider the open interval $(0,2)$. Then the sequence $(1/n)_{n\in\mathbb{N}}$ is contained in $(0,2)$, yet the only possible cluster point, namely the number 0, is not contained in $(0,2)$. Similarly, the entire set $\mathbb{R}$ is not sequentially compact. For example the sequence $(n)_{n\in\mathbb{N}}$ has no cluster points.

Theorem 3.4 (Heine-Borel: 1-dimensional version)**.** *A subset $K \subset \mathbb{R}$ is sequentially compact if and only if it is closed and bounded.*

Proof. Let $K \subset \mathbb{R}$ be sequentially compact. If K were not bounded, then for each $n \in \mathbb{N}$ there would be some $x_n \in K$ with $|x_n| > n$. Clearly the sequence $(x_n)_{n\in\mathbb{N}}$ has no cluster points. Thus K must be bounded. Assume now that K is not closed. That is, $M \setminus K$ is not open. Therefore there must be some $x \in (M \setminus K)$, such that for all $\epsilon > 0$, $B(x,\epsilon) \not\subset (M \setminus K)$. Or, put another way, for all $\epsilon > 0$, there exists some $y_\epsilon \in K$ with $d(x, y_\epsilon) < \epsilon$. In particular, there exists a sequence $(z_n)_{n\in\mathbb{N}}$ with $z_n \in K$ for all n, and $d(x, z_n) < 1/n$.

Could it be that there exists some cluster point $v \in K$ for this sequence? If so, then since $x \notin K$, we must have $x \neq v$, and so $d(x,v) > 0$. Since the sequence $(z_n)_{n\in\mathbb{N}}$ converges to x, there exists some $N \in \mathbb{N}$ with $d(z_n, x) < d(x,v)/2$, for all $n \geq N$. But then for all such n, we must have $z_n \notin B(v, d(x,v)/2)$, since otherwise we would have

$$d(x,v) \leq d(x,z_n) + d(z_n,v) < \frac{d(x,v)}{2} + \frac{d(x,v)}{2} = d(x,v).$$

But this is impossible, since $d(x,v) > 0$.

On the other hand, assume that $K \subset \mathbb{R}$ is closed and bounded, and let $(a_n)_{n\in\mathbb{N}}$ be some sequence in K. Since the sequence is bounded, there must exist some convergent subsequence (Bolzano-Weierstraß, theorem 2.4). Therefore there is some cluster point $x \in \mathbb{R}$ of the sequence. We must show that $x \in K$. If not, then x is in the open set $\mathbb{R} \setminus K$. But then there would exist some $\epsilon > 0$ such that $B(x, \epsilon) \cap K = \emptyset$. However, since all the elements a_n are contained in K, it would follow that x could not be a cluster point of the sequence $(a_n)_{n\in\mathbb{N}}$. A contradiction. □

With only a few changes, the same proof also works in the case of $\mathbb{R}^n$. For this we begin by observing that if typical points of $\mathbb{R}^n$ are $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, then

$$d(x, y) = \|x - y\| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

From this, one sees that for each $i = 1, \ldots, n$ we have

$$|x_i - y_i| \leq d(x, y).$$

Therefore let $(a_m)_{m\in\mathbb{N}}$ be a sequence of points in $\mathbb{R}^n$. For each m, we have that $a_m = (a_{m1}, \ldots, a_{mn})$, with $a_{ij} \in \mathbb{R}$ for each $m \in \mathbb{N}$ and $j \in \{1, \ldots, n\}$. Then if the sequence converges to a point $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, it follows that we have $\lim_{m\to\infty} a_{mj} = x_j$ for each j. Thus it is clear that a sequence $(a_m)_{m\in\mathbb{N}}$ in $\mathbb{R}^n$ converges if and only if the n separate sequences of coordinates $(a_{mj})_{m\in\mathbb{N}}$ each converge in $\mathbb{R}$, for $j = 1, \ldots, n$. Slightly more general is the question of convergent subsequences.

Theorem 3.5. *Every bounded sequence in $\mathbb{R}^n$ contains a convergent subsequence.*

Proof. Let $(a_m)_{m\in\mathbb{N}}$ be a bounded sequence in $\mathbb{R}^n$. Therefore each of the sequences of coordinates $(a_{mj})_{m\in\mathbb{N}}$ is bounded in $\mathbb{R}$. Let $(a_{m_1(k)1})_{k\in\mathbb{N}}$ be a convergent subsequence for the sequence $(a_{m1})_{m\in\mathbb{N}}$ (Bolzano-Weierstraß, theorem 2.4). But then $(a_{m_1(k)2})_{k\in\mathbb{N}}$ is another bounded sequence in $\mathbb{R}$. So let $(a_{m_2(k)2})_{k\in\mathbb{N}}$ be a convergent subsequence. Note that then, $(a_{m_2(k)1})_{k\in\mathbb{N}}$ is still a convergent sequence in $\mathbb{R}$. Again, choose $(a_{m_3(k)3})_{k\in\mathbb{N}}$ to be a convergent subsequence of $(a_{m_2(k)3})_{k\in\mathbb{N}}$, and so on. Eventually we obtain the subsequence $(a_{m_n(k)})_{k\in\mathbb{N}}$, such that all of the sequences of the coordinates converge. Thus $(a_{m_n(k)})_{k\in\mathbb{N}}$ converges in $\mathbb{R}^n$. □

Theorem 3.6 (Heine-Borel: n-dimensional version). *A subset $K \subset \mathbb{R}^n$ is sequentially compact if and only if it is closed and bounded.*

Proof. First assume that K is sequentially compact. The proof that K must be closed and bounded is the same as before.

On the other hand, under the assumption that K is closed and bounded, let $(a_m)_{m\in\mathbb{N}}$ be some sequence in K. According to theorem 3.5, there exists a convergent subsequence $(a_{m(k)})_{k\in\mathbb{N}}$ with $\lim_{k\to\infty} a_{m(k)} = x$, say. If $x \notin K$ then since

K is closed, it follows that $\mathbb{R}^n \setminus K$ is open, and so there exists some $\epsilon > 0$ with $B(x, \epsilon) \subset (\mathbb{R}^n \setminus K)$. Yet $\lim_{k\to\infty} a_{m(k)} = x$. That means that if we look at the coordinates of the points, then we see that $\lim_{k\to\infty} a_{m(k)j} = x_j$, for $j = 1, \ldots, n$. In particular, there must exist an $N \in \mathbb{N}$ such that for all $k \geq N$ we have $|a_{m(k)j} - x_j| < \epsilon/\sqrt{n}$, for all $j = 1, \ldots, n$. But then we have

$$\|a_{m(k)} - x\| = \sqrt{\sum_{j=1}^{n} \left(a_{m(k)j} - x_j\right)^2} < \sqrt{\sum_{j=1}^{n} \left(\frac{\epsilon}{\sqrt{n}}\right)^2} = \sqrt{\epsilon^2} = \epsilon.$$

That would mean that $a_{m(k)} \in (\mathbb{R}^3 \setminus K)$ for all $k \geq N$, a contradiction, since $a_m \in K$ for *all* $m \in \mathbb{N}$. Therefore we must have $x \in K$. □

The usual idea of "compactness" differs somewhat from this idea of "sequential compactness". In 1929, the two mathematicians Pavel Alexandrov and Pavel Urysohn realized that the following definition is more general, and often more useful.

Definition. *Let (M, d) be a metric space, and let $K \subset M$. A collection of open sets $U_i \subset M$, with $i \in I$ some index set, such that $K \subset \bigcup_{i\in I} U_i$, is called an* open covering *of K. The set K is* compact *if for any open covering of K, there exists a finite open sub-covering. That is to say, there exist $i_1, i_2, \ldots, i_n$, for some $n \in \mathbb{N}$, and $i_j \in I$ for all j, such that $K \subset \bigcup_{j=1}^{n} U_{i_j}$.*

Theorem 3.7. *Let (M, d) be a metric space, and let $K \subset M$. Then K is sequentially compact if and only if it is compact.*

Proof. Begin by assuming K is compact. Could it be that K is not sequentially compact? If so, then there exists a sequence $(a_n)_{n\in\mathbb{N}}$ in K which has no cluster points in K. Thus, since each point $x \in K$ is not a cluster point of the sequence, there exists an $\epsilon_x > 0$ (depending on x), such that there are only finitely elements of the sequence in $B(x, \epsilon_x)$. But $\bigcup_{x\in K} B(x, \epsilon_x)$ is an open covering of K, and since K is compact, there exists a finite sub-covering, say

$$K \subset B(x_1, \epsilon_{x_1}) \cup \cdots \cup B(x_m, \epsilon_{x_m}).$$

But then there could only be finitely many elements in the whole sequence, which is a contradiction. Therefore K must be sequentially compact.

Now assume that K is sequentially compact. We must show that it is compact. To start with, for each $n \in \mathbb{N}$ we find a finite collection of points $x_{n1}, x_{n2}, \ldots, x_{n_m}$ in K as follows. Take $x_{n1} \in K$ to be some arbitrary point. Then if $K \not\subset B(x_{n1}, 1/n)$, take $x_{n2} \in K \setminus B(x_{n1}, 1/n)$. And so on, with $x_{nk} \in K \setminus \bigcup_{j=1}^{k-1} B(x_{nj}, 1/n)$. Eventually we must reach some $n_m \in \mathbb{N}$ with $K \subset \bigcup_{j=1}^{n_m} B(x_{nj}, 1/n)$, for otherwise $(x_{nj})_{j\in\mathbb{N}}$ would be a sequence in K containing no convergent subsequence, contradicting the fact that K is sequentially compact.

Therefore, taken together, there are only countably many possible open balls of the form $B(x_{nj}, 1/n)$.

To obtain a contradiction, assume that K is not compact. Assume that there is an open covering $\bigcup_{i\in I} U_i$ of K such that no finite sub-covering exists. For each element $x \in K$, one of the open sets U_i contains x. And for a sufficiently large n, there is an open ball $B(x_{nj}, 1/n)$ containing x, which is also contained within U_i. For each $x \in K$ we can take such a ball, and in this way we obtain a *countable* open covering, say $\bigcup_{k=1}^{\infty} V_k$ of K, such that each V_k is contained within some U_i of the original open covering. Therefore, for all $m \in \mathbb{N}$ we have $K \setminus \bigcup_{k=1}^{m} V_k \neq \emptyset$.

We can construct a sequence $(a_k)_{k\in\mathbb{N}}$ in K by taking

$$a_k \in K \setminus \bigcup_{j=1}^{k} V_j,$$

for each $k \in \mathbb{N}$. Since K is assumed to be sequentially compact, the sequence must have a cluster point, $a \in K$. But since

$$K \subset \bigcup_{k=1}^{\infty} V_k,$$

there must be some $k_a \in \mathbb{N}$ with $a \in V_{k_a}$. Thus V_{k_a} contains infinitely many elements of the sequence, in particular elements of the form a_j for $j > k_a$. This is a contradiction. □

Therefore we see that for metric spaces, the ideas of "sequentially compact", and "compact", are the same. Since Euclidean space $\mathbb{R}^n$ is a metric space, it is usual to state Heine-Borel's theorem by saying that a subset $K \in \mathbb{R}^n$ is compact if and only if it is closed and bounded.

A counterexample

But remember, Heine-Borel's theorem only applies to Euclidean spaces! In general it is not true that closed and bounded implies compact. For example, consider the metric space $C_0([0,1], \mathbb{R})$, consisting of the set of continuous functions $f : [0,1] \to \mathbb{R}$. Let $K \subset C_0([0,1], \mathbb{R})$ be the set of functions whose supremum norm is less than or equal to 1. That is

$$K = \{f \in C_0([0,1], \mathbb{R}) : \sup_{x\in[0,1]} \{|f(x)|\} \leq 1\}.$$

It is an exercise to show that K is both closed and bounded, and yet it is not sequentially compact — therefore also not compact.

Of course, given any metric space (M, d), then the whole set M is also a subset of itself. Therefore the definition of compactness can also apply to the whole space.

Theorem 3.8. *If a metric space (M, d) is compact, then any subset $V \subset M$ is compact if and only if it is closed in M.*

Proof. Assume that V is closed. Then $M \setminus V$ is open. Let $\{U_i : i \in I\}$ be an open covering of V. Then $\{U_i : i \in I\} \cup \{M \setminus V\}$ is an open covering of M. Since M is compact, there exists a finite sub-covering $\{U_{i_1}, \ldots, U_{i_n}, M \setminus V\}$. Then $\{U_{i_1}, \ldots, U_{i_n}\}$ must be a finite sub-covering of V. Thus V is compact.

Now assume that V is compact. We must show that V is closed, that is, that $M \setminus V$ is open. Choose some $x \in (M \setminus V)$. For each $v \in V$, take the open set $B(v, d(v,x)/3)$ around v, and also take the open set $B(x, d(v,x)/3)$ around x. So we have $B(v, d(v,x)/3) \cap B(x, d(v,x)/3) = \emptyset$.[1] But then we have

$$V \subset \bigcup_{v \in V} B(v, d(v,x)/3),$$

so that $\{B(v, d(v,x)/3) : v \in V\}$ is an open covering of V. Since V is compact, there exists a finite sub-covering $\{B(v_1, d(v_1,x)/3), \ldots, B(v_1, d(v_n,x)/3)\}$. Then the intersection

$$\bigcap_{j=1}^{n} B(x, d(v_j,x)/3)$$

is an open set containing x, and we must have

$$V \cap \left(\bigcap_{j=1}^{n} B(x, d(v_j,x)/3) \right) = \emptyset.$$

Thus $M \setminus V$ is open, and so V must be closed. □

3.1.3 Continuous mappings between metric spaces

Definition. *Let (X, d_X) and (Y, d_Y) be metric spaces. A mapping $f : X \to Y$ is* continuous *at the point $x_0 \in X$ if for all $\epsilon > 0$, a $\delta > 0$ exists, such that for all $x \in X$ with $d_X(x, x_0) < \delta$ we have $d_Y(f(x), f(x_0)) < \epsilon$. The mapping is everywhere continuous — that is to say, it is continuous — if it is continuous at all points of X.*

This definition is entirely analogous with the definition we had last semester for real-valued functions $f : U \to \mathbb{R}$, where U is some interval along $\mathbb{R}$. Continuity involved the "distance function", which was given by $d(x,y) = |x - y|$. But as we see, the generalization to metric spaces is a natural one.

When considering mappings $f : X \to Y$, one makes much use of the sets of the form $f^{-1}(V)$, for subsets $V \subset Y$. That is,

$$f^{-1}(V) = \{x \in X : f(x) \in V\}.$$

Thus, for each subset $V \subset Y$, we have that $f^{-1}(V) \subset X$.

Theorem 3.9. *The mapping $f : X \to Y$ is continuous if and only if $f^{-1}(V)$ is open in X, for every open set $V \subset Y$.*

[1] Here we are using the fact that every metric space is a "Hausdorff space". This part of the theorem is not true for non-Hausdorff topological spaces.

Proof. Assume first that $f : X \to Y$ is continuous. Let $V \subset Y$ be open and let $x \in f^{-1}(V)$. Since $f(x) \in V$ and V is open, there exists some $\epsilon > 0$ with $B(f(x), \epsilon) \subset V$. Since f is continuous, there is some $\delta > 0$ with

$$f(B(x, \delta)) \subset B(f(x), \epsilon) \subset V.$$

That is, $B(x, \delta) \subset f^{-1}(V)$. Since x was arbitrarily chosen in $f^{-1}(V)$, it follows that $f^{-1}(V)$ must be an open subset of X.

In the other direction, assume that $f^{-1}(V)$ is open in X, for every open subset V of Y. Let $x \in X$ be some arbitrary point. Let $\epsilon > 0$ be given. Then take the open set $B(f(x), \epsilon) \subset V$. We must have $f^{-1}(B(f(x), \epsilon)) \subset X$ being open in X. Since $x \in f^{-1}(B(f(x), \epsilon))$, there is some open ball $B(x, \delta)$ around x which is contained within $f^{-1}(B(f(x), \epsilon))$. That is, we have shown that for all $\epsilon > 0$ there exists a $\delta > 0$ with $d_Y(f(x'), f(x)) < \epsilon$, for all $x' \in X$ with $d_X(x', x) < \delta$. Therefore f is continuous at x, and since x was arbitrary, it is continuous everywhere. □

It is also interesting to see what happens to compact subsets under continuous mappings.

Theorem 3.10. *Let $f : X \to Y$ be a continuous mapping between metric spaces and let $K \subset X$ be compact. Then $f(K) = \{f(x) : x \in K\}$ is compact in Y.*

Proof. Let $\{U_i : i \in I\}$ be an open covering of $f(K)$ in Y. Then since f is continuous, $\{f^{-1}(U_i) : i \in I\}$ must be an open covering of K in X. Therefore, since K is compact, there must be a finite sub-covering $\{f^{-1}(U_{i_1}), \ldots, f^{-1}(U_{i_n})\}$ of K, and so $\{U_{i_1}, \ldots, U_{i_n}\}$ is a finite sub-covering of $f(K)$ in V. □

The proof of the following theorem follows the proofs of the analogous theorems which we have seen in the last semester.

Theorem 3.11. *Let (M, d) be a metric space and let $K \subset M$ be compact. If $f : M \to \mathbb{R}$ is continuous, then $f(K)$ is compact (that is, closed and bounded) in $\mathbb{R}$. Furthermore, there exist points $x_1 \in K$ with $f(x_1) = \inf\{f(x) : x \in K\}$ and $x_2 \in K$ with $f(x_2) = \sup\{f(x) : x \in K\}$. Also f is uniformly continuous on K.*

Theorem 3.12 (The fundamental theorem of algebra)**.** *Let*

$$f(z) = a_0 + a_1 z + \cdots a_n z^n$$

be a polynomial with $a_j \in \mathbb{C}$ for all $j = 0, \ldots, n$, $n \geq 1$ and $a_n \neq 0$. Then there exists some $z_0 \in \mathbb{C}$ with $f(z_0) = 0$.

Proof. For $z \neq 0$ we have

$$\begin{aligned} |f(z)| &= |a_n||z|^n \left|1 + \frac{a_{n-1}}{a_n z} + \cdots + \frac{a_0}{a_n z^n}\right| \\ &\geq |a_n||z|^n \left(1 - \left|\frac{a_{n-1}}{a_n z}\right| - \cdots - \left|\frac{a_0}{a_n z^n}\right|\right). \end{aligned}$$

Then for $z \in \mathbb{C}$ with

$$|z| > R = \max\{1, 2n|a_j/a_n| : 0 \leq j < n\},$$

we must have

$$|f(z)| > \frac{|a_n||z|^n}{2} > 0.$$

Furthermore, it is clear that for any $C > 0$, we can choose an $R_0 > 0$ such that for all $z \in \mathbb{C}$ with $|z| > R_0$, we have $|f(z)| > C$. In particular, choose C to be larger than $|f(z)|$, for some $z \in \mathbb{C}$. Therefore, if there exists any solution z_0 with $f(z_0) = 0$, then it must be contained within the closed disc

$$D = \{z \in \mathbb{C} : |z| \leq R_0\}.$$

Furthermore, for all $z \notin D$, we know that $|f(z)|$ is not minimal.

According to the theorem of Heine-Borel, D must be compact. Also the function $\phi : D \to \mathbb{R}$ with $\phi(z) = |f(z)|$ is continuous. Thus there exists some $z_0 \in D$ with

$$\phi(z_0) = \inf\{\phi(z) : z \in D\} = \inf\{|f(z)| : z \in \mathbb{C}\}.$$

We must show that $\phi(z_0) = 0$.

To obtain a contradiction, assume that $|f(z_0)| > 0$. We can write

$$f(z) = a_0 + a_1 z + \cdots a_n z^n = c_0 + c_1(z - z_0) + \cdots + c_n(z - z_0)^n$$

for some particular complex numbers $c_0, \ldots, c_n \in \mathbb{C}$. Since $|f(z_0)| > 0$, we must have $|f(z_0)| = |c_0| \neq 0$. Another way to look at this is to take the new function f_1, with $f_1(z) = f(z + z_0)$, so that

$$f_1(z) = c_0 + c_m z^m + z^{m+1} g(z),$$

where $m \geq 1$ is the smallest number such that $c_m \neq 0$, and g is a further polynomial in $\mathbb{C}$.

Let $z_1 \in \mathbb{C}$ be such that $z_1^m = -c_0/c_m$, and for $0 \leq \lambda \leq 1$, consider

$$\begin{aligned} f_1(\lambda z_1) &= c_0 - \lambda^m c_0 + \lambda^{m+1} z_1^{m+1} g(\lambda z_1) \\ &= c_0 \left(1 - \lambda^m + \lambda^{m+1} z_1^{m+1} c_0^{-1} g(\lambda z_1)\right). \end{aligned}$$

Since the interval $[0, 1]$ is compact, there exists some $L > 0$ with

$$|z_1^{m+1} c_0^{-1} g(\lambda z_1)| \leq L,$$

for all $0 \leq \lambda \leq 1$. Therefore

$$|f_1(\lambda z_1)| \leq |c_0| \left(1 - \lambda^m + L\lambda^{m+1}\right).$$

But if we choose $\lambda < 1/L$, then we have $L\lambda^{m+1} < \lambda^m$, so that

$$-\lambda^m + L\lambda^{m+1} < 0.$$

Therefore, for such λ we have

$$|f_1(\lambda z_1)| = |f(\lambda z_1 + z_0)| < |c_0| = \inf\{|f(z)| : z \in \mathbb{C}\}.$$

This contradiction proves the theorem. □

3.1.4 Topological spaces

If the idea of metric spaces is a generalization of the "usual" geometry of the real numbers $\mathbb{R}$, or the Euclidean spaces $\mathbb{R}^n$, then a further generalization is to consider topological spaces. In fact, the study of topology is one of the major branches of pure mathematics.

Definition. *Let X be any non-empty set. A set $\mathcal{O}$ of subsets of X is a* topology *on X if*

1. *Both $\emptyset \in \mathcal{O}$ and also $X \in \mathcal{O}$.*
2. *Every finite intersection of elements of $\mathcal{O}$ is also an element of $\mathcal{O}$.*
3. *Every union of elements of $\mathcal{O}$ is also an element of $\mathcal{O}$.*

Given a *topological space* $(X, \mathcal{O})$, that is to say a non-empty set, together with an appropriate set of subsets, then the elements of $\mathcal{O}$ are called the *open sets* of X. Furthermore, a set $V \subset X$ is a *closed set* if $X \setminus V$ is open.

It is obvious that many of the ideas we have developed for metric spaces can be generalized into the framework of topological spaces. For example theorem 3.9 shows how the idea of continuous mappings between topological spaces should be defined. Also the definition of compact sets can be directly generalized into the theory of topology. Students who wish to pursue such ideas may enjoy taking part in the topology lectures which are offered each year in the Faculty.

3.2 Convolutions

Definition. *Let (M, d) be a metric space and let $f : M \to \mathbb{R}$ be a real-valued function. The* support *$Supp(f)$ of f is the closure of the subset $\{x \in X : f(x) \neq 0\}$. If the support is compact, then the function f is said to have* compact support.

A function $f : \mathbb{R} \to \mathbb{R}$ is called "piecewise continuous" if it is continuous at all points of $\mathbb{R}$ except possibly for some finite set of points $\{p_1, \ldots, p_n\} \subset \mathbb{R}$ where it might be discontinuous.

Definition. *Let $f : \mathbb{R} \to \mathbb{R}$ be piecewise continuous. Then given a Riemann integrable function $g : \mathbb{R} \to \mathbb{R}$ of compact support, the* convolution *$g * f : \mathbb{R} \to \mathbb{R}$ is the function given by*

$$g * f(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt.$$

3.2.1 Dirac sequences

We will say that a sequence of functions of compact support $K_n : \mathbb{R} \to \mathbb{R}$, with $n \in \mathbb{N}$, is a *Dirac sequence* if for all n:

1. $K_n(x) \geq 0$, for all $x \in \mathbb{R}$,
2. K_n is Riemann integrable and we have $\int_{-\infty}^{\infty} K_n(x)dt = 1$, and
3. for each $\delta > 0$, there exists an $N \in \mathbb{N}$ such that $Supp(K_n) \subset [-\delta, \delta]$, for all $n \geq N$.

Theorem 3.13. *Let $f : \mathbb{R} \to \mathbb{R}$ be a piecewise continuous, bounded function, and let $(K_n)_{n\in\mathbb{N}}$ be a Dirac sequence. For each n we define $f_n = K_n * f$. Then for each closed interval $S = [a, b] \subset \mathbb{R}$, such that f is continuous in an open neighborhood of S, we have that the sequence $(f_n)_{n\in\mathbb{N}}$ converges uniformly to f on S.*

Proof. The substitution rule gives

$$f_n(x) = \int_{-\infty}^{\infty} f(x - t)K_n(t)dt,$$

and since $\int_{-\infty}^{\infty} K_n(x)dt = 1$, we have

$$f(x) = f(x) \int_{-\infty}^{\infty} K_n(t)dt = \int_{-\infty}^{\infty} f(x)K_n(t)dt.$$

Therefore, for each $x \in \mathbb{R}$ we have

$$f_n(x) - f(x) = \int_{-\infty}^{\infty} (f(x - t) - f(x)) K_n(t)dt.$$

Let $S = [a, b] \subset \mathbb{R}$ be a closed interval such that f is continuous in an open neighborhood of S. In particular, there exist $a' < a$ and $b' > b$ such that f is continuous in the closed interval $[a', b']$. Then since f is uniformly continuous on $[a', b']$, we have that for all $\epsilon > 0$, some $\delta > 0$ exists with $\delta < \min\{a - a', b' - b\}$, such that for all t with $|t| < \delta$ we have

$$|f(x - t) - f(x)| < \epsilon,$$

for all $x \in S$. Now choose $N \in \mathbb{N}$ so large that $Supp(K_n) \subset (-\delta, \delta)$. Then we have

$$\begin{aligned} |f_n(x) - f(x)| &\leq \int_{-\infty}^{\infty} |f(x - t) - f(x)|K_n(t)dt \\ &= \int_{-\delta}^{\delta} |f(x - t) - f(x)|K_n(t)dt \\ &< \epsilon \cdot \left(\int_{-\delta}^{\delta} K_n(t)dt \right) \\ &= \epsilon. \end{aligned}$$

□

Remark. *It is an exercise (using the Intermediate Value Theorem for Integrals, theorem 2.42) to show that the idea of Dirac sequences can be generalized in the following way. Rather than assuming that the functions in the sequence K_n are of compact support, we assume instead that they satisfy the condition that for all $\epsilon > 0$ and all $\delta > 0$ there exists an $N \in \mathbb{N}$ such that for all $n \geq N$, we have*

$$\int_{-\infty}^{-\delta} K_n(x)dx + \int_{\delta}^{\infty} K_n(x)dx < \epsilon.$$

Therefore we have:

Corollary. *With this alternative formulation of Dirac sequences, theorem 3.13 is also true.*

3.2.2 Weierstrass' convergence theorem

Theorem 3.14. *Let $f : [0,1] \to \mathbb{R}$ be continuous, with $f(0) = f(1) = 0$. Then there exists a sequence of polynomials $(P_n)_{n\in\mathbb{N}}$ which converges to f uniformly on $[0,1]$.*

Proof. We can extend f to a function $f : \mathbb{R} \to \mathbb{R}$ by simply taking $f(x) = 0$, for $x \notin [0,1]$. Consider the sequence of functions $(K_n)_{n\in\mathbb{N}}$ with $K_n : \mathbb{R} \to \mathbb{R}$ for all $n \in \mathbb{N}$, such that

$$K_n(t) = \begin{cases} \frac{(1-t^2)^n}{c_n}, & |t| \leq 1, \\ 0, & |t| > 1, \end{cases}$$

where

$$c_n = \int_{-1}^{1} (1-t^2)^n dt.$$

Then, noting that $K_n(-t) = K_n(t)$, for all $t \in \mathbb{R}$, we have

1. $K_n(t) \geq 0$, for all t.
2. $\int_{-\infty}^{\infty} K_n(t)dt = 1$, since the constant c_n was chosen to ensure that this is true. Note further that
$$\frac{c_n}{2} = \int_0^1 (1-t^2)^n dt = \int_0^1 (1+t)^n(1-t)^n dt \geq \int_0^1 (1-t)^n dt = \int_0^1 t^n dt = \frac{1}{n+1}.$$
3. For $\delta > 0$ with $\delta < 1$ we have
$$\begin{aligned} \int_{-\infty}^{-\delta} K_n(t)dt + \int_{\delta}^{\infty} K_n(t)dt &= 2\int_{\delta}^{1} K_n(t)dt \\ &= 2\int_{\delta}^{1} \frac{(1-t^2)^n}{c_n} dt \\ &\leq 2\int_{\delta}^{1} \frac{(n+1)}{2}(1-\delta^2)^n dt \\ &= (n+1)(1-\delta^2)^n(1-\delta). \end{aligned}$$

But[2] since $0 < (1-\delta^2) < 1$, we have $(n+1)(1-\delta^2)^n \underset{n\to\infty}{\longrightarrow} 0$.

Therefore, $(K_n)_{n\in\mathbb{N}}$ is a Dirac sequence. But according to the corollary to theorem 3.13, we must have the sequence $(f_n)_{n\in\mathbb{N}}$ with

$$f_n(x) = \int_{-\infty}^{\infty} f(t)K_n(x-t)dt$$

converging uniformly to f on $[0,1]$. Since $f(t) = 0$ for $t \notin [0,1]$, we have

$$f_n(x) = \int_0^1 f(t)K_n(x-t)dt.$$

Since K_n is a polynomial of degree $2n$, we can write

$$K_n(x-t) = g_0(t) + g_1(t)x + \cdots + g_{2n}(t)x^{2n},$$

where the $g_j(t)$ are some polynomials in t. Thus if we write

$$a_j = \int_0^1 f(t)g_j(t)dt$$

for each $j = 0, \ldots, 2n$, we obtain

$$f_n(x) = a_0 + a_1x + \cdots + a_{2n}x^{2n}.$$

□

Theorem 3.15 (Weierstrass' convergence theorem). *Let $f : [a,b] \to \mathbb{R}$ be piecewise continuous. Then there exists a sequence of polynomials which converges uniformly to f on compact intervals which contain no points of discontinuity of f.*

Proof. Let $[a,b]$ be an interval where f is continuous. Instead of the function f, consider the function

$$F(x) = f((b-a)x + a) - f(a) - x(f(b) - f(a)).$$

This new function fulfills the requirements of theorem 3.14, so there exits a sequence of polynomials $(P_n)_{n\in\mathbb{N}}$ which converges uniformly to F on $[0,1]$. Then the sequence of polynomials $(Q_n)_{n\in\mathbb{N}}$ with

$$Q_n(x) = P_n\left(\frac{x-a}{b-a}\right) + f(a) + \frac{x-a}{b-a}(f(b) - f(a))$$

converges uniformly to f. □

[2] For n sufficiently large we have $\frac{n+1}{n+2} > (1-\delta^2)$, or $\frac{n+1}{(n+1)+1}(1-\delta^2)^n > (1-\delta^2)^{n+1}$. Thus $(n+1)(1-\delta^2)^n > ((n+1)+1)(1-\delta^2)^{n+1}$, and it follows that the sequence is monotonically decreasing. Since

$$\frac{((n+1)+1)(1-\delta^2)^{n+1}}{(n+1)(1-\delta^2)^n} = \frac{n+2}{n+1}(1-\delta^2) \underset{n\to\infty}{\longrightarrow} (1-\delta^2) < 1,$$

The sequence must converge to zero.

3.3 Periodic functions

3.3.1 Fourier polynomials

A function $f : \mathbb{R} \to \mathbb{R}$ with the property that there exists some constant $L > 0$ such that $f(x+L) = f(x)$ for all $x \in \mathbb{R}$ is called *periodic*, with period L. The most obvious examples are the trigonometric functions: sine and cosine. According to theorem 2.38, both of these functions are periodic, with period 2π.

Of course there are many other possibilities. For example the "sawtooth" function $f(x) = x - [x]$, where $[x] \in \mathbb{Z}$ is the largest whole number[3] which is not larger than x, is periodic, with period 1.

In the theory of Fourier series we seek to represent periodic functions as sums of the trigonometric functions. Therefore, given a periodic function with period L, we first need to alter it so that its period becomes 2π. If, namely $f : \mathbb{R} \to \mathbb{R}$ has period L, then the new function $F : \mathbb{R} \to \mathbb{R}$ given by

$$F(x) = f\left(\frac{L}{2\pi}x\right)$$

has period 2π. Alternatively, we could change the period of the trigonometric functions to L by taking instead

$$\sin\left(\frac{2\pi}{L}x\right) \quad \text{and} \quad \cos\left(\frac{2\pi}{L}x\right).$$

Thus, for the sake of simplicity, and without loosing generality, we will only consider periodic functions with period 2π.

Definition. *A* Fourier polynomial *of order n is a function of the form*

$$\phi(x) = \sum_{k=0}^{n} (a_k \sin(kx) + b_k \cos(kx)).$$

Just as is the case with the "usual" polynomials, it is also true that both sums and products of Fourier polynomials are again Fourier polynomials.[4]

[3] This function is called the "floor function" in English; it is called the "Gauss Klammer" in German.

[4] Recall from last semester that we have the following formulas. For $k > 1$:

$$\begin{aligned} \sin(kx) = \sin(x + (k-1)x) &= \cos(x)\sin((k-1)x) + \sin(x)\cos((k-1)x) \\ \cos(kx) = \cos(x + (k-1)x) &= \cos(x)\cos((k-1)x) - \sin(x)\sin((k-1)x) \end{aligned}$$

Also

$$\begin{aligned} \sin(kx)\sin(lx) &= \frac{1}{2}(\cos((k-l)x) - \cos((k+l)x)) \\ \cos(kx)\cos(lx) &= \frac{1}{2}(\cos((k-l)x) + \cos((k+l)x)) \\ \sin(kx)\cos(lx) &= \frac{1}{2}(\sin((k-l)x) + \sin((k+l)x)) \end{aligned}$$

A particular class of Fourier polynomials can be written

$$D_m(x) = \frac{1}{2\pi}\left(1 + \sum_{k=1}^{m} 2\cos(kx)\right) = \frac{1}{2\pi}\sum_{k=-m}^{m} e^{ikx} = \frac{1}{2\pi}\left(-1 + 2Re\sum_{k=0}^{m} e^{ikx}\right).$$

When $x \neq 2\pi l$, $l \in \mathbb{Z}$, we have

$$\sum_{k=0}^{m} e^{ikx} = \frac{1 - e^{i(m+1)x}}{1 - e^{ix}} = \frac{e^{-ix/2} - e^{i(m+1/2)x}}{e^{-ix/2} - e^{ix/2}} = \frac{e^{-ix/2} - e^{i(m+1/2)x}}{-2i\sin(x/2)}.$$

Therefore

$$-1 + 2Re\sum_{k=0}^{m} e^{ikx} = -1 + 2\left(\frac{\sin(-x/2) - \sin((m+1/2)x)}{-2\sin(x/2)}\right) = \frac{\sin((m+1/2)x)}{\sin(x/2)}.$$

We also have the relation

$$\sum_{m=0}^{n-1} e^{imx} = \frac{1 - e^{inx}}{1 - e^{ix}} = \frac{1 - \cos(nx) - i\sin(nx)}{e^{ix/2}(e^{-ix/2} - e^{ix/2})}.$$

Multiplying both sides with $e^{ix/2}$, we obtain

$$\sum_{m=0}^{n-1} e^{i(m+1/2)x} = \frac{1 - \cos(nx) - i\sin(nx)}{e^{-ix/2} - e^{ix/2}} = \frac{1 - \cos(nx) - i\sin(nx)}{-2i\sin(x/2)}.$$

Comparing the imaginary parts, we must have

$$\sum_{m=0}^{n-1} \sin((m+1/2)x) = \frac{1 - \cos(nx)}{2\sin(x/2)} = \frac{\sin^2(nx/2)}{\sin(x/2)}.$$

For the last equation here, we have used the formula which we found last semester, namely

$$\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b).$$

Therefore

$$\begin{aligned} 1 - \cos(nx) &= 1 - (\cos^2(nx/2) - \sin^2(nx/2)) \\ &= (\cos^2(nx/2) + \sin^2(nx/2)) - (\cos^2(nx/2) - \sin^2(nx/2)) \\ &= 2\sin^2(nx/2). \end{aligned}$$

Next, we consider the Fourier polynomial

$$K_n(x) = \frac{1}{n}\sum_{m=0}^{n-1} D_m(x) = \frac{1}{2\pi n}\sum_{m=0}^{n-1} \frac{\sin((m+1/2)x)}{\sin(x/2)} = \frac{1}{2\pi n}\frac{\sin^2(nx/2)}{\sin^2(x/2)}.$$

For the sequence $(K_n)_{n\in\mathbb{N}}$ we have

1. $K_n(x) \geq 0$ for all x with $|x| < \pi$. (For $x = 0$ we have $D_m(0) = (1/2\pi)(1 + \sum_{k=1}^{m} 2.)$

2.
$$\int_{-\pi}^{\pi} K_n(x)dx = 1$$
This is true since each of the terms D_m is $1/2\pi$ plus a sum of functions of the form $\cos(kx)$, where $k \neq 0$. But for such functions, the integral from $-\pi$ to π is zero.

3. For each $0 \leq \delta < \pi$ we have
$$\frac{1}{n}\int_{\delta}^{\pi} \left(\frac{\sin(nt/2)}{\sin(t/2)}\right)^2 dt \leq \frac{1}{n}\int_{\delta}^{\pi} \frac{1}{\sin^2(t/2)}dt.$$
But the last integral[5] gives a constant, independent of n. Therefore, for all $\epsilon > 0$ and $\delta > 0$ there exists an $N \in \mathbb{N}$, such that
$$\int_{-\pi}^{-\delta} K_n(x)dx + \int_{\delta}^{\pi} K_n(x)dx < \epsilon,$$
for all $n \geq N$.

We can now follow the proof of theorem 3.13 of the last section, but confining our function f, and the functions in a Dirac sequence, to the interval $[-\pi, \pi]$, rather than $(-\infty, \infty)$. That is to say, we alter K_n outside $[-\pi, \pi]$, so that for all $x \in \mathbb{R}$ with $|x| > \pi$, we take $K_n(x)$ to be zero. Given this, then our sequence $(K_n)_{n\in\mathbb{N}}$ is a Dirac sequence, and we follow the proof of theorem 3.14 to obtain

Theorem 3.16. *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous periodic function with period 2π. Then there exists a sequence of trigonometric polynomials which converges uniformly to f.*

Proof. In the proof of theorem 3.14 we used the fact that $K_n(x-t)$ has a particular form. In the present proof, K_n is a Fourier polynomial, consisting of terms of the form $\cos(kx)$. But then we have
$$\cos(k(x-t)) = \cos(kx-kt) = \cos(kx)\cos(kt) + \sin(kx)\sin(kt).$$
Therefore as in the proof of theorem 3.14, the terms with t can be integrated to obtain the coefficients of the appropriate Fourier polynomial.

Since both f and also K_n are periodic, with period 2π, it follows that the restriction to the interval $[-\pi, \pi]$ can be removed, and so the convergence also holds throughout $\mathbb{R}$. □

Remark. *As with Weierstrass' convergence theorem, the present theorem can be extended to include all piecewise continuous periodic functions with period 2π.*

[5] We have
$$\int_{\delta}^{\pi} \frac{1}{\sin^2(t/2)}dt = \frac{2}{\tan(\delta/2)}.$$

3.3.2 Fourier series

In the theory of Fourier series it is most convenient to think in terms of complex valued functions. Let us say that V is the set of all piecewise continuous functions $f : \mathbb{R} \to \mathbb{C}$ with $f(x) = f(x + 2\pi)$, for all $x \in \mathbb{R}$. Therefore V is a vector space over the complex numbers with the usual addition and scalar multiplication of functions. In addition to this, we have the scalar product:

$$\langle f, g \rangle = \frac{1}{2\pi} \int_0^{2\pi} \overline{f(x)} g(x) dx,$$

for elements $f, g \in V$. The following theorem follows trivially from this definition.

Theorem 3.17. *For f, g, $h \in V$ and $\lambda \in \mathbb{C}$ we have*

- $\langle f + g, h \rangle = \langle f, h \rangle + \langle g, g \rangle$,
- $\langle f, g + h \rangle = \langle f, g \rangle + \langle f, h \rangle$,
- $\langle \lambda f, g \rangle = \overline{\lambda} \langle f, g \rangle$,
- $\langle f, \lambda g \rangle = \lambda \langle f, g \rangle$,
- $\langle f, g \rangle = \overline{\langle g, f \rangle}$.
- $\langle f, f \rangle \geq 0$,
- *if $\|f\|_2$ is defined to be $\|f\|_2 = \sqrt{\langle f, f \rangle}$ then $\|\lambda f\|_2 = |\lambda| \|f\|_2$,*
- $\|f + g\|_2 \leq \|f\|_2 + \|g\|_2$. *(This is theorem 2.62 from last semester.)*

The last two properties suggest that the function $\| \cdot \|_2 : V \to \mathbb{R}$ might be a norm. But it isn't. The problem is that we might have $f \neq 0$, but nevertheless, $\|f\|_2 = 0$. If f were assumed to be continuous then $\|f\|_2$ could only be zero if f was the zero function. However since we only assumed that f was piecewise continuous, it might be that f happens to be non-zero at some finite number of points, but zero everywhere else. If this were the case then we would still have $\|f\|_2 = 0$. Thus one says that $\| \cdot \|_2$ is a "semi-norm", rather than a norm.

Theorem 3.18. *Assume that f and $g \in V$ with $\langle f, g \rangle = 0$. Then*

$$\|f + g\|_2^2 = \|f\|_2^2 + \|g\|_2^2.$$

Proof.

$$\langle f + g, f + g \rangle = \langle f, f \rangle + \langle g, g \rangle + \underbrace{\langle f, g \rangle + \langle g, f \rangle}_{=0} = \langle f, f \rangle + \langle g, g \rangle.$$

□

For each $k \in \mathbb{Z}$ we define

$$e_k(x) = e^{ikx}.$$

Theorem 3.19.

$$\langle e_k, e_l \rangle = \begin{cases} 1, & k = l, \\ 0, & k \neq l. \end{cases}$$

Proof.

$$\langle e_k, e_k \rangle = \frac{1}{2\pi} \int_0^{2\pi} \overline{e^{ikx}} e^{ikx} dt = \frac{1}{2\pi} \int_0^{2\pi} 1 \, dt = \frac{1}{2\pi} 2\pi = 1.$$

For $k \neq l$

$$\begin{aligned} \langle e_k, e_l \rangle &= \frac{1}{2\pi} \int_0^{2\pi} \overline{e^{ikx}} e^{ilx} dt \\ &= \frac{1}{2\pi} \int_0^{2\pi} e^{i(l-k)x} dt \\ &= \frac{1}{2\pi} \int_0^{2\pi} \cos((l-k)x) dt + \frac{1}{2\pi} i \int_0^{2\pi} \sin((l-k)x) dt = 0. \end{aligned}$$

□

Given any function $f \in V$, then the k-th *Fourier coefficient* (for any $k \in \mathbb{Z}$) of f is defined to be

$$c_k = \langle e_k, f \rangle.$$

Theorem 3.20. *Given $f \in V$, let $F_n = \sum_{k=-n}^{n} c_k e_k$. Then for any $P = \sum_{k=-n}^{n} a_k e_k$, we have*

$$\langle f - F_n, P \rangle = 0.$$

Proof. For any k with $-n \leq k \leq n$, we have

$$\langle f - F_n, e_k \rangle = \langle f, e_k \rangle - \langle F_n, e_k \rangle = \overline{c_k} - \sum_{j=-n}^{n} \overline{c_j} \langle e_j, e_k \rangle = \overline{c_k} - \overline{c_k} = 0.$$

Therefore

$$\langle f - F_n, P \rangle = \langle f - F_n, \sum_{k=-n}^{n} a_k e_k \rangle = \sum_{k=-n}^{n} a_k \langle f - F_n, e_k \rangle = 0.$$

□

Theorem 3.21. *Again take $F_n = \sum_{k=-n}^{n} c_k e_k$ and $P = \sum_{k=-n}^{n} a_k e_k$. Then we have*

$$\|f - F_n\|_2 \leq \|f - P\|_2.$$

Proof. According to theorem 3.20, we have $\langle f - F_n, F_n - P \rangle = 0$. Therefore according to theorem 3.18 we have

$$\|f - P\|_2^2 = \|(f - F_n) + (F_n - P)\|_2^2 = \|f - F_n\|_2^2 + \underbrace{\|F_n - P\|_2^2}_{\geq 0}.$$

□

Therefore theorem 3.16 implies:

Theorem 3.22. $\lim_{n\to\infty} \|f - F_n\|_2 = 0$.

One says that the sequence $(F_n)_{n\in\mathbb{N}}$ converges to f in *quadratic mean.*

Remark. *In theorem 3.16 we were concerned with periodic* real *functions, and we approximated them with Fourier polynomials, which were again real-valued functions consisting of linear combinations of terms of the form* $\sin(kx)$ *and* $\cos(kx)$*, for various* non-negative *integer values of* k*. So let*

$$P(x) = a_0 + \sum_{k=1}^{n} (a_k \cos(kx) + b_k \sin(kx))$$

be such a Fourier polynomial. Then we have

$$P(x) = \sum_{k=-n}^{n} c_k e^{ikx} = \sum_{k=-n}^{n} c_k(\cos(kx) + i\sin(kx)),$$

where $c_0 = a_0$ *and we have* $c_k = \frac{1}{2}(a_k - ib_k)$ *and* $c_{-k} = \frac{1}{2}(a_k + ib_k)$*, for all* $k \geq 1$*.*

It is now an exercise to show that theorem 3.16 is also true when applied to complex-valued periodic functions.

Definition. *Let* $f \in V$*. The* Fourier series *of* f *is*

$$\lim_{n\to\infty} F_n = \sum_{k=-\infty}^{\infty} c_k e_k.$$

Theorem 3.23. *Let* $f \in V$ *and* c_k *be the* k*-th Fourier coefficient of* f *for each* $k \in \mathbb{N}$*. Then*

$$\left(\left\| f - \sum_{k=-n}^{n} c_k e_k \right\|_2 \right)^2 = (\|f\|_2)^2 - \sum_{k=-n}^{n} |c_k|^2.$$

Proof. Let $g = \sum_{k=-n}^{n} c_k e_k$. Then

$$\langle f, g\rangle = \sum_{k=-n}^{n} c_k \langle f, e_k\rangle = \sum_{k=-n}^{n} c_k \overline{c_k} = \sum_{k=-n}^{n} |c_k|^2.$$

Also

$$\langle g, g\rangle = \sum_{k=-n}^{n} c_k \langle g, e_k\rangle = \sum_{k=-n}^{n} c_k \overline{c_k} = \sum_{k=-n}^{n} |c_k|^2.$$

Therefore

$$\begin{aligned} \|f - g\|_2^2 &= \langle f - g, f - g\rangle \\ &= \langle f, f\rangle - \langle f, g\rangle - \langle g, f\rangle + \langle g, g\rangle \\ &= \|f\|_2^2 - \sum_{k=-n}^{n} |c_k|^2 - \sum_{k=-n}^{n} |c_k|^2 + \sum_{k=-n}^{n} |c_k|^2 \\ &= \|f\|_2^2 - \sum_{k=-n}^{n} |c_k|^2. \end{aligned}$$

□

Theorem 3.24 (Bessel's inequality). *For $f \in V$ with Fourier coefficients c_k we have the inequality*

$$\sum_{k=-\infty}^{\infty} |c_k|^2 \leq \|f\|_2^2 = \frac{1}{2\pi} \int_0^{2\pi} |f(x)|^2 dx.$$

Proof. This is a direct consequence of theorem 3.23, since we must have

$$\left(\left\| f - \sum_{k=-n}^{n} c_k e_k \right\|_2 \right)^2 \geq 0.$$

□

Theorem 3.25. *Let f be a continuous periodic function with period 2π which is also piecewise continuously differentiable with f' being bounded. Then the Fourier series of f converges uniformly to f. Therefore*

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e_k(x),$$

for all $x \in \mathbb{R}$.

Proof. With Bessel's inequality, we have[6]

$$\sum_{k=-\infty}^{\infty} |c_k| < \infty.$$

Since $|e_k(x)| = |\exp(ikx)| = 1$, it follows that the Fourier series is absolutely convergent for each x. Let $\lim_{n\to\infty} F_n(x) = g(x)$, thus defining a function

$$g : \mathbb{R} \to \mathbb{C}.$$

For all $x \in \mathbb{R}$ and $n \in \mathbb{N}$ we have

$$|g(x) - F_n(x)| = \left| \sum_{|k|=n+1}^{\infty} c_k e_k(x) \right| \leq \sum_{|k|=n+1}^{\infty} |c_k e_k(x)| = \sum_{|k|=m+1}^{\infty} |c_k|.$$

[6]Partial integration gives

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} dx = \frac{i}{2\pi k} f(x) e^{-ikx} \Big|_0^{2\pi} - \frac{i}{2\pi k} \int_0^{2\pi} f'(x) e^{-ikx} dx = \frac{\gamma_k}{k}$$

say, where

$$\gamma_k = \frac{i}{2\pi} \int_0^{2\pi} f'(x) e^{-ikx} dx.$$

Then we have

$$|c_k| = \left| \frac{1}{k} \right| \cdot |\gamma_k| \leq \frac{1}{2} \left(\left| \frac{1}{k} \right|^2 + |\gamma_k|^2 \right).$$

But both $\sum_{k=1}^{\infty} 1/k^2$ and $\sum_{k=-\infty}^{\infty} |\gamma_k|^2$ converge.

Thus the sequence of continuous functions $(F_n)_{n\in\mathbb{N}}$ is uniformly convergent. According to theorem 2.47, it follows that the function g must be continuous.

Therefore (remembering theorem 3.22), we have both $\lim_{n\to\infty}\|f-F_n\|_2=0$ and $\lim_{n\to\infty}\|F_n-g\|_2=0$. In other words, for all $\epsilon>0$ there exists an $N\in\mathbb{N}$ with both $\|f-F_n\|_2<\epsilon/2$ and $\|F_n-g\|_2<\epsilon/2$. Using Minkowski's inequality, we have

$$\|f-g\|_2=\|(f-F_n)+(F_n-g)\|_2\leq\|f-F_n\|_2+\|F_n-g\|_2<\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon.$$

Since this is true for all $\epsilon>0$ we must have $\|f-g\|_2=0$. That is,

$$\frac{1}{2\pi}\int_0^{2\pi}|f(x)-g(x)|^2dx=0.$$

This can only be true if $f=g$. Therefore

$$F_n\underset{n\to\infty}{\longrightarrow}f.$$

□

3.3.3 $\zeta(2)=\pi^2/6$

The Riemann "zeta" function is defined to be

$$\zeta(z)=\sum_{n=1}^{\infty}n^{-z},$$

for $z=x+iy\in\mathbb{C}$ with $x>1$. In section 2.23.1 of these notes (from last semester), we saw that this series converges for all such z. It diverges for all z with real part less than or equal to 1. Yet within the theory of complex analysis, it may be extended to the whole complex plane (except for the isolated singularity at $z=1$). The most famous unsolved problem in present-day mathematics is the Riemann Hypothesis. That is that all the non-trivial zeros of the zeta function are confined to the line $z=x+iy$, with $x=1/2$. Anybody who is able to prove the Riemann Hypothesis will achieve immortal fame!

A far simpler question is that of obtaining the values of $\zeta(n)$, for various integers greater than 1. In particular we can use the theory of Fourier series to calculate the value of $\zeta(2)$. For this, we take the function $f:\mathbb{R}\to\mathbb{R}$ with $f(x)=x^2$, for $|x|\leq\pi$, and we specify that $f(x+2\pi)=f(x)$, for all $x\in\mathbb{R}$. Thus f is a periodic, continuous function with period 2π. Therefore we must have

$$f(x)=\sum_{k=-\infty}^{\infty}c_ke^{ikx}=\sum_{k=-\infty}^{\infty}c_k(\cos(kx)+i\sin(kx))=c_0+2\sum_{k=1}^{\infty}c_k\cos(kx).$$

Here we use the fact that f is symmetric ($f(x)=f(-x)$ for all x). But

$$c_0=\frac{1}{2\pi}\int_{-\pi}^{\pi}t^2dt=\frac{\pi^2}{3}.$$

It is an exercise to show that for $k \geq 1$ we have

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} t^2 \cos(kt) dt = (-1)^k \frac{2}{k^2}.$$

Thus

$$x^2 = \frac{\pi^2}{3} + 2 \sum_{k=1}^{\infty} (-1)^k \frac{2}{k^2} \cos(kx).$$

Putting $x = \pi$ into this equation and noting that $\cos(k\pi) = (-1)^k$, we obtain

$$\pi^2 = \frac{\pi^2}{3} + 4 \sum_{k=1}^{\infty} \frac{1}{k^2}.$$

Therefore

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}.$$

3.4 Partial derivatives

Let $G \subset \mathbb{R}^n$ be some open set, and let the function $f : G \to \mathbb{R}$ be given. Then if we take some arbitrary element $\mathbf{x} \in G$, we can write $\mathbf{x} = (x_1, \ldots, x_n)$. Take some $j \in \{1, \ldots, n\}$ and consider the elements $(x_1, \ldots, x_j + h, \ldots, x_n)$, for various values of $h \in \mathbb{R}$. Since G is open, there must exist some $\delta > 0$, such that for all h with $|h| < \delta$, we have $(x_1, \ldots, x_j + h, \ldots, x_n) \in G$. Or we can use the notation of linear algebra: let $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ be the canonical basis for $\mathbb{R}^n$, so that $(x_1, \ldots, x_j + h, \ldots, x_n) = \mathbf{x} + h\mathbf{e}_j$. Then if

$$\lim_{\substack{h \to 0 \\ h \neq 0}} \frac{f(\mathbf{x} + h\mathbf{e}_j) - f(\mathbf{x})}{h}$$

exists, it is called the *partial derivative* of f with respect to x_j, and it is written $\partial_j f(\mathbf{x})$, or $D_j f(\mathbf{x})$. Sometimes it is also written as if it were a fraction, namely

$$\frac{\partial f(\mathbf{x})}{\partial x_j}.$$

If the partial derivative $\partial_j f(\mathbf{x})$ exists for all $\mathbf{x} \in G$, then we can further think about whether or not the partial derivative in the x_k direction exists, for some $k \in \{1, \ldots, n\}$, when applied to the function $\partial_j f : G \to \mathbb{R}$. If so, then we obtain a new function $\partial_k \partial_j f : G \to \mathbb{R}$. In particular, we write

$$\partial_j^2 f(\mathbf{x})$$

if $k = j$.

One also writes

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_k \partial x_j},$$

or

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_j^2},$$

if $k = j$.

Physicists enjoy using these partial derivatives in order to describe the various laws of classical physics. For this, they have developed a number of traditional words to describe certain special combinations of partial derivatives. For example, if we have the function $f : G \to \mathbb{R}$ such that all the partial derivatives $\partial_j f(\mathbf{x})$ exist at some point $\mathbf{x} \in G$, then the vector

$$\text{grad } f(\mathbf{x}) = (\partial_1 f(\mathbf{x}), \ldots, \partial_n f(\mathbf{x}))$$

is called the "gradient" of f at $\mathbf{x}$. Sometimes people also write "$\nabla f(\mathbf{x})$" for the gradient.

A *vector field* is a mapping $F : G \to \mathbb{R}^n$. Then, since $F(\mathbf{x}) \in \mathbb{R}^n$, for each $\mathbf{x} \in G$, we can write

$$F(\mathbf{x}) = (F_1(\mathbf{x}), \ldots, F_n(\mathbf{x})),$$

so that we obtain n new functions $F_i : G \to \mathbb{R}$, for $i = 1, \ldots, n$. If they all have partial derivatives, then we can take

$$\text{div } F(\mathbf{x}) = \partial_1 F_1(\mathbf{x}) + \cdots + \partial_n F_n(\mathbf{x}).$$

This is called the "divergence" of F at $\mathbf{x}$.

These two things can be combined by observing that if we have a twice differentiable function $f : G \to \mathbb{R}$, then the gradient is a vector field, and the divergence of that is again simply a real function. This is called the "Laplace operator", namely

$$\text{div grad } f(\mathbf{x}) = \partial_1^2 f(\mathbf{x}) + \cdots + \partial_n^2 f(\mathbf{x}).$$

It is often written $\Delta f(\mathbf{x})$, and it plays an important role in "potential theory" of mathematical analysis.

Also, particularly in Maxwell's equations of classical electrodynamics, if we have the special case of a vector field in 3-dimensional Euclidean space $\mathbb{R}^3$, then physicists use another combination of partial derivatives, called the "curl" of the vector field. This is sometimes written "$\nabla \times F$", where $F : G \to \mathbb{R}^3$ is the vector field. But the curl operator is not really a part of mathematics, so I will simply ignore it from now on.[7]

[7] It is interesting to know that much of this, and particularly the curl operator, arises in a very natural and elegant way if we consider analysis based on the system of quaternion numbers. This is a kind of 4-dimensional generalization of the 2-dimensional complex number system which we have already gotten to know. In the quaternion system, the "imaginary" part has 3-dimensions, while the "real" part has just one dimension, as with $\mathbb{C}$. When Hamilton discovered the quaternions in 1843, he believed that he had found the true secret behind all of physics. The world consisted simply of quaternions, with "space" being the imaginary part of the quaternions, and "time" being the real part. It all seemed quite compelling, but unfortunately, physics has now progressed beyond such things, and quaternions play no role in modern physics. However, in order to honor the memory of Sir William Hamilton, today's physicists continuously use something called the "Hamiltonian" in their descriptions of quantum field theory.

3.4.1 Partial derivatives commute if they are continuous

Theorem 3.26. *Let $G \subset \mathbb{R}^n$ be open, and let $f : G \to \mathbb{R}$ be such that all second partial derivatives exist and are continuous. Then for all $\mathbf{x} \in G$, and for all $i, j = 1, \ldots, n$ we have*

$$\partial_i \partial_j f(\mathbf{x}) = \partial_j \partial_i f(\mathbf{x}).$$

Proof. Without loss of generality, we prove the theorem in the case $n = 2$ and $i = 1$, $j = 2$. Let $\mathbf{x} = (x_1, x_2)$. For simplicity, and again without loss of generality, we also just prove the theorem in the special case $\mathbf{x} = \mathbf{0} = (0, 0)$.

Therefore, since G is open, and $\mathbf{x} = \mathbf{0}$ is contained within G, there exists some $\delta > 0$, such that the square

$$H = (-\delta, +\delta) \times (-\delta, +\delta)$$

is contained in G. In particular, for all h with $|h| < \delta$, we have that (h, h) is contained in G.

Let the function $F : (-\delta, +\delta) \to \mathbb{R}$ be defined to be

$$F(h) = (f(h, h) - f(h, 0)) - (f(0, h) - f(0, 0)).$$

We can write this as

$$F(h) = g(h) - g(0),$$

where

$$g(t) = f(t, h) - f(t, 0).$$

Then the mean value theorem (2.34), shows that there must exist some ξ between 0 and h ($h \neq 0$), with

$$\frac{g(h) - g(0)}{h} = g'(\xi) = \partial_1 f(\xi, h) - \partial_1 f(\xi, 0).$$

Using the mean value theorem again on the continuously differentiable function

$$\partial_1 f(\xi, \cdot) : (-\delta, +\delta) \to \mathbb{R},$$

we find some μ between 0 and h with

$$\frac{\partial_1 f(\xi, h) - \partial_1 f(\xi, 0)}{h} = \partial_2 \partial_1 f(\xi, \mu).$$

That is

$$F(h) = g(h) - g(0) = g'(\xi)h = (\partial_1 f(\xi, h) - \partial_1 f(\xi, 0))h = \partial_2 \partial_1 f(\xi, \mu) \cdot h^2,$$

or, noting that $(\xi, \mu) \to (0, 0)$ as $h \to 0$, we see that

$$\lim_{\substack{h \to 0 \\ h \neq 0}} \frac{F(h)}{h^2} = \partial_2 \partial_1 f(0, 0).$$

But we could start the other way around, by observing that

$$F(h) = k(h) - k(0),$$

where

$$k(t) = f(h,t) - f(0,t).$$

Then there exists some $\tilde{\mu}$ between 0 and h, such that

$$\frac{k(h)-k(0)}{h} = k'(\tilde{\mu}) = \partial_2 f(h,\tilde{\mu}) - \partial_2 f(0,\tilde{\mu}).$$

Arguing as before, we obtain a $\tilde{\xi}$ between 0 and h with

$$F(h) = k(h) - k(0) = k'(\tilde{\xi})h = (\partial_2 f(h,\tilde{\mu}) - \partial_2 f(0,\tilde{\mu}))h = \partial_1\partial_2 f(\tilde{\xi},\tilde{\mu}) \cdot h^2.$$

But then, again, we have

$$\lim_{\substack{h\to 0\\ h\neq 0}} \frac{F(h)}{h^2} = \partial_1\partial_2 f(0,0).$$

Since the limit

$$\lim_{\substack{h\to 0\\ h\neq 0}} \frac{F(h)}{h^2}$$

is the same in both cases, we finally obtain

$$\partial_1\partial_2 f(\mathbf{0}) = \partial_2\partial_1 f(\mathbf{0}).$$

□

Corollary. *Given that f has sufficiently many continuously differentiable partial derivatives, then for a given m, and a given permutation $\sigma : \{1,\dots,m\} \to \{1,\dots,m\}$, we have*

$$\partial_{i_1}\partial_{i_2}\cdots\partial_{i_m} f(\mathbf{x}) = \partial_{i_{\sigma(1)}}\partial_{i_{\sigma(2)}}\cdots\partial_{i_{\sigma(m)}} f(\mathbf{x}),$$

for all $\mathbf{x} \in G$.

3.4.2 Total derivatives

Let $G \subset \mathbb{R}^n$ be open, and let $f : G \to \mathbb{R}^m$ be a function. That is to say, for each $\mathbf{x} \in G$, $f(\mathbf{x}) \in \mathbb{R}^m$. Therefore we can write $f(\mathbf{x}) = (f_1(\mathbf{x}),\dots,f_m(\mathbf{x}))$, where $f_i : G \to \mathbb{R}$, for each $i = 1,\dots,m$. It may be that each of these functions has partial derivatives. If so, then we can consider

$$\partial_j f_i(\mathbf{x}) = \lim_{\substack{h\to 0\\ h\neq 0}} \frac{f_i(x_1,\dots,x_j+h,\dots,x_n) - f_i(x_1,\dots,x_j,\dots,x_n)}{h},$$

for each $j = 1, \ldots, n$ and $i = 1, \ldots, m$. This gives us an $m \times n$ matrix, namely

$$\begin{pmatrix} \partial_1 f_1(\mathbf{x}) & \cdots & \partial_n f_1(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \partial_1 f_m(\mathbf{x}) & \cdots & \partial_n f_m(\mathbf{x}) \end{pmatrix},$$

which is called the *Jacobi matrix* for the function f at the point $\mathbf{x} \in U$. If f is *totally differentiable* at $\mathbf{x}$, then we write $Df(\mathbf{x})$ to denote its total derivative, and in fact, the total derivative is the Jacobi matrix. But let's begin with the general definition. First note that since G is an open set, there exists some $\delta > 0$, such that $\mathbf{x} + \xi \in G$, for all $\xi \in \mathbb{R}^n$ with $\|\xi\| < \delta$.

Definition. *Let $f : G \to \mathbb{R}^m$ be a function, and take some point $\mathbf{x} \in G$. Then f is said to be* totally differentiable *at $\mathbf{x}$ if there exists an $m \times n$ matrix A, such that if we take $\delta > 0$ to be sufficiently small that $\mathbf{x} + \xi \in U$, for all $\xi \in \mathbb{R}^n$ with $\|\xi\| < \delta$, then the function $\varphi : B(\mathbf{x}, \delta) \to \mathbb{R}^m$ from the ball around $\mathbf{x}$ with radius δ to $\mathbb{R}^m$ given by*

$$f(\mathbf{x} + \xi) = f(\mathbf{x}) + A\xi + \varphi(\xi)$$

is such that

$$\lim_{\substack{\xi \to 0 \\ \xi \neq 0}} \frac{\varphi(\xi)}{\|\xi\|} = 0.$$

Rather than writing the complicated expression $\lim_{\substack{\xi \to 0 \\ \xi \neq 0}} \frac{\varphi(\xi)}{\|\xi\|} = 0$, it is usual to write

$$\varphi(\xi) = o(\|\xi\|).$$

Remark. *Although this definition may look more complicated than the familiar definition for the derivative of a function in one dimension, in reality it is just the same. For if we have the function $f : (a, b) \to \mathbb{R}$ being differentiable at the point $x \in (a, b)$, with derivative $f'(x)$, then let a new function φ be defined for sufficiently small h to be*

$$\varphi(h) = (f(x + h) - f(x)) - f'(x)h.$$

But we have

$$\lim_{\substack{h \to 0 \\ h \neq 0}} \frac{f(x + h) - f(x)}{h} = f'(x),$$

or, put another way

$$\lim_{\substack{h \to 0 \\ h \neq 0}} \frac{\varphi(h)}{h} = \lim_{h \to 0} \frac{f(x + h) - f(x) - f'(x)h}{h} = 0.$$

That is to say, also here we have that f is differentiable at the point x if there exists some real number $f'(x)$, such that

$$f(x + h) = f(x) + f'(x)h + o(|h|).$$

Theorem 3.27. *Let $G \subset \mathbb{R}^n$ be open, and let $f : G \to \mathbb{R}^m$ be a function. Assume that f is differentiable at the point $\mathbf{x} \in G$, with matrix A. Then f is continuous at x, and furthermore, all partial derivatives $\partial_j f_i(\mathbf{x})$ exist at $\mathbf{x}$, and we have $a_{ij} = \partial_j f_i(\mathbf{x})$.*

Proof. Since $\varphi(\xi) = o(\|\xi\|)$, we have $\lim_{\xi \to \mathbf{0}} \varphi(\xi) = 0$. But also $\lim_{\xi \to \mathbf{0}} A\xi = \mathbf{0}$. The fact that f is continuous at $\mathbf{x}$ then follows, since

$$\lim_{\xi \to \mathbf{0}} f(\mathbf{x} + \xi) = \lim_{\xi \to \mathbf{0}} (f(\mathbf{x}) + A\xi + \varphi(\xi)) = f(\mathbf{x}).$$

Given $\xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} \in \mathbb{R}^n$, and some $i = 1, \ldots, m$, let $\varphi_i(\xi)$ be defined to be

$$\varphi_i(\xi) = f_i(\mathbf{x} + \xi) - f_i(\mathbf{x}) - \sum_{k=1}^{n} a_{ik}\xi_k.$$

In particular, if we take $\xi = h\mathbf{e}_j$, then we have

$$f_i(\mathbf{x} + h\mathbf{e}_j) = f_i(\mathbf{x}) + ha_{ij} + \varphi_i(h\mathbf{e}_j),$$

with $\varphi(\xi) = o(\|\xi\|)$, that is $\varphi_i(h\mathbf{e}_j) = o(|h|)$. Therefore

$$\partial_j f_i(\mathbf{x}) = \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{f_i(\mathbf{x} + h\mathbf{e}_j) - f_i(\mathbf{x})}{h} = \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{ha_{ij} + \varphi_i(h\mathbf{e}_j)}{h} = a_{ij}.$$

□

Theorem 3.28. *Again, $f : G \to \mathbb{R}^n$. This time assume that all partial derivatives $\partial_j f_i$ exist and are continuous in some neighborhood of $\mathbf{x} \in G$. Then f is totally differentiable at $\mathbf{x}$.*

Proof. Let $\delta > 0$ be sufficiently small that the ball around $\mathbf{x}$ with radius δ is contained within G. That is, $B(\mathbf{x}, \delta) \subset G$. Let $\xi = (\xi_1, \ldots, \xi_n) \in B(\mathbf{x}, \delta)$. Thus, $\|\xi\| < \delta$. For each $k = 0, 1, \ldots, n$, let

$$\mathbf{p}_k = \mathbf{x} + \sum_{l=1}^{k} \xi_k \mathbf{e}_k,$$

where $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is the canonical basis for $\mathbb{R}^n$. So $\mathbf{p}_0 = \mathbf{x}$ and $\mathbf{p}_n = \mathbf{x} + \xi$.

According to the intermediate value theorem, for each k, there exists some $\theta_k \in [0, 1]$, such that

$$f_i(\mathbf{p}_k) - f_i(\mathbf{p}_{k-1}) = \partial_k f_i(\mathbf{p}_{k-1} + \theta_k \xi_k \mathbf{e}_k)\xi_k.$$

That is, if $\xi_k \neq 0$, then we can write this in the more familiar form

$$\frac{f_i(\mathbf{p}_{k-1} + \xi_k \mathbf{e}_k) - f_i(\mathbf{p}_{k-1})}{\xi_k} = \partial_k f_i(\mathbf{p}_{k-1} + \theta_k \xi_k \mathbf{e}_k).$$

Therefore, we have

$$\begin{aligned} f_i(\mathbf{x}+\xi) - f_i(\mathbf{x}) &= \sum_{k=1}^{m} (f_i(\mathbf{p}_k) - f_i(\mathbf{p}_{k-1})) \\ &= \sum_{k=1}^{m} \partial_k f_i(\mathbf{p}_{k-1} + \theta_k \xi_k \mathbf{e}_k)\xi_k \\ &= \sum_{k=1}^{m} \partial_k f_i(\mathbf{x})\xi_k + \varphi_i(\xi), \end{aligned}$$

where

$$\varphi_i(\xi) = \sum_{k=1}^{m} (\partial_k f_i(\mathbf{p}_{k-1} + \theta_k \xi_k \mathbf{e}_k) - \partial_k f_i(\mathbf{x}))\xi_k.$$

Then the fact that the function $\partial_k f_i$ is continuous at $\mathbf{x}$ means that we must have $\varphi_i(\xi) = o(\|\xi\|)$ for each i. So finally, if we take $A = Df$ to be the Jacobi matrix of partial derivatives, we obtain the desired expression:

$$f(\mathbf{x}+\xi) = f(\mathbf{x}) + A\xi + \varphi(\xi).$$

That is

$$\begin{pmatrix} f_1(\mathbf{x}+\xi) \\ \vdots \\ f_n(\mathbf{x}+\xi) \end{pmatrix} = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{pmatrix} + A \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} + \begin{pmatrix} \varphi_1(\xi) \\ \vdots \\ \varphi_n(\xi) \end{pmatrix},$$

with

$$\begin{pmatrix} \varphi_1(\xi) \\ \vdots \\ \varphi_n(\xi) \end{pmatrix} = \varphi(\xi) = o(\|\xi\|).$$

□

3.4.3 The chain rule in higher dimensions

Theorem 3.29. *Let $G \subset \mathbb{R}^n$ and $H \subset \mathbb{R}^m$ be open subsets, and let $g : G \to \mathbb{R}^m$ and $f : H \to \mathbb{R}^k$ be functions such that $g(G) \subset H$. Therefore, we can consider the combined function $f \circ g : G \to \mathbb{R}^k$, with $(f \circ g)(\mathbf{x}) = f(g(\mathbf{x}))$ for all $\mathbf{x} \in G$. Now let $\mathbf{x}$ be some point particular point in G, and assume that g is totally differentiable at $\mathbf{x}$, and furthermore, f is totally differentiable at $g(\mathbf{x})$. Thus the differential of g at $\mathbf{x}$ is the $m \times n$ matrix $Dg(\mathbf{x})$, and the differential of f at $g(\mathbf{x})$ is the $k \times m$ matrix $Df(g(\mathbf{x}))$.*

Then $f \circ g$ is totally differentiable at $\mathbf{x}$, and we have that $D(f \circ g)$ is the $k \times n$ matrix

$$Df(g(x)) \cdot Dg(x).$$

Proof. Let $\xi \in \mathbb{R}^n$ be a sufficiently small vector so that

$$g(\mathbf{x}+\xi) = g(\mathbf{x}) + Dg(\mathbf{x})\xi + \varphi(\xi)$$

with $\varphi(\xi) = o(\|\xi\|)$. Then let

$$\zeta = g(\mathbf{x}+\xi) - g(\mathbf{x}) = Dg(\mathbf{x})\xi + \varphi(\xi),$$

so that $f(g(\mathbf{x})+\zeta) = f(g(\mathbf{x})) + Df(g(\mathbf{x}))\zeta + \psi(\zeta)$, with $\psi(\zeta) = o(\|\zeta\|)$.

We obtain

$$\begin{aligned}
(f\circ g)(\mathbf{x}+\xi) &= f(g(\mathbf{x}+\xi)) \\
&= f(g(\mathbf{x})+\zeta) \\
&= f(g(\mathbf{x})) + Df(g(\mathbf{x}))\zeta + \psi(\zeta) \\
&= f(g(\mathbf{x})) + Df(g(\mathbf{x}))(Dg(\mathbf{x})\xi + \varphi(\xi)) + \psi(\zeta) \\
&= f(g(\mathbf{x})) + Df(g(\mathbf{x}))Dg(\mathbf{x})\xi + Df(g(\mathbf{x}))\varphi(\xi) + \psi(\zeta) \\
&= f(g(\mathbf{x})) + Df(g(\mathbf{x}))Dg(\mathbf{x})\xi + \chi(\xi)
\end{aligned}$$

where

$$\begin{aligned}
\chi(\xi) &= Df(g(\mathbf{x}))\varphi(\xi) + \psi(\zeta) \\
&= Df(g(\mathbf{x}))\varphi(\xi) + \psi(Dg(\mathbf{x})\xi + \varphi(\xi)).
\end{aligned}$$

So the problem is to show that $\chi(\xi) = o(\|\xi\|)$.

To begin with, since $Df(g(\mathbf{x}))$ is a matrix, representing a linear mapping, we have that for any vector $\mathbf{v} \in \mathbb{R}^m$ there is a constant L such that

$$\|Df(g(\mathbf{x}))\mathbf{v}\| \leq L\|\mathbf{v}\|.$$

Therefore since

$$\lim_{\substack{\xi\to 0\\ \xi\neq 0}} \frac{\varphi(\xi)}{\|\xi\|} = 0,$$

it follows that

$$\lim_{\substack{\xi\to 0\\ \xi\neq 0}} \frac{Df(g(\mathbf{x}))\varphi(\xi)}{\|\xi\|} = 0.$$

The problem now is to show that

$$\lim_{\substack{\xi\to 0\\ \xi\neq 0}} \frac{\psi(Dg(\mathbf{x})\xi + \varphi(\xi))}{\|\xi\|} = 0.$$

For this, we begin by observing that since $\varphi(\xi) = o(\|\xi\|)$, there must exist a constant K with $\|\varphi(\xi)\| \leq K\|\xi\|$, and also there exists a constant H with $\|Dg(\mathbf{x})\xi\| \leq H\|\xi\|$, for all ξ within a given neighborhood of $\mathbf{0}$. Thus

$$\|\zeta\| = \|Dg(\mathbf{x})\xi + \varphi(\xi)\| \leq H\|\xi\| + K\|\xi\| = (H+K)\|\xi\|.$$

On the other hand, since $\psi(\zeta) = o(\|\zeta\|)$, if we write

$$\psi_1(\zeta) = \frac{\psi(\zeta)}{\|\zeta\|},$$

then $\psi_1(\zeta) \to 0$ as $\zeta \to 0$.

We can also write $\psi(\zeta) = \|\zeta\|\psi_1(\zeta)$, so that $\|\psi(\zeta)\| = \|\zeta\| \cdot \|\psi_1(\zeta)\|$. We have

$$\|\psi(\zeta)\| = \|\psi(Dg(\mathbf{x})\xi + \varphi(\xi))\| \leq (H+K)\|\xi\| \cdot \|\psi_1(Dg(\mathbf{x})\xi + \varphi(\xi))\|,$$

or

$$\frac{\|\psi(Dg(\mathbf{x})\xi + \varphi(\xi))\|}{\|\xi\|} \leq (H+K)\|\psi_1(Dg(\mathbf{x})\xi + \varphi(\xi))\|.$$

Therefore

$$\frac{\|\psi(Dg(\mathbf{x})\xi + \varphi(\xi))\|}{\|\xi\|} \xrightarrow[\xi \to 0]{} 0.$$

□

3.4.4 The directional derivative

This is a simple case of the chain rule. Let $G \subset \mathbb{R}^n$ be an open subset and let $f : G \to \mathbb{R}$ be a continuously differentiable function. Now take any vector $\mathbf{v} \in \mathbb{R}^n$ with $\|\mathbf{v}\| = 1$. So $\mathbf{v}$ points us in some specific *direction* in the space $\mathbb{R}^n$. The *directional derivative* of f in the direction $\mathbf{v}$ at the point $\mathbf{x} \in G$ is then defined to be

$$D_{\mathbf{v}} f(\mathbf{x}) = \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{1}{h}(f(\mathbf{x} + h\mathbf{v}) - f(\mathbf{x})).$$

Theorem 3.30. $D_{\mathbf{v}} f(\mathbf{x}) = \langle \mathrm{grad} f(\mathbf{x}), \mathbf{v} \rangle$. *That is, it is the scalar product of* $\mathbf{v}$ *with* $\mathrm{grad} f(x)$.

Proof. We define the function $g : \mathbb{R} \to \mathbb{R}^n$ to be

$$g(t) = \mathbf{x} + t\mathbf{v}.$$

Then clearly g is totally differentiable everywhere, and in particular we have

$$Dg(0) = \mathbf{v}.$$

Writing it out in coordinates, this is

$$Dg(0) = \begin{pmatrix} g_1'(0) \\ \vdots \\ g_n'(0) \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}.$$

But also we have

$$Df(\mathbf{y}) = \Big(\partial_1 f(\mathbf{y}) \cdots \partial_n f(\mathbf{y}) \Big),$$

a $1 \times n$ matrix, for arbitrary points $\mathbf{y} \in G$. The directional derivative of f at $\mathbf{x}$ is given by the derivative of the real function $f \circ g$ at zero. Therefore we have

$$\begin{aligned} D_{\mathbf{v}} f(\mathbf{x}) &= Df(g(0))Dg(0) \\ &= \Big(\partial_1 f(g(0)) \cdots \partial_n f(g(0)) \Big) \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \\ &= \langle \mathrm{grad} f(\mathbf{x}), \mathbf{v} \rangle. \end{aligned}$$

□

3.5 Taylor's formula in higher dimensions

Taylor's formula in higher dimensions is really nothing more than a simple application of the chain rule. Unfortunately it *looks* unpleasantly complicated owing to the fact that everything must be formulated in terms of the messy notation of linear algebra.

The situation to be described is the following. Let $G \subset \mathbb{R}^n$ be some open set, and let $f : G \to \mathbb{R}$ be a function. To say that f is m-times continuously differentiable means that at all points of G, all partial derivatives $\partial_1^{i(1)} \cdots \partial_n^{i(n)} f(x_1, \ldots, x_n)$ exist and are continuous for each combination of partial derivatives such that $0 \leq i(j)$ for all $j = 1, \ldots, n$ and $i(1) + \cdots + i(n) \leq m$. (Of course if we have $i(j) = 0$ for some j, then that simply means that the j-th partial derivative is *not* taken at all.)

Theorem 3.31. *Let $f : G \to \mathbb{R}$ be m-times continuously differentiable. Let $\mathbf{x} \in G$ be some given point, and let $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n$ be such that*

$$\{\mathbf{x} + t\xi : 0 \leq t \leq 1\} \subset G.$$

A new function $g : [0, 1] \to \mathbb{R}$ is now given by the rule $g(t) = f(\mathbf{x} + t\xi)$. We have that g is m times continuously differentiable, and

$$g^{(m)}(t) = \frac{d^m}{dt^m} g(t) = \sum_{i(1)+\cdots+i(n)=m} \frac{m!}{i(1)! \cdots i(n)!} \partial_1^{i(1)} \cdots \partial_n^{i(n)} f(\mathbf{x} + t\xi) \xi_1^{i(1)} \cdots \xi_n^{i(n)}.$$

Proof. Induction on m. For $m = 1$, the derivative is nothing more than the directional derivative, namely

$$g'(t) = \sum_{j=1}^{n} \partial_j f(\mathbf{x} + t\xi) \xi_j = \sum_{i(1)+\cdots+i(n)=1} \frac{1}{1 \cdots 1} \partial_1^{i(1)} \cdots \partial_n^{i(n)} f(\mathbf{x} + t\xi) \xi_1^{i(1)} \cdots \xi_n^{i(n)}.$$

Now assume $m \geq 1$ and that the theorem is true for m. Then

$$\begin{aligned} g^{(m+1)}(t) &= \sum_{j=1}^{n} \partial_j \left(\sum_{i(1)+\cdots+i(n)=m} \frac{m!}{i(1)! \cdots i(n)!} \partial_1^{i(1)} \cdots \partial_n^{i(n)} f(\mathbf{x} + t\xi) \xi_1^{i(1)} \cdots \xi_n^{i(n)} \right) \xi_j \\ &= \sum_{j=1}^{n} \left(\sum_{i(1)+\cdots+i(n)=m} \frac{m!}{i(1)! \cdots i(n)!} \partial_j \partial_1^{i(1)} \cdots \partial_n^{i(n)} f(\mathbf{x} + t\xi) \xi_j \xi_1^{i(1)} \cdots \xi_n^{i(n)} \right) \\ &= \sum_{i(1)+\cdots+i(n)=m+1} \frac{(m+1)!}{i(1)! \cdots i(n)!} \partial_1^{i(1)} \cdots \partial_n^{i(n)} f(\mathbf{x} + t\xi) \xi_1^{i(1)} \cdots \xi_n^{i(n)}. \end{aligned}$$

Here we have used theorem 3.26, and also a standard theorem of combinatorics, namely the *multinomial* theorem. At the beginning of the last semester, we saw

the *binomial* theorem. That was that we have

$$(a+b)^m = \sum_{k=0}^{m} \binom{m}{k} a^{m-k}b^k = \sum_{k=0}^{m} \frac{m!}{(m-k)!k!} a^{m-k}b^k.$$

The multinomial theorem is the appropriate generalization for the expression

$$(a+b+\cdots+c)^m.$$

Or, in other words,

$$(a_1+\cdots+a_n)^m = \sum_{i(1)+\cdots+i(n)=m} \frac{m!}{i(1)!\cdots i(n)!} a_1^{i(1)} a_2^{i(2)} \cdots a_n^{i(n)}.$$

In the present instance we have

$$(\partial_1+\cdots+\partial_n)^m f(\mathbf{x}+\xi) = \sum_{i(1)+\cdots+i(n)=m} \frac{m!}{i(1)!\cdots i(n)!} \partial_1^{i(1)} \partial_2^{i(2)} \cdots \partial_n^{i(n)} f(\mathbf{x}+\xi).$$

□

But now the Taylor formula (theorem 2.49) gives, in its alternative formulation

Theorem 3.32. *Given the conditions of theorem 3.31, then there exists some* θ *with* $0 \leq \theta \leq 1$*, such that*

$$f(\mathbf{x}+\xi) = g(1) = \sum_{k=0}^{m-1} \frac{g^{(k)}(0)}{k!} \cdot 1^k + \frac{g^{(m)}(\theta)}{m!} \cdot 1^m = \sum_{k=0}^{m-1} \frac{g^{(k)}(0)}{k!} + \frac{g^{(m)}(\theta)}{m!}.$$

Substituting the appropriate expressions from theorem 3.31 for the terms $g^{(k)}(0)$ gives the complicated-looking formulation of Taylor's formula found in most textbooks.

Note that the last term in this formula can be written

$$\frac{g^{(m)}(\theta)}{m!} = \frac{g^{(m)}(0)}{m!} + R(\xi),$$

where

$$\begin{aligned} R(\xi) &= \frac{g^{(m)}(\theta)}{m!} - \frac{g^{(m)}(0)}{m!} \\ &= \sum_{i(1)+\cdots+i(n)=m} \frac{m!}{i(1)!\cdots i(n)!} \partial_1^{i(1)} \cdots \partial_n^{i(n)} \left(f(\mathbf{x}+\theta\xi) - f(\mathbf{x}) \right) \xi_1^{i(1)} \cdots \xi_n^{i(n)} \\ &= o(\|\xi\|^m) \end{aligned}$$

since f is taken to be m-times continuously partially differentiable.

Therefore we have

$$f(\mathbf{x}+\xi) = \sum_{k=0}^{m-1} \frac{g^{(k)}(0)}{k!} + \frac{g^{(m)}(0)}{m!} + R(\xi),$$

where $R(\xi) = o(\|\xi\|^m)$. That is, $\lim_{\xi\to\infty} R(\xi)/\|\xi\| = 0$.

3.5.1 The Hessian Matrix

It is often considered interesting to take the Taylor formula for the case $m = 2$. So let $f : G \to \mathbb{R}$ be twice continuously partially differentiable at the point $\mathbf{x} \in G$, and open subset of $\mathbb{R}^n$. Let $\xi \in \mathbb{R}^n$ be such that $\mathbf{x} + t\xi \in G$, for all $t \in [0, 1]$. Then writing $g(t) = f(\mathbf{x} + t\xi)$, we obtain a function $g : [0, 1] \to \mathbb{R}$ which is continuous, and twice continuously differentiable in $(0, 1)$. According to Taylor's formula, we then have

$$\begin{aligned} f(\mathbf{x}+\xi) &= g(1) \\ &= g(0) + g'(0)(1-0) + \frac{1}{2}g''(\theta)(1-0)^2 \\ &= g(0) + g'(0) + \frac{1}{2}g''(0) + R(\xi) \end{aligned}$$

where $0 < \theta < 1$ and $R(\xi) = o(\|\xi\|^2)$.

But $g'(0)$ is simply the directional derivative at $\mathbf{x}$ in the direction of ξ. Furthermore, according to theorem 3.31, we must have

$$g''(0) = \sum_{i=1}^{n}\sum_{j=1}^{n} \partial_i\partial_j f(\mathbf{x})\xi_i\xi_j.$$

Therefore, we obtain

$$f(\mathbf{x}+\xi) = f(\mathbf{x}) + \langle \mathrm{grad} f(\mathbf{x}), \xi\rangle + \frac{1}{2}\langle \xi, A\xi\rangle + R(\xi),$$

where $R(\xi) = o(\|\xi\|^2)$ and A is the *Hessian matrix*. That is:

Definition. *The $n \times n$ matrix*

$$\begin{pmatrix} \partial_1\partial_1 f(\mathbf{x}) & \cdots & \partial_1\partial_n f(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \partial_n\partial_1 f(\mathbf{x}) & \cdots & \partial_n\partial_n f(\mathbf{x}) \end{pmatrix}$$

is called the Hessian matrix *of f at the point $\mathbf{x}$.*

Since for all i and j we have $\partial_i\partial_j f(\mathbf{x}) = \partial_j\partial_i f(\mathbf{x})$, we see that the Hessian matrix is symmetric. But from linear algebra we know that every real symmetric matrix is similar to a diagonal matrix. Thus there exists an orthonormal basis for $\mathbb{R}^n$, with respect to which the Hessian matrix is diagonal.

What this means is that we can find a new basis for the vector space $\mathbb{R}^n$, such that with respect to this new basis, the Hessian matrix is diagonal

$$A = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}.$$

Expressing the vector ξ as a linear combination of these new basis vectors (and simply writing $\xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix}$ again), we have

$$\frac{1}{2}\langle \xi, A\xi \rangle = \frac{1}{2}\sum_{i=1}^{n} \lambda_i \xi_i^2.$$

If $\lambda_i > 0$ for all $i = 1, \ldots, n$, then we say that the matrix A is *positive definite*. If we only have $\lambda_i \geq 0$ for all $i = 1, \ldots, n$, then A is called *positive semi-definite*. Similarly, A is called *negative definite* if $\lambda_i < 0$ for all i, and *negative semi-definite* if $\lambda_i \leq 0$ for all i. Otherwise, the matrix A is called *indefinite*.

Put another way, if A is positive definite, then $\langle \xi, A\xi \rangle > 0$ for all $\xi \neq 0$. Also if A is negative definite, then $\langle \xi, A\xi \rangle < 0$, for all $\xi \neq 0$. This is true regardless of which basis is chosen for representing vectors in $\mathbb{R}^n$, so we may simply return to the canonical basis.

All of this is of most interest in the case that $\mathrm{grad} f(\mathbf{x}) = 0$. Then we have:

Theorem 3.33. *Let $G \subset \mathbb{R}^n$ be open, and let $f : G \to \mathbb{R}$ be twice continuously differentiable at some point $\mathbf{x} \in G$, such that $\mathrm{grad} f(\mathbf{x}) = 0$. If the Hessian matrix is positive definite, then $\mathbf{x}$ is an isolated local minimum of the function. On the other hand, if the Hessian matrix is negative definite, then $\mathbf{x}$ is an isolated local maximum.*

Remark. *If the Hessian matrix is indefinite, then one says that $\mathbf{x}$ is a* saddle-point *of the function.*

3.6 Implicit Functions

3.6.1 An example

Let $F : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$F(x, y) = x^2 + y^2.$$

The set of points (x, y) satisfying $F(x, y) = 1$ is then obviously the unit circle. On the other hand, we can ask the question: What function g satisfies the relation

$$F(x, g(x)) = 1 \ ?$$

Simply by looking at the circle, we see that the answer is given by taking the function $g : [-1, +1] \to \mathbb{R}$ with either $g(x) = -\sqrt{1 - x^2}$ or $g(x) = +\sqrt{1 - x^2}$. Thus the function g is given *implicitly* by the conditions $F(x, g(x)) = 1$ and $F(x, y) = x^2 + y^2$.

But assuming that we were not able to so easily see what g was, how should we proceed?

Assuming that there is some solution g, let $h(x) = F(x, g(x))$. Since we assume that g is such that $F(x, g(x)) = 1$, it follows that $h(x) = 1$, for all relevant x. Therefore, $h'(x) = 0$, or, using the chain rule, we get

$$h'(x) = \partial_1 F(x, g(x))x' + \partial_2 F(x, g(x))g'(x) = 2x + 2g(x)g'(x) = 0.$$

Or

$$g'(x) = -\frac{x}{g(x)}.$$

Clearly, the functions $g(x) = \pm\sqrt{1 - x^2}$ satisfy this equation.

This little calculation can be generalized to higher dimensional spaces in the following, rather complicated way.

3.6.2 The same method in higher dimensions

This time, let $G_1 \subset \mathbb{R}^k$ and $G_2 \subset \mathbb{R}^m$ be open subsets, and let

$$F : G_1 \times G_2 \to \mathbb{R}^m$$

be a mapping into $\mathbb{R}^m$ such that F is totally differentiable at some point $(\mathbf{a}, \mathbf{b}) \in G_1 \times G_2$. Thus $DF(\mathbf{a}, \mathbf{b})$ is an $m \times (k + m)$ matrix.

It is convenient to consider this matrix as consisting of two parts, namely the first k columns, giving an $m \times k$ matrix, and then the m columns after that, giving an $m \times m$ matrix. Thus for various points $\mathbf{x} = (x_1, \ldots, x_k) \in G_1$ and $\mathbf{y} = (y_1, \ldots, y_m) \in G_2$, the total derivative of F is

$$DF = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_k} & \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_m} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial x_1} & \cdots & \frac{\partial F_m}{\partial x_k} & \frac{\partial F_m}{\partial y_1} & \cdots & \frac{\partial F_m}{\partial y_m} \end{pmatrix} = \begin{pmatrix} \frac{\partial F}{\partial \mathbf{x}} & \frac{\partial F}{\partial \mathbf{y}} \end{pmatrix},$$

where

$$\frac{\partial F}{\partial \mathbf{x}}(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} \frac{\partial F_1(\mathbf{x},\mathbf{y})}{\partial x_1} & \cdots & \frac{\partial F_1(\mathbf{x},\mathbf{y})}{\partial x_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m(\mathbf{x},\mathbf{y})}{\partial x_1} & \cdots & \frac{\partial F_m(\mathbf{x},\mathbf{y})}{\partial x_k} \end{pmatrix}, \quad \frac{\partial F}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial F_1(\mathbf{x},\mathbf{y})}{\partial y_1} & \cdots & \frac{\partial F_1(\mathbf{x},\mathbf{y})}{\partial y_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m(\mathbf{x},\mathbf{y})}{\partial y_1} & \cdots & \frac{\partial F_m(\mathbf{x},\mathbf{y})}{\partial y_m} \end{pmatrix}.$$

Given all this, then we have...

Theorem 3.34. *Assume $g : G_1 \to G_2$ is a mapping, totally differentiable[8] at $\mathbf{a}$, with $g(\mathbf{a}) = \mathbf{b}$, such that $F(\mathbf{x}, g(\mathbf{x})) = 0$, for all $\mathbf{x} \in G_1$. Assume furthermore*

[8]The proof of this theorem is made somewhat more complicated if we only assume that g is continuous at $(\mathbf{a}, \mathbf{b})$, rather than being totally differentiable. However, with this seemingly more general assumption, we can still prove that g is totally differentiable there, so nothing is gained. In particular, our proof of theorem 3.38 will only be sufficient to show that the function g which is obtained is continuous. Interested students are referred to the appropriate part of Forster's *Analysis 2* for the relevant proof.

that $\frac{\partial F}{\partial \mathbf{y}}(\mathbf{a}, \mathbf{b})$ *is a non-singular matrix and that* $F(\mathbf{x}, g(\mathbf{x})) = \mathbf{0}$*, for all* $\mathbf{x} \in G_1$*. Then we have*

$$\frac{\partial g}{\partial \mathbf{x}}(\mathbf{a}) = -\left(\frac{\partial F}{\partial \mathbf{y}}(\mathbf{a}, \mathbf{b})\right)^{-1} \frac{\partial F}{\partial \mathbf{x}}(\mathbf{a}, \mathbf{b}),$$

where

$$Dg(\mathbf{a}) = \frac{\partial g}{\partial \mathbf{x}}(\mathbf{a}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(\mathbf{a}) & \cdots & \frac{\partial g_1}{\partial x_k}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial x_1}(\mathbf{a}) & \cdots & \frac{\partial g_m}{\partial x_k}(\mathbf{a}) \end{pmatrix}.$$

Proof. Despite all these complicated matrices, the situation is really the same as in the more simple case when $k = m = 1$, which we have already seen. We can consider the mapping $h : G_1 \to G_2$ given by

$$h(\mathbf{x}) = F(\mathbf{x}, g(\mathbf{x})) = \mathbf{0}.$$

Since h is the constant mapping to $\mathbf{0}$, we have that $Dh(\mathbf{x})$ exists everywhere, and it is simply the $m \times m$ zero matrix, which we can denote by 0. Therefore, using the chain rule, we have

$$0 = h'(\mathbf{a}) = \frac{\partial F}{\partial \mathbf{x}}(\mathbf{a}, \mathbf{b}) \frac{\partial \mathbf{a}}{\partial \mathbf{x}} + \frac{\partial F}{\partial \mathbf{y}}(\mathbf{a}, \mathbf{b}) \frac{\partial g}{\partial \mathbf{x}}(\mathbf{a}).$$

But $\frac{\partial \mathbf{x}}{\partial \mathbf{x}}$ is the $m \times m$ unit matrix. Therefore we have

$$\frac{\partial F}{\partial \mathbf{x}}(\mathbf{a}, \mathbf{b}) + \frac{\partial F}{\partial \mathbf{y}}(\mathbf{a}, \mathbf{b}) \frac{\partial g}{\partial \mathbf{x}}(\mathbf{a}) = 0.$$

□

3.6.3 Finding an implicitly given function

The technique used to find an implicitly given function involves finding a series of functions which converge to the specific function which we are looking for. The same technique is also used when we prove that certain kinds of differential equations have unique solutions.

The functions are considered to be vectors in a real vector space. From theorem 2.47 we know that if the functions are continuous, then the vector space is complete with respect to the supremum norm.

Theorem 3.35 (Banach's fixed point theorem)**.** *Let* V *be a complete normed vector space,*[9] *and let* $f : V \to V$ *be such that there exists some constant* $0 \leq L < 1$ *with*

$$\|f(u) - f(v)\| \leq L\|u - v\|,$$

for all $u, v \in V$*. Then there exists a unique fixed point* $w \in V$*, with* $f(w) = w$*.*

[9] That is, all Cauchy sequences converge.

Proof. Choose some arbitrary vector $v_0 \in V$. Then recursively define $v_n = f(v_{n-1})$, for all $n \in \mathbb{N}$. Thus the sequence $(v_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, whose limit is some particular vector $w \in V$. It is now an exercise to show that $f(w) = w$.

If w' is some other vector with $f(w') = w'$ then we have

$$\|w' - w\| = \|f(w') - f(w)\| \leq L\|w' - w\|.$$

Since $L < 1$, this can only be true if $\|w' - w\| = 0$, that is, $w' = w$. □

The next idea which we need is a generalization of the mean value theorem (2.34) to higher dimensions. To begin with, recall that the one-dimensional version of the mean value theorem can be formulated in the following way. Let $f : [a, b] \to \mathbb{R}$ be continous and differentiable in (a, b). Then there exists some θ with $0 < \theta < 1$ such that

$$\frac{f(b) - f(a)}{b - a} = \frac{f(a + (b - a)) - f(a)}{b - a} = f'(a + \theta(b - a)).$$

Or, taking $x = a$ and $h = b - a$, we have

$$f(x + h) - f(x) = f'(x + \theta h) \cdot h.$$

On the other hand, using the fundamental theorem of calculus, and the substitution rule for integrals, we have

$$f(x + h) - f(x) = \int_x^{x+h} f'(u)du = \int_0^1 f'(x + th)hdt = \left(\int_0^1 f'(x + th)dt\right) h.$$

The mean value theorem in higher dimensions will be a generalization of this formula.

Theorem 3.36 (Mean value theorem for higher dimensions)**.** *Let $G \subset \mathbb{R}^n$ be open, and let $f : G \to \mathbb{R}^m$ be continuously differentiable. Take some $\mathbf{x} \in G$ and $\xi \in \mathbb{R}^n$ such that $\mathbf{x} + t\xi \in G$, for all t with $t \in [0, 1]$. Then we have*

$$f(\mathbf{x} + \xi) - f(\mathbf{x}) = \left(\int_0^1 Df(\mathbf{x} + t\xi)dt\right) \xi.$$

Proof. To begin with, note that the matrix inside the integral here consists of an $m \times n$ array of real functions, namely the functions $\partial_j f_i(\mathbf{x}+t\xi)$. Taking the integral of this matrix involves integrating each of these individual functions, giving us an $m \times n$ matrix of real numbers, representing a linear mapping $\mathbb{R}^n \to \mathbb{R}^m$.

For each $i \in \{1, \ldots, m\}$, let $g_i(t) = f_i(\mathbf{x} + t\xi)$, for $t \in [0, 1]$. Then for $0 < t < 1$ we have

$$g_i'(t) = \langle \mathrm{grad} f_i(\mathbf{x} + t\xi), \xi\rangle = \sum_{j=1}^n \partial_j f_i(\mathbf{x} + t\xi)\xi_j.$$

Therefore

$$\begin{aligned} f_i(\mathbf{x}+\xi) - f_i(\mathbf{x}) &= g_i(1) - g_i(0) \\ &= \int_0^1 g'(t)dt \\ &= \int_0^1 \left(\sum_{j=1}^n \partial_j f_i(\mathbf{x}+t\xi)\xi_j \right) dt \\ &= \sum_{j=1}^n \left(\int_0^1 \partial_j f_i(\mathbf{x}+t\xi)dt \right) \xi_j, \end{aligned}$$

and $\int_0^1 \partial_j f_i(\mathbf{x}+t\xi)dt$ is the i,j-th element of the matrix $\int_0^1 Df(\mathbf{x}+t\xi)dt$, □

Theorem 3.37. *With the same conditions as in the previous theorem, let*

$$M = \sup_{0\le t\le 1} \{\|Df(\mathbf{x}+t\xi)\cdot \mathbf{y}\| : \mathbf{y}\in\mathbb{R}^n \text{ with } \|y\| = 1\}.$$

Then

$$\|f(\mathbf{x}+\xi) - f(\mathbf{x})\| \le M\|\xi\|.$$

Proof. For each $t\in[0,1]$ let us say that $M(t)$ is the *norm* of the matrix $Df(\mathbf{x}+t\xi)$. That is to say, given some linear mapping

$$\psi : \mathbb{R}^n \to \mathbb{R}^m,$$

the norm of ψ (or of the matrix representing ψ) is defined to be

$$\|\psi\| = \sup_{\substack{\zeta\in\mathbb{R}^n \\ \|\zeta\|=1}} \|\psi(\zeta)\|.$$

Then given any non-zero vector $\mathbf{y}\in\mathbb{R}^n$, we have

$$\|\psi(\mathbf{y})\| = \|\mathbf{y}\| \left\| \psi\left(\frac{\mathbf{y}}{\|\mathbf{y}\|} \right) \right\| \le \|\mathbf{y}\|\|\psi\|.$$

It is a simple exercise in linear algebra to show that for a linear mapping between two finite dimensional normed vector spaces, the norm of the mapping, as defined here, must exist.[10]

[10] For example, if the canonical basis vectors of $\mathbb{R}^n$ are $\mathbf{e}_1,\dots,\mathbf{e}_n$, then we can write

$$\zeta = \zeta_1\mathbf{e}_1 + \cdots + \zeta_n\mathbf{e}_n.$$

Since $\|\zeta\| = 1$ we must have $|\zeta_j| \le 1$ for all $j = 1,\dots,n$. Then

$$\|\psi(\zeta)\| \le |\zeta_1|\|\psi(\mathbf{e}_1)\| + \cdots + |\zeta_n|\|\psi(\mathbf{e}_n)\| \le \|\psi(\mathbf{e}_1)\| + \cdots + \|\psi(\mathbf{e}_n)\|.$$

Since our mapping $f : G \to \mathbb{R}^m$ is assumed to be continuously differentiable, it follows that the function $t \to M(t)$ is continuous on $[0,1]$, thus the supremum M must exist. Therefore we have

$$\begin{aligned} \|f(\mathbf{x}+\xi) - f(\mathbf{x})\| &= \left\| \left(\int_0^1 Df(\mathbf{x}+t\xi)dt \right) \cdot \xi \right\| \\ &= \left\| \int_0^1 Df(\mathbf{x}+t\xi) \cdot \xi dt \right\| \\ &\leq \int_0^1 \|Df(\mathbf{x}+t\xi) \cdot \xi\| dt \\ &\leq \int_0^1 M\|\xi\| dt \\ &= M\|\xi\|. \end{aligned}$$

(Note that the first inequality here follows from the triangle inequality in $\mathbb{R}^m$.) $\square$

We are now able to prove the theorem on implicit functions.

Theorem 3.38. *Let $G_1 \subset \mathbb{R}^k$ and $G_2 \subset \mathbb{R}^m$ be open subsets such that the product $G_1 \times G_2$ contains a particular point $(\mathbf{a}, \mathbf{b}) \in G_1 \times G_2$. Let $F : G_1 \times G_2 \to \mathbb{R}^m$ be a continuously differentiable function such that $F(\mathbf{a}, \mathbf{b}) = \mathbf{0}$, and such that $\frac{\partial F}{\partial \mathbf{y}}(\mathbf{a}, \mathbf{b})$ is an invertible $m \times m$ matrix. (We use the same notation here as in theorem 3.34.) Then there are open neighborhoods $V_1 \subset G_1$ and $V_2 \subset G_2$ with $\mathbf{a} \in V_1$ and $\mathbf{b} \in V_2$, and a continuous mapping $g : V_1 \to V_2$ with $F(\mathbf{x}, g(\mathbf{x})) = \mathbf{0}$ for all $\mathbf{x} \in V_1$. Furthermore, for all points $(\mathbf{x}, \mathbf{y})$ with $F(\mathbf{x}, \mathbf{y}) = \mathbf{0}$), we have $\mathbf{y} = g(\mathbf{x})$.*

Proof. For simplicity, and without loss of generality, we assume that $\mathbf{a} = \mathbf{0} \in \mathbb{R}^k$ and $\mathbf{b} = \mathbf{0} \in \mathbb{R}^m$. And then we will simply denote the matrix $\frac{\partial F}{\partial \mathbf{y}}(\mathbf{0}, \mathbf{0})$ by B.

A new mapping $H : G_1 \times G_2 \to \mathbb{R}^m$ is given by the rule

$$H(\mathbf{x}, \mathbf{y}) = \mathbf{y} - B^{-1}F(\mathbf{x}, \mathbf{y}).$$

Clearly if $F(\mathbf{x}, \mathbf{y}) = \mathbf{0}$, then we must have $H(\mathbf{x}, \mathbf{y}) = \mathbf{y}$. But also if $H(\mathbf{x}, \mathbf{y}) = \mathbf{y}$ then we must have $F(\mathbf{x}, \mathbf{y}) = \mathbf{0}$, since after all, $F(\mathbf{x}, \mathbf{y})$ is simply a vector in $\mathbb{R}^m$, and since B is an invertible matrix, if $F(\mathbf{x}, \mathbf{y})$ were non-zero, then also $B^{-1}F(\mathbf{x}, \mathbf{y})$ would be non-zero. Therefore

$$F(\mathbf{x}, \mathbf{y}) = \mathbf{0} \quad \Leftrightarrow \quad H(\mathbf{x}, \mathbf{y}) = \mathbf{y},$$

and so our goal is now to find a function g with $H(\mathbf{x}, g(\mathbf{x})) = g(\mathbf{x})$.

Because F is continuously differentiable, it follows that the same is true of H, and we have

$$\frac{\partial H}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y}) = 1 - B^{-1}\frac{\partial F}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y}).$$

Here, 1 stands for the $m \times m$ identity matrix. So we have

$$\frac{\partial H}{\partial \mathbf{y}}(\mathbf{0}, \mathbf{0}) = 0,$$

where 0 is the $m \times m$ zero matrix.

Since F was continuously differentiable, it follows that also the functions which are the elements of the matrix $\frac{\partial H}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y})$ are continuous. They are all zero at $(\mathbf{0}, \mathbf{0})$, therefore there exist $\delta_1 > 0$ and $\delta_2 > 0$ such that

$$\left\| \frac{\partial H}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y}) \right\| < \frac{1}{2},$$

for all $\|\mathbf{x}\| < \delta_1$ and $\|\mathbf{y}\| < \delta_2$. Here, $\left\| \frac{\partial H}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y}) \right\|$ is the norm of the matrix.

But also, since H is continuous, we may choose δ_1 sufficiently small that also

$$\|H(\mathbf{x}, \mathbf{0})\| < \frac{\delta_2}{2},$$

for all $\|\mathbf{x}\| < \delta_1$.

Let $V_1 = \{\mathbf{x} \in \mathbb{R}^k : \|\mathbf{x}\| < \delta_1\}$ and $V_2 = \{\mathbf{y} \in \mathbb{R}^m : \|\mathbf{y}\| < \delta_2\}$. The function

$$g : V_1 \to V_2$$

is then constructed by means of an iteration. To begin, let

$$g_0(\mathbf{x}) = \mathbf{0},$$

for all $\mathbf{x} \in V_1$. Then for each $n \in \mathbb{N}$, let

$$g_n(\mathbf{x}) = H(\mathbf{x}, g_{n-1}(\mathbf{x})).$$

We obtain

$$\begin{aligned}
\|g_{n+1}(\mathbf{x}) - g_n(\mathbf{x})\| &= \|H(\mathbf{x}, g_n(\mathbf{x})) - H(\mathbf{x}, g_{n-1}(\mathbf{x}))\| \\
&\leq \frac{1}{2}\|(x, g_n(\mathbf{x})) - (x, g_{n-1}(\mathbf{x}))\| \\
&= \frac{1}{2}\|g_n(\mathbf{x}) - g_{n-1}(\mathbf{x})\| \\
&\leq \frac{1}{2^n}\|g_1(\mathbf{x}) - g_0(\mathbf{x})\| \\
&= \frac{1}{2^n}\|g_1(\mathbf{x})\| \\
&< \frac{1}{2^n}\frac{\delta_2}{2}.
\end{aligned}$$

Note that the first inequality here follows from theorem 3.37, and the further inequalities follow from the fact that

$$\|g_1(\mathbf{x}) - g_0(\mathbf{x})\| = \|g_1(\mathbf{x})\| < \frac{\delta_2}{2},$$

and for each subsequent iteration, the difference is halved.

Therefore, for all $\mathbf{x} \in V_1$ and $n \in \mathbb{N}$, we have

$$\begin{aligned} \|g_n(\mathbf{x})\| &= \|g_n(\mathbf{x}) - g_0(\mathbf{x})\| \\ &= \left\| \sum_{l=1}^{n} (g_l(\mathbf{x}) - g_{l-1}(\mathbf{x})) \right\| \\ &\leq \sum_{l=1}^{n} \|g_l(\mathbf{x}) - g_{l-1}(\mathbf{x})\| \\ &< \sum_{l=1}^{n} \frac{1}{2^{l-1}} \frac{\delta_2}{2} \\ &< \delta_2. \end{aligned}$$

Therefore $g_n(\mathbf{x}) \in V_2$, for all $\mathbf{x} \in V_1$ and for all $n \in \mathbb{N}$.

All of the functions g_n are continuous, and according to theorem 3.35, they converge uniformly to the unique continuous function $g : V_1 \to V_2$ which must satisfy $H(\mathbf{x}, g(\mathbf{x})) = g(\mathbf{x})$ for all $\mathbf{x} \in V_1$. That is, $F(\mathbf{x}, g(\mathbf{x})) = \mathbf{0}$ for all $\mathbf{x} \in V_1$. □

We have only shown that g is continuous. However, as noted in theorem 3.34, the fact that $F(\mathbf{x}, g(\mathbf{x})) = \mathbf{0}$ can be shown to imply that also g is totally differentiable at $(\mathbf{a}, \mathbf{b})$, and thus the formula there will also apply. ‘

3.7 Lagrange Multipliers

Theorem 3.39. *Let $G \subset \mathbb{R}^n$ be an open set and let $f : G \to \mathbb{R}$ be continuously differentiable. Assume that $M = \{\mathbf{x} \in G : f(\mathbf{x}) = \mathbf{0}\} \neq \emptyset$. Let $\mathbf{a} \in M$ with $\nabla f(\mathbf{a}) \neq \mathbf{0}$. Assume furthermore that $h : G \to \mathbb{R}$ is another continuously differentiable function such that in some open neighborhood $V \subset G$ with $\mathbf{a} \in V$ we have $h(\mathbf{a}) \geq h(\mathbf{x})$, for all $\mathbf{x} \in M \cap V$. Then there exists some $\lambda \in \mathbb{R}$ with*

$$\nabla h(\mathbf{a}) = \lambda \nabla f(\mathbf{a}).$$

Remark. *The number λ in this theorem is called a* Lagrange multiplier.

Proof. Since $\nabla f(\mathbf{a}) \neq \mathbf{0}$, we must have $\partial_i f(\mathbf{a}) \neq 0$, for some $i \in \{1, \ldots, n\}$. Without loss of generality, assume that $i = n$. For $\mathbf{a} = (a_1, \ldots, a_n)$, take $\mathbf{a}' = (a_1, \ldots, a_{n-1})$, so that $\mathbf{a} = (\mathbf{a}', a_n)$.

According the the theorem on implicit functions (3.38), there exists a neighborhood $V' \times V''$ of $\mathbf{a}$ (so that both $V' \subset \mathbb{R}^{n-1}$ and $V'' \subset \mathbb{R}$ are open sets, and $\mathbf{a}' \in V'$ and $a_n \in V''$) such that there exists a continuously differentiable function $g : V' \to V''$, with

$$M \cap (V' \times V'') = \{\mathbf{x} \in V' \times V'' : x_n = g(x_1, \ldots, x_{n-1})\}.$$

This means that we have

$$f(\mathbf{x}', g(\mathbf{x}')) = 0,$$

for points $\mathbf{x}' = (x_1, \ldots, x_{n-1})$ in an open neighborhood of $\mathbf{a}'$. In particular, for each $i \in \{1, \ldots, n-1\}$ we must have

$$\partial_i f(\mathbf{a}', g(\mathbf{a}')) = \partial_i f(\mathbf{a}) + \partial_n f(\mathbf{a}) \partial_i g(\mathbf{a}') = 0.$$

Let the function $H : V' \to \mathbb{R}$ be

$$H(\mathbf{x}') = h(\mathbf{x}', g(\mathbf{x}')).$$

Then the condition on h, namely that $h(\mathbf{a}) \geq h(\mathbf{x})$, for all $\mathbf{x} \in M \cap V$, means that $H(\mathbf{x}') \geq H(\mathbf{a}')$, for all $\mathbf{x}'$ in an open neighborhood of $\mathbf{a}'$ in V'. Therefore we have

$$\partial_i H(\mathbf{a}') = 0,$$

for all $i \in \{1, \ldots, n-1\}$. But

$$\partial_i H(\mathbf{a}') = \partial_i h(\mathbf{x}', g(\mathbf{x}')) = \partial_i h(\mathbf{a}) + \partial_n h(\mathbf{a}) \partial_i g(\mathbf{a}') = 0.$$

Therefore, since $\partial_n f(\mathbf{a}) \neq 0$, we can write

$$\partial_i h(\mathbf{a}) = -\partial_n h(\mathbf{a}) \partial_i g(\mathbf{a}') = \frac{\partial_n h(\mathbf{a})}{\partial_n f(\mathbf{a})} \left(-\partial_n f(\mathbf{a}) \partial_i g(\mathbf{a}')\right) = \lambda \partial_i f(\mathbf{a}),$$

with

$$\lambda = \frac{\partial_n h(\mathbf{a})}{\partial_n f(\mathbf{a})}.$$

Furthermore, we obviously have

$$\partial_n h(\mathbf{a}) = \frac{\partial_n h(\mathbf{a})}{\partial_n f(\mathbf{a})} \partial_n f(\mathbf{a}) = \lambda \partial_n f(\mathbf{a}).$$

Therefore, taking all the $i = 1, \ldots, n$ together, we have

$$\nabla h(\mathbf{a}) = \lambda \nabla f(\mathbf{a}).$$

□

The condition $f(\mathbf{x}) = 0$ represents a *constraint* on the set of possible points which are to be brought into consideration in the given situation, constraining things to the set M. Then h is a function whose value we are interested in on the constrained set M. The point $\mathbf{a}$ is "optimal"[11] under h with respect to the other points of M. (Of course the theorem also works just as well if we say that $\mathbf{a}$ is a minimal — rather than a maximal — value under h.) Then the theorem says that the gradient, ∇f, which, according to theorem 3.30, gives the direction of the greatest increase of the function, is the same as the gradient ∇h.

The proof used a number of theorems which we proved a while ago, and so you may find it difficult to get a clear picture of what's going on here. Let's think

[11] That is, it is a (local) maximal value.

about the function f first. Since $f(\mathbf{x}) = 0$ everywhere throughout M, it is obvious that the gradient of f at $\mathbf{a}$ must lie perpendicularly to M. But it M has dimension $n-1$; therefore there is only one *single* direction perpendicular to M. On the other hand, thinking about h, consider the set $M_{h(\mathbf{a})} = \{\mathbf{x} \in G : h(\mathbf{x}) = h(\mathbf{a})\}$. If $\nabla h(\mathbf{a}) \neq \mathbf{0}$, then the fact that $f(\mathbf{a})$ is a local extremum in M means that we can't have $M_{h(\mathbf{a})}$ crossing through M at $\mathbf{a}$. The two "hyper-planes" must be tangent to one another at $\mathbf{a}$, and so the perpendicular direction is the same for both. On the other hand, if $\nabla h(\mathbf{a}) = \mathbf{0}$ then obviously the theorem is true if we simply take the Lagrange multiplier λ to be zero.

All of this gives a method for finding a necessary condition that a point be a (locally) extreme point for the function h under the constraint $f(\mathbf{x}) = 0$. It is namely the case that for such a point, the gradients of h and f must have the same directions (or opposite directions if the Lagrange multiplier λ is a negative number).

3.8 Ordinary differential equations

The kinds of differential equations which we will investigate here are of the form

$$y' = f(x, y),$$

where $f : G \to \mathbb{R}$ is some continuous function and $G \subset \mathbb{R}^2$ is an open subset. A *solution* to such a differential equation is a differentiable function $\varphi : I \to \mathbb{R}$, where $I \subset \mathbb{R}$ is some open interval and $(x, \varphi(x)) \in G$ for all $x \in I$, such that

$$\varphi'(x) = f(x, \varphi(x)),$$

for all $x \in I$.

The simplest case is that the function f depends only upon x. That is, we have the differential equation

$$y' = f(x).$$

But we already know how to solve this equation. The solution is simply an anti-derivative to the function f. And we already know that all possible anti-derivatives are given by the integral of f, plus a constant. That is, the solution to this simple form of differential equation is

$$\varphi(x) = \int_{x_0}^{x} f(t)dt + y_0.$$

Here, $y_0 \in \mathbb{R}$ is a constant, and the solution φ has the *initial value* $\varphi(x_0) = y_0$.

Of course if we express the anti-derivative as an integral in this way, we only obtain values of $\varphi(x)$ for $x \geq x_0$. But we can also extend the anti-derivative to values of x less than x_0 by considering the integral

$$-\int_{x}^{x_0} f(t)dt.$$

But what can we do in the more general case? For example consider the differential equation

$$y' = y.$$

Remembering the properties of the exponential function, we can guess that a solution is

$$\varphi(x) = \exp(x).$$

But then, a little further thought convinces us that also $k\exp(x)$ is a solution, for any constant $k \in \mathbb{R}$. Are there further solutions? And more generally, can we solve differential equations of the form $y' = g(y)$, where g is any continuous function?

3.8.1 Separation of variables

The natural thing is to investigate differential equations of the form

$$y' = f(x) \cdot g(y),$$

where both f and g are continuous functions. This is a differential equation which has *separation of its variables*.

Theorem 3.40. *Let I, $J \subset \mathbb{R}$ be open intervals, $f : I \to \mathbb{R}$ and $g : J \to \mathbb{R}$ continuous functions with $g(y) \neq 0$ for all $y \in J$. Let $(x_0, y_0) \in I \times J$ be some "initial value", and take*

$$F(x) = \int_{x_0}^{x} f(t)dt \quad \text{and} \quad G(y) = \int_{y_0}^{y} \frac{ds}{g(s)},$$

for $x \in I$ and $y \in J$. Further, assume that $I' \subset I$ is some open interval contained in I such that $x_0 \in I'$ and $F(I') \subset G(J)$. Then there exists a unique continuously differentiable function $\varphi : I' \to \mathbb{R}$, such that $\varphi(x_0) = y_0$ and

$$\varphi'(x) = f(x)g(\varphi(x)),$$

for all $x \in I'$. And we have $G(\varphi(x)) = F(x)$ for all $x \in I'$.

Proof. Assuming such a φ exists, then we have

$$F(x) = \int_{x_0}^{x} f(t)dt = \int_{x_0}^{x} \frac{\varphi'(t)}{g(\varphi(t))}dt = \int_{y_0}^{\varphi(x)} \frac{ds}{g(s)} = G(\varphi(x)).$$

That is to say, $G(\varphi(x)) = F(x)$. The second equation here follows from the assumed equation $\varphi'(x) = f(x)g(\varphi(x))$, and the third equation follows from the substitution rule for integrals.

Next we prove that φ is unique. Since $G'(y) = \frac{1}{g(y)} \neq 0$, for all $y \in J$, and since G is continuous, it follows that G is a bijection between J and its image $G(J) \subset \mathbb{R}$. Thus there must be an inverse function $H : G(J) \to J$, with $H(G(y)) = y$, for all $y \in J$. But then

$$\varphi(x) = H(G(\varphi(x))) = H(F(x)) = H\left(\int_{x_0}^{x} f(t)dt\right),$$

and it follows that $\varphi(x)$ is uniquely determined.

So the final question is: is $\varphi(x) = H\left(\int_{x_0}^{x} f(t)dt\right)$ really a solution of the differential equation?

Well, we need only differentiate the equation $G(\varphi(x)) = F(x)$ in order to obtain

$$\varphi'(x)G'(\varphi(x)) = \frac{\varphi'(x)}{g(\varphi(x))} = F'(x) = f(x),$$

or $\varphi'(x) = f(x)g(\varphi(x))$, as required. Furthermore, we have

$$\varphi(x_0) = H\left(\int_{x_0}^{x_0} f(t)dt\right) = H(0).$$

But $G(y_0) = 0$. Thus

$$\varphi(x_0) = H(0) = H(G(y_0)) = y_0.$$

□

3.8.2 An example: $y' = x \cdot y$

The equation $y' = x \cdot y$ obviously has separation of variables. We take $I = \mathbb{R}$ and $J = \mathbb{R}_+ = \{x \in \mathbb{R} : x > 0\}$. Then we have

$$F(x) = \int_{x_0}^{x} t\,dt = \frac{x^2 - x_0^2}{2},$$

and

$$G(y) = \int_{y_0}^{y} \frac{dt}{t} = \ln(y) - \ln(y_0) = \ln\left(\frac{y}{y_0}\right).$$

Since the function G is the logarithm, its inverse function H must be the exponential function. In fact, we have

$$y_0 \cdot \exp\left(\ln\left(\frac{y}{y_0}\right)\right) = y,$$

for all $y > 0$. Therefore the solution with the initial value $\varphi(x_0) = y_0$, where $y_0 > 0$, is

$$\varphi(x) = H(F(x)) = y_0 \exp\left(\int_{x_0}^{x} t\,dt\right) = y_0 \exp\left(\frac{x^2 - x_0^2}{2}\right).$$

3.8.3 Another example: homogeneous linear differential equations

The general *first order homogeneous linear differential equation* has the form

$$y' = a(x) \cdot y,$$

where a is a continuous function. This is again a case of separation of variables, and so, using the methods we have developed, the general solution

$$\varphi(x) = y_0 \cdot \exp\left(\int_{x_0}^{x} a(t)dt\right)$$

is immediately obtained. Note that if $\varphi(x_0) = y_0 \neq 0$, then since $\exp(w)$ is always positive for all $w \in \mathbb{R}$, it follows that $\varphi(x) \neq 0$, for all possible x.

3.8.4 Variation of constants

This is the method used to solve *inhomogeneous* first order linear differential equations. That is, equations of the form

$$y' = a(x) \cdot y + b(x).$$

To begin with, let φ be a solution to the *homogeneous* linear differential equation $y' = a(x) \cdot y$, with initial value $\varphi(x_0) = 1$. Thus

$$\varphi'(x) = a(x)\varphi(x),$$

with solution

$$\varphi(x) = \exp\left(\int_{x_0}^{x} a(t)dt\right).$$

Next, we *assume* that the inhomogeneous equation with the extra term $b(x)$ has some solution ψ, so that

$$\psi'(x) = a(x) \cdot \psi(x) + b(x).$$

Given this, then we simply define a new function ζ to be

$$\zeta(x) = \frac{\psi(x)}{\varphi(x)}.$$

That is, $\psi(x) = \zeta(x)\varphi(x)$; but remember that $\varphi'(x) = a(x)\varphi(x)$. Therefore, putting it all together, we obtain

$$\begin{aligned}
\psi'(x) &= \zeta'(x)\varphi(x) + \zeta(x)\varphi'(x) \\
&= \zeta'(x)\varphi(x) + \zeta(x)a(x)\varphi(x) \\
&= a(x)\psi(x) + b(x) \\
&= a(x)\zeta(x)\varphi(x) + b(x).
\end{aligned}$$

Subtracting the term $\zeta(x)a(x)\varphi(x)$ from both sides, we the obtain

$$\zeta'(x)\varphi(x) = b(x),$$

or

$$\zeta'(x) = \frac{b(x)}{\varphi(x)}.$$

Thus ζ is simply an anti-derivative of $\frac{b(x)}{\varphi(x)}$, that is

$$\zeta(x) = \int_{x_0}^{x} \frac{b(t)}{\varphi(t)} dt + K,$$

where $K \in \mathbb{R}$ is some suitable constant. Choosing $K = y_0$ gives us the solution

$$\psi(x) = \zeta(x)\varphi(x) = \exp\left(\int_{x_0}^{x} a(t)dt\right) \cdot \left(\int_{x_0}^{x} \frac{b(t)}{\exp\left(\int_{x_0}^{t} a(s)ds\right)} dt + y_0\right),$$

which satisfies the initial value $\psi(x_0) = y_0$.

3.8.5 The equation $y' = f\left(\frac{y}{x}\right)$

To round off our discussion of special classes of first-order ordinary differential equations, we consider the equation

$$y' = f\left(\frac{y}{x}\right).$$

We are looking for a solution $\varphi : I \to \mathbb{R}$ with an interval $I \subset \mathbb{R}$, such that $0 \notin I$. Here again, f is taken to be continuous and defined on an appropriate open interval of $\mathbb{R}$. Given this, then we have:

Theorem 3.41. *There exists a solution $\varphi : I \to \mathbb{R}$ with*

$$\varphi'(x) = f\left(\frac{\varphi(x)}{x}\right)$$

if and only if

$$\psi'(x) = \frac{f(\psi(x)) - \psi(x)}{x},$$

where $\psi(x) = \frac{\varphi(x)}{x}$.

Proof. Assume first that $\varphi'(x) = f\left(\frac{\varphi(x)}{x}\right)$. Then we have

$$\begin{aligned} \psi'(x) &= \frac{\varphi'(x)}{x} - \frac{\varphi(x)}{x^2} \\ &= \frac{1}{x}\left(f\left(\frac{\varphi(x)}{x}\right) - \frac{\varphi(x)}{x}\right) \\ &= \frac{1}{x}(f(\psi(x)) - \psi(x)). \end{aligned}$$

Conversely, if we assume $\psi'(x) = \frac{f(\psi(x)) - \psi(x)}{x}$, then since we have $\varphi(x) = \psi(x) \cdot x$, it follows

$$\begin{aligned} \varphi'(x) &= \psi'(x) \cdot x + \psi(x) \\ &= \frac{(f(\psi(x)) - \psi(x))}{x} \cdot x + \psi(x) \\ &= f(\psi(x)) \\ &= f\left(\frac{\varphi(x)}{x}\right). \end{aligned}$$

□

Therefore, in order to solve the equation

$$y' = f\left(\frac{y}{x}\right),$$

the first thing to do is to solve the equation

$$z' = \frac{1}{x}(f(z) - z).$$

The equation with z is a case of separation of variables, and we have already seen how to solve such equations. Therefore we obtain a solution z, and the solution y for the original equation becomes $y = x \cdot z$.

3.9 The theorem of Picard and Lindelöf

In our discussion of the method of the variation of constants, we simply *assumed* that some solution to the differential equation must exist. But how do we know if this assumption is a reasonable one? To answer this question we need to give some thought to the general theory of differential equations.

3.9.1 Systems of first order differential equations

In the discussion so far, we have considered single equations of the form $y' = f(x, y)$, where we are looking for a solution of the form $\varphi : I \to \mathbb{R}$. More generally, we can look at a set of n equations which are all linked together.

$$\begin{aligned} y_1' &= f_1(x, y_1, \ldots, y_n) \\ y_2' &= f_2(x, y_1, \ldots, y_n) \\ &\vdots \\ y_1' &= f_1(x, y_1, \ldots, y_n) \end{aligned}$$

We can think of these n components $y_1, \ldots, y_n$ as being the coordinates of a vector $\mathbf{y} \in \mathbb{R}^n$, and so the differential equation can be written as if it were a kind of vector equation: $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$, or in other words

$$\begin{pmatrix} y_1' \\ \vdots \\ y_n' \end{pmatrix} = \begin{pmatrix} f_1(x, \mathbf{y}) \\ \vdots \\ f_n(x, \mathbf{y}) \end{pmatrix}.$$

This differential equation is determined by the function f, so it is necessary to say what it is.

Let $G \subset \mathbb{R} \times \mathbb{R}^n$ be an open subset (of $\mathbb{R}^{n+1}$), and $f : G \to \mathbb{R}^n$ a continuous function. Given some $x_0 \in \mathbb{R}$ and $\mathbf{y}_0 \in \mathbb{R}^n$ with $(x_0, \mathbf{y}_0) \in G$, then a solution to

the differential equation $\mathbf{y}' = f(x, \mathbf{y})$, with initial value $(x_0, \mathbf{y}_0)$, is a differentiable function $\varphi : I \to \mathbb{R}^n$, for some open interval $I \subset \mathbb{R}$, such that $x_0 \in I$, $\varphi(x_0) = \mathbf{y}_0$, and $(x, \varphi(x)) \in G$ for all $x \in I$, and finally, the function φ satisfies the differential equation. That is,

$$\varphi'(x) = \mathbf{f}(x, \varphi(x)),$$

for all $x \in I$.

3.9.2 The Lipschitz condition

Definition. *Again, let $G \subset \mathbb{R} \times \mathbb{R}^n$ be an open subset, and let $\mathbf{f} : G \to \mathbb{R}^n$ be a function. The function $\mathbf{f}$ is said to satisfy a* Lipschitz condition *with* Lipschitz constant $L \geq 0$ *if for all $(x, \mathbf{y})$, $(x, \tilde{\mathbf{y}}) \in G$, we have $\|\mathbf{f}(x, \mathbf{y}) - \mathbf{f}(x, \tilde{\mathbf{y}})\| \leq L\|\mathbf{y} - \tilde{\mathbf{y}}\|$.*

In the theory of differential equations, we usually generalize things somewhat, assuming that the function $\mathbf{f}$ only satisfies a *local* Lipschitz condition. That is to say, the function satisfies a local Lipschitz condition if for every $(x, \mathbf{y}) \in G$, there exists some open neighborhood $U \subset G$ with $(x, \mathbf{y}) \in U$, such that $\mathbf{f}$ satisfies a Lipschitz condition in U.

For simplicity in the discussion here, let us assume that we have a *global* Lipschitz condition, and furthermore it will be assumed that we have just a *single* first order ordinary differential equation. Thus $G \subset \mathbb{R}^2$.

3.9.3 Uniqueness of solutions

Theorem 3.42. *Let $G \subset \mathbb{R}^2$ be an open subset and let $f : G \to \mathbb{R}$ be a continuous function satisfying a Lipschitz condition with Lipschitz constant $L > 0$. Assume $(x_0, y_0) \in G$, $I \subset \mathbb{R}$ is an open interval with $x_0 \in I$, and we have two functions φ, $\psi : I \to \mathbb{R}$ which are both solutions of the differential equation $y' = f(x, y)$, with initial value (x_0, y_0). That is, $\varphi(x_0) = \psi(x_0) = y_0$. Then we have $\varphi(x) = \psi(x)$ for all $x \in I$.*

Proof. We have $\varphi'(x) = f(x, \varphi(x))$. Therefore $\varphi(x) = \int_{x_0}^{x} f(t, \varphi(t))dt + y_0$, and the same is true of the function ψ. Thus for each $x \geq x_0$ we have

$$\begin{aligned} |\varphi(x) - \psi(x)| &= \left| \int_{x_0}^{x} \Big(f(t, \varphi(t)) - f(t, \psi(t)) \Big) dt \right| \\ &\leq \int_{x_0}^{x} \Big| f(t, \varphi(t)) - f(t, \psi(t)) \Big| dt \\ &\leq L \cdot \int_{x_0}^{x} |\varphi(t)) - \psi(t)| dt \end{aligned}$$

For each $x \in I$ with $x \geq x_0$, let

$$M(x) = \sup\{|\varphi(t) - \psi(t)| : x_0 \leq t \leq x\}.$$

In particular, for all t between x_0 and x, we have

$$|\varphi'(t) - \psi'(t)| = |f(t, \varphi(t)) - f(t, \psi(t))| \leq L \cdot |\varphi(t) - \psi(t)| \leq L \cdot M(x).$$

Therefore

$$|\varphi(t) - \psi(t)|' \leq L \cdot M(x).$$

Then, using the intermediate value theorem and noting that $\varphi(x_0) = \psi(x_0)$, we see that

$$|\varphi(t) - \psi(t)| \leq |t - x_0| \cdot L \cdot M(x),$$

for all t between x_0 and x. In particular, this implies that

$$M(x) \leq |x - x_0| \cdot L \cdot M(x).$$

But if we choose x to be sufficiently close to x_0 so that

$$|x - x_0| < \frac{1}{2L},$$

then we obtain

$$M(x) \leq \frac{1}{2} M(x).$$

This can only be true if $M(x) = 0$, or in other words, $\varphi(t) = \psi(t)$ for all $t \geq x_0$, with $|t - x_0| < 1/2L$.

Now take $x_1 = \sup\{\xi \in I : \varphi(t) = \psi(t), \forall t \in [x_0, \xi]\}$. There cannot be any elements of I greater than x_1 since for all points t of I nearer than $1/2L$ to x_1, we must have $\varphi(t) = \psi(t)$. Thus, for all elements of I greater than x_0, we must have φ and ψ being equal.

The argument can also be extended to show that for all elements of I less than x_0, the two functions are equal. For this we need only note that we would have

$$\varphi(x) = -\int_x^{x_0} f(t, \varphi(t)) dt + y_0,$$

and the analogous expression for $\psi(x)$. □

Examples

- Linear differential equations $y' = a(x) \cdot y + b(x)$ obviously satisfy a local Lipschitz condition. For let x be an element of the open interval I where the equation is defined. Then let $I' \subset I$ be a finite closed interval such that x is contained in the interior of I'. Since the function a is assumed to be continuous, it is uniformly continuous on I'. Let $L > 0$ be chosen with $L > a(x')$, for all $x' \in I'$. Then we have

$$|(a(x) \cdot y + b(x)) - (a(x) \cdot \tilde{y} + b(x))| \leq L|y - \tilde{y}|.$$

- The standard example of a differential equation which does not satisfy a Lipschitz condition is

$$y' = \sqrt{|y|}.$$

For example, if we take $\tilde{y} = 0$, then in order to have

$$\left|\sqrt{|y|} - \sqrt{|\tilde{y}|}\right| = \sqrt{|y|} \leq L|y - \tilde{y}| = L|y|,$$

we would have to have

$$L \geq \frac{1}{\sqrt{|y|}}.$$

Yet, as $y \to 0$, the fraction $1/\sqrt{|y|} \to \infty$.

Specifically, we can think of a number of different solutions. For example the function $\varphi_0(x) = 0$ obviously satisfies the equation $y' = \sqrt{|y|}$. Another solution is $\varphi_1(x) = x^2/4$. Obviously we have $\varphi_0(0) = \varphi_1(0) = 0$. That is to say, given the initial value $(x_0, y_0) = (0, 0)$, then we have two *different* solutions of the differential equation, starting from the same initial value. More generally, for all $k \in \mathbb{R}$, the function

$$\varphi(x) = \begin{cases} 0, & x < -k \\ \frac{(x+k)^2}{4}, & x \geq -k \end{cases}$$

is a solution to the differential equation $y' = \sqrt{|y|}$. (But note that this differential equation is of the form "seperation of variables". Thus, according to theorem 3.40, the solution is unique if we start with an initial value such that $y \neq 0$, and confine the solution to a region where it remains not equal to zero.)

3.9.4 Existence of solutions

Theorem 3.43. *Again, $G \subset \mathbb{R}^2$ open; $f : G \to \mathbb{R}$ continuous, satisfying a Lipschitz condition with constant $L > 0$. Let $(x_0, y_0) \in G$. Then there exists an open interval $I \subset \mathbb{R}$ with $x_0 \in I$, and a continuously differentiable function $\varphi : I \to \mathbb{R}$, such that $\varphi(x_0) = y_0$, $(x, \varphi(x)) \in G$ and $\varphi'(x) = f(x, \varphi(x))$, for all $x \in I$.*

Proof. We show how to find $\varphi(x)$, for $x > x_0$. The procedure for $x < x_0$ is analogous.

To begin, since G is open, there must exist some $\delta > 0$ such that the square

$$S_{(x_0,y_0)}(\delta) = \{(x, y) : |x - x_0| \leq \delta \text{ and } |y - y_0| \leq \delta\} \subset G.$$

Since f is continuous, there must exist some $M > 0$, such that $|f(x, y)| \leq M$, for all $(x, y) \in S_{(x_0,y_0)}(\delta)$.) So let

$$\epsilon = \min\left\{\delta, \frac{\delta}{M}, \frac{1}{2L}\right\}$$

and then take

$$I = (x_0 - \epsilon, x_0 + \epsilon).$$

The next thing to do is to define recursively a sequence of functions $\varphi_n : I \to \mathbb{R}$ as follows. We start with the constant function

$$\varphi_0(x) = y_0.$$

Then, for each $n \in \mathbb{N}$, we take

$$\varphi_n(x) = \int_{x_0}^{x} f(t, \varphi_{n-1}(t))dt + y_0.$$

Obviously $\varphi_n(x_0) = y_0$, for all n. Furthermore, we also have

$$(x, \varphi_n(x)) \in S_{(x_0,y_0)}(\delta) \subset G,$$

for all n. In order to see this, we begin by observing that $(x, \varphi_0(x)) = (x, y_0) \in S_{(x_0,y_0)}(\delta)$ for all $x \in I$, since we must have $|x - x_0| < \epsilon \leq \delta$.

So now let $n \in \mathbb{N}$ be given, and we assume inductively that $(x, \varphi_{n-1}(x)) \in S_{(x_0,y_0)}(\delta)$ for all $x \in I$. Then we have

$$\begin{aligned} |\varphi_n(x) - y_0| &= \left| \int_{x_0}^{x} f(t, \varphi_{n-1}(t))dt \right| \\ &\leq \int_{x_0}^{x} |f(t, \varphi_{n-1}(t))|dt \\ &\leq |x - x_0| \cdot M \\ &\leq \frac{\delta}{M} \cdot M \\ &= \delta. \end{aligned}$$

Therefore $(x, \varphi_n(x)) \in G$, for all n.

The next step is to show that the sequence of functions φ_n converges uniformly to a function $\varphi : I \to \mathbb{R}$ which is a solution to the differential equation $y' = f(x, y)$. Writing $\|\cdot\|$ for the supremum norm, we have

$$\begin{aligned} |\varphi_{n+1}(x) - \varphi_n(x)| &= \left| \int_{x_0}^{x} (f(t, \varphi_n(t)) - f(t, \varphi_{n-1}(t)))dt \right| \\ &\leq \int_{x_0}^{x} L|\varphi_n(t) - \varphi_{n-1}(t)|dt \\ &\leq L \cdot |x - x_0| \cdot \|\varphi_n - \varphi_{n-1}\| \\ &\leq L \cdot \frac{1}{2L} \cdot \|\varphi_n - \varphi_{n-1}\| \\ &= \frac{1}{2}\|\varphi_n - \varphi_{n-1}\|. \end{aligned}$$

Since this is true for all $x \in I$ with $x > x_0$, we have

$$\|\varphi_{n+1} - \varphi_n\| \leq \frac{1}{2}\|\varphi_n - \varphi_{n-1}\|.$$

Thus we see that the sequence of continuous functions φ_n is a Cauchy sequence with respect to the supremum norm. Therefore it converges uniformly to a function $\varphi : I \to \mathbb{R}$. We have $\varphi(x_0) = y_0$ and $(x, \varphi(x)) \in G$, for all $x \in I$. Furthermore, using theorem 2.48, we obtain

$$\varphi(x) = \lim_{n\to\infty} \varphi_n(x) = \lim_{n\to\infty} \int_{x_0}^{x} f(t, \varphi_{n-1}(t))dt = \int_{x_0}^{x} f(t, \varphi(t))dt,$$

and so we must have

$$\varphi'(x) = f(x, \varphi(x))$$

for all $x \in I$. □

Remarks

- In this proof, we have assumed that $x > x_0$, but as has been repeatedly remarked, it is a simple matter to alter the proof in order to deal with the values of x in I which are less than x_0.

- Since we confined things to the small square $S_{(x_0,y_0)}(\delta)$ around the point $(x_0, y_0) \in G$, it is clear that we only needed to have a Lipschitz condition in that square. That is, the theorem is also true if the function f only satisfies a *local* Lipschitz condition.

- Our interval I, which contains the initial value x_0, is taken to be small in order to ensure that the sequence of functions φ_n do not bring us out of the region G. Also I must be sufficiently small to ensure that we have the contraction $\|\varphi_{n+1} - \varphi_n\| \leq \frac{1}{2}\|\varphi_n - \varphi_{n-1}\|$. But then, given that the solution φ is defined along the interval I, we can take a point near the end of I and use that as the initial value, constructing an extension of the domain os φ. In general this procedure allows us to extend the interval along which φ is defined, in fact going out to the edge of the region G. Such ideas are dealt with more fully in the many books on differential equations in the library, and also in the lecture devoted to differential equations in our faculty.

- The method of proof describes a practical method for finding solutions of differential equations. Given an initial value $(x_0, y_0) \in G$, we take the first approximation to be simply the constant function $\varphi_0(x) = y_0$, for all $x \in I$. Then the sequence φ_n, for $n \in \mathbb{N}$ should converge to a solution. This is called the *Picard-Lindelöf iteration method.*

- When dealing with systems of first order differential equations, we have vectors in $\mathbb{R}^n$, rather than just numbers in $\mathbb{R}$. The iteration step is then a vector equation

$$\varphi_n(x) = \int_{x_0}^{x} \mathbf{f}(t, \varphi_{n-1}(t))dt + \mathbf{y}_0.$$

Here $x \in I \subset \mathbb{R}$, but $\mathbf{f}(t, \varphi_{n-1}(t)) \in \mathbb{R}^n$ and $\mathbf{y}_0 \in \mathbb{R}^n$. The integral becomes an integral over a vector-valued function.

$$\int_{x_0}^{x} \mathbf{f}(t, \varphi_{n-1}(t))dt = \int_{x_0}^{x} \begin{pmatrix} f_1(t, \varphi_{n-1}(t)) \\ \vdots \\ f_n(t, \varphi_{n-1}(t)) \end{pmatrix} dt,$$

and each of the components $f_i(t, \varphi_{n-1}(t))$ is just a function $f_i : I \to \mathbb{R}$. So we integrate each of the components separately.

3.10 Ordinary differential equations of higher order

These are equations of the form

$$y^{(n)} = f(x, y, y', \ldots, y^{(n-1)}),$$

where $y^{(n)}$ is the n-th derivative. That is, given an initial value (x_0, y_0), then we are looking for a solution $\varphi : I \to \mathbb{R}$, with $\varphi(x_0) = y_0$ and

$$\varphi^{(n)}(x) = f(x, \varphi(x), \varphi'(x), \ldots, \varphi^{(n-1)}(x)),$$

for all $x \in I$.

The method is to convert this into a system of n first-order differential equations in the variables $y_1, \ldots, y_n$. To begin with, let $y_1 = y$. then take

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= y_3 \\ &\vdots \\ y_{n-1}' &= y_n \\ y_n' &= f(x, y_1, \ldots, y_n). \end{aligned}$$

This reduces the problem to that of solving systems of first order equations. And given a solution

$$\varphi(x) = \begin{pmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \vdots \\ \varphi_n(x) \end{pmatrix},$$

then $\varphi_1 : I \to \mathbb{R}$ is clearly a solution to the original equation

$$y^{(n)} = f(x, y, y', \ldots, y^{(n-1)}).$$

Example

Consider the simple equation $y'' = -y$. This describes (without bothering about additional constants) the harmonic oscillator. In order to solve the equation, we reduce it to a system of two first order equations, namely

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= -y_1 \end{aligned}$$

But we have already seen in an exercise that the solution (with the initial value $\varphi(0) = 1$) is $\varphi(x) = \cos(x)$.

3.11 Partial differential equations

Ordinary differential equations depend on one parameter, x. The most general form for such an equation would be

$$F(x, y, y', \dots, y^{(n)}) = 0.$$

Thus, for example if we have an equation of the form

$$y' = f(x, y),$$

this becomes

$$F(x, y, y') = y' - f(x, y) = 0.$$

Perhaps is is natural to think of such equations as describing the movement of a particle through space, where the parameter x describes the time. For example, Gauss spent huge amounts of time, involving hundreds of thousands — even millions — of arithmetical operations, trying to calculate the paths of various astroids in their movements about the sun.

It is also possible to consider differential equations which depend on more than one parameter, say $x_1, x_2, \dots, x_n$. We then have the theory of *partial differential equations*. Such an equation will be of the form

$$F(x_1, x_2, \dots, x_n, \mathbf{y}, \partial_1\mathbf{y}, \dots, \partial_n\mathbf{y}, \partial_1\partial_1\mathbf{y}, \partial_1\partial_2\mathbf{y}, \dots, \partial_n\partial_n\mathbf{y}, \dots) = \mathbf{0}.$$

Here, a solution to the equation is a function $\varphi : G \to \mathbb{R}^m$, where $G \subset \mathbb{R}^n$ is some open subset, and the function F has some finite number of possible partial derivatives.

Special classes of partial differential equations, such as the Laplace equation

$$\partial_1^2\varphi(x_1, \dots, x_n) + \cdots + \partial_n^2\varphi(x_1, \dots, x_n) = 0,$$

with given boundary values, have been studied theoretically. But in general, it is impossible to find exact solutions. Instead, people use numerical methods to find approximate solutions.

Such things are very important in many practical situations. For example when people design an airplane, then it is necessary to calculate the flow of air over the wings using the partial differential equations of fluid dynamics. And then, of course the stresses within the wing itself must be calculated in order to determine what strengths the various components must have. For this, one uses the method of "finite element analysis", which again is a way of finding an approximate solution to a system of partial differential equations. In the early 1800s, when Gauss was active, such calculations were hardly feasible, at least for "normal" people. But these days, such methods are applied all the time, using computers and standard software libraries.

All of this is beyond the scope of the Analysis 2 lectures. Still, it may be interesting to have a quick look at some methods which are used for dealing with ordinary differential equations.

3.12 Numerical methods for solving ordinary differential equations

3.12.1 Euler's method

Given the differential equation $y' = f(x, y)$, and the initial value (x_0, y_0), then Euler's method for finding an approximate solution is to look at things in a discrete sequence of steps

$$x_0,\ x_0 + \Delta x,\ x_0 + 2\Delta x,\ x_0 + 3\Delta x, \ldots$$

That is to say, things are calculated at the points

$$x_0,\ x_1,\ x_2,\ x_3, \ldots$$

where

$$x_n = x_{n-1} + \Delta x,$$

and Δx is some fixed distance between one calculation and the next.

But what are the corresponding values of y for each of these x_n? The rule is:

$$y_n = y_{n-1} + \Delta x \cdot f(x_{n-1}, y_{n-1}),$$

progressing through increasing values of n in $\mathbb{N}$. In this way we obtain a sequence of points

$$(x_0, y_0),\ (x_1, y_1),\ (x_2, y_2), \ldots$$

and then connecting the points with straight line segments, we hope to get some sort of approximation to the correct solution.

A simple example: $y' = x$

If the initial value is $(0, 0)$, then we are looking for the function $\varphi : \mathbb{R} \to \mathbb{R}$ with $\varphi(0) = 0$ and $\varphi'(x) = x$, for all $x \in \mathbb{R}$. Obviously, the correct solution is

$$\varphi(x) = \frac{1}{2}x^2$$

as can be seen by observing that this is the anti-derivative of the function x.

So what does Euler's method make of this problem if we take the discrete step length to be $\Delta x = 1$? We obtain the sequence of points

$(0,\ 0)$
$(1,\ 0)$
$(2,\ 1)$
$(3,\ 3)$
$(4,\ 6)$
etc.

But the correct solution $\varphi(x) = \frac{1}{2}x^2$ goes through the points

$(0,\ 0)$
$(1,\ \frac{1}{2})$
$(2,\ 2)$
$(3,\ 4\frac{1}{2})$
$(4,\ 8)$
etc.

So we see that Euler's method is not particularly good in this case.

3.12.2 The Runga-Kutta method

The simplest version of the Runga-Kutta method is to use the rule

$$y_n = y_{n-1} + \frac{\Delta x}{2}\Big(f(x_{n-1}, y_{n-1}) + f(x_{n-1} + \Delta x, y_{n-1} + \Delta x f(x_{n-1}, y_{n-1}))\Big).$$

This gives the sequence of points

$(0,\ 0)$
$(1,\ \frac{1}{2})$
$(2,\ 2)$
$(3,\ 4\frac{1}{2})$
$(4,\ 8)$
etc.

And we see that this is gives us precisely the points of the correct solution!

Of course this example is rather special. Experimenting with more general examples, one usually finds that this simple Runga-Kutta method is superior to the Cauchy method, but it is also not particularly efficient.

There are various possibilities for obtaining a better calculation, depending on the details of the given equation which is to be solved. One such method uses a 4-step iteration.

Given the equation $y' = f(x, y)$, with initial value (x_0, y_0), let φ be a solution. In particular, we have that the initial value is satisfied: $\varphi(x_0) = y_0$.

Now let $h > 0$ be the discrete step length. The problem is to calculate a sensible approximation to the number k, such that $\varphi(x_0 + h) = y_0 + k$. Then, of course we can set $x_1 = x_0 + h$ and $y_1 = y_0 + k$ and continue the calculation from there. By taking h to be small, and using the fastest computer obtainable, one hopes to piece together a reasonably good solution to the given equation.

The method for finding k, given f and (x_0, y_0), is given by the following scheme. Begin by setting $x = x_0$ and $y = y_0$. Then we have

$$\begin{aligned} y_I &= y + k_I/2, & k_I &= f(x, y)h \\ y_{II} &= y + k_{II}/2, & k_{II} &= f(x + h/2, y_I)h \\ y_{III} &= y + k_{III}, & k_{III} &= f(x + h/2, y_{II})h \\ & & k_{IV} &= f(x + h, y_{III})h \end{aligned}$$

and then finally,

$$k = \frac{1}{6}(k_I + 2k_{II} + 2k_{III} + k_{IV}),$$

so that we have the starting point for the next step in the calculation, namely

$$x_1 = x + h, \text{ and } y_1 = y + k.$$

3.13 The variational calculus: a sketch

The most general way to think about the variational calculus is to imagine that we have some abstract set X, together with a real-valued function $F : X \to \mathbb{R}$ which is bounded below. The problem is then: find some $x_0 \in X$ (if such a thing exists!) such that $F(x_0) \leq F(x)$, for all possible $x \in X$. That is, x_0 is an element with the *minimal* possible value.

If we are looking for an element with the *maximal* value, then that is the same as looking for some x_0 such that $-F(x_0)$ has a *minimal* value.

For example, in the theory of economics, it might be imagined that we have a factory which produces various things which can be sold at various prices. Should more workers be employed, or should some be made redundant? Which combinations of raw materials at what prices should be bought? And so on and so forth. Each of the possible combinations is an element of the set X of different possible ways of running the factory. In the end, the amount of profit the factory makes

is some number $F(x)$, which might be calculated for each of the possible elements of X. Economists then imagine that the factory manager will choose to run the factory according to the method $x_0 \in X$, which gives the greatest profit.

But such a level of generality brings us away from practical mathematics. Let us therefore restrict ourselves to the kind of variational calculus which describes practical situations in the physical world, and which are described in terms of differential equations.

Examples

- The problem which was posed by the mathematician Johann Bernoulli in the journal *Acta Eruditorum* in June 1696, and which led to the formulation of the theory of the variational calculus, was the *Problem of the Brachystochrone*. He wrote:

  "I, Johann Bernoulli, address the most brilliant mathematicians in the world. Nothing is more attractive to intelligent people than an honest, challenging problem, whose possible solution will bestow fame and remain as a lasting monument. Following the example set by Pascal, Fermat, etc., I hope to gain the gratitude of the whole scientific community by placing before the finest mathematicians of our time a problem which will test their methods and the strength of their intellect. If someone communicates to me the solution of the proposed problem, I shall publicly declare him worthy of praise."

  The problem was the following:

  "Given two points A and B in a vertical plane, what is the curve traced out by a point acted on only by gravity, which starts at A and reaches B in the shortest time."

  Many mathematicians accepted the challenge. For example it is said that Newton (who at that time was the Director of the Royal Mint)

  "in the midst of the hurry of the great recoinage, did not come home till four (in the afternoon) from the Tower very much tired, but did not sleep till he had solved it, which was by four in the morning."

- Another problem, which is perhaps more practical, is the following: What is the shape of a telephone wire which hangs steadily, in equilibrium under gravity between two points A and B?

The general form of such problems is: find some function y of x such that the value of

$$F(y) = \int f(x, y, y')dx$$

is as small as possible.

For example, looking at Bernoulli's problem, imagine that the point A has the coordinates (x_1, y_1) in the Euclidean plane $\mathbb{R}^2$, and the point B has the coordinates

(x_2, y_2). It is natural to imagine that $x_1 < x_2$ and that $y_1 > y_2$. Furthermore, it seems clear that the optimal curve would not switch directions, or loop around itself. Thus it could be described as a function $\varphi : [x_1, x_2] \to \mathbb{R}$, presumably sufficiently smooth to be differentiable to any desired degree, such that $\varphi(x_1) = y_1$ and $\varphi(x_2) = y_2$. Then for each x, the value of the function $\varphi(x)$ gives the point (x, y) through which the curve passes, where $y = \varphi(x)$. If $y < y_1$, then the speed at the point $y' = \varphi'(x)$ is given by equating the potential energy which has been lost with the kinetic energy which the point would have in its passage through (x, y). Since the problem is to find the curve giving the shortest time from x_1 to x_2, the function f should measure the speed in the horizontal direction.

Returning to the more general problem, let $G \subset \mathbb{R}^3$ be some open subset, and let $f : G \to \mathbb{R}$ be a function which is at least twice continuously partially differentiable. Then the problem is to find a function $\varphi : I \to \mathbb{R}$ such that $(x, \varphi(x), \varphi'(x)) \in G$, for all $x \in I$ with $I = [a, b]$, such that

$$F(\varphi) = \int_a^b f(x, \varphi(x), \varphi'(x))dx$$

is as small as possible.

One way to do this is to think of other possible functions $\tilde{\varphi} : I \to \mathbb{R}$, and compute the values of $F(\tilde{\varphi})$, checking to see if they are always greater than, or equal to $F(\varphi)$. Writing

$$\psi = \tilde{\varphi} - \varphi,$$

we obtain a new function $\psi : I \to \mathbb{R}$ which is such that $\psi(a) = \psi(b) = 0$. (It is assumed that all of these functions are at least continuously differentiable.)

Generalizing things slightly, let us take $(-\delta, +\delta)$ to be a small open interval around zero. Then we can examine the functions $\varphi + s\psi$, for various values of $s \in (-\delta, +\delta)$. This gives us a new function

$$\Gamma : (-\delta, +\delta) \to \mathbb{R},$$

such that

$$\Gamma(s) = F(\varphi + s\psi) = \int_a^b f\Big(x, \varphi(x) + s\psi(x), \varphi'(x) + s\psi'(x)\Big)dx.$$

Since f is continuous, it follows that Γ is differentiable, and if φ is a solution to our variational problem then it must be that

$$\Gamma'(0) = 0.$$

We then have

$$\begin{aligned}
\Gamma'(0) &= \left.\frac{d}{ds}\right|_{s=0} \int_a^b f\Big(x, \varphi(x) + s\psi(x), \varphi'(x) + s\psi'(x)\Big)dx \\
&= \int_a^b \left.\frac{d}{ds}\right|_{s=0} f\Big(x, \varphi(x) + s\psi(x), \varphi'(x) + s\psi'(x)\Big)dx \\
&= \int_a^b \left(\psi(x)\frac{\partial}{\partial y} f(x, \varphi(x), \varphi'(x)) + \psi'(x)\frac{\partial}{\partial y'} f(x, \varphi(x), \varphi'(x))\right) dx \\
&= \int_a^b \left(\psi(x)\frac{\partial}{\partial y} f(x, \varphi(x), \varphi'(x))\right) dx + \int_a^b \left(\psi'(x)\frac{\partial}{\partial y'} f(x, \varphi(x), \varphi'(x))\right) dx \\
&= \int_a^b \left(\psi(x)\frac{\partial}{\partial y} f(x, \varphi(x), \varphi'(x))\right) dx - \int_a^b \left(\psi(x)\frac{d}{dx}\frac{\partial}{\partial y'} f(x, \varphi(x), \varphi'(x))\right) dx \\
&= \int_a^b \psi(x)\left(\frac{\partial}{\partial y} f(x, \varphi(x), \varphi'(x)) - \frac{d}{dx}\frac{\partial}{\partial y'} f(x, \varphi(x), \varphi'(x))\right) dx \\
&= 0
\end{aligned}$$

Here:

- The first equation is just the definition of the function Γ.
- The second equation follows by observing that if we have a function g which depends on two variables, x and s, then

$$\begin{aligned}
\frac{d}{ds}\int_a^b g(x,s)dx &= \lim_{h\to 0}\frac{1}{h}\int_a^b \Big(g(x, s+h) - g(x,s)\Big)dx \\
&= \lim_{h\to 0}\int_a^b \frac{g(x,s+h) - g(x,s)}{h}dx.
\end{aligned}$$

  And if g is continuously partially differentiable, then as $h \to 0$ we have uniform convergence of the fraction

$$\frac{g(x,s+h) - g(x,s)}{h}$$

  to

$$\frac{\partial}{\partial s} g(x,s).$$

- In the third equation, the notation $\frac{\partial}{\partial y} f(x, \varphi(x), \varphi'(x))$ means the partial derivative with respect to the second component of f, and $\frac{\partial}{\partial y'}$ is the partial derivative with respect to the third component. The fact that the third equation is true is a consequence of the chain rule for derivatives.
- The fourth equation is trivial.
- The fifth equation follows using partial integration and the fact that $\psi(a) = \psi(b) = 0$.

- Finally, the sixth equation is trivial.

Since this must hold for all possible variational functions ψ, we conclude that the *Euler-Lagrange differential equation*

$$\frac{\partial}{\partial y} f(x, \varphi(x), \varphi'(x)) - \frac{d}{dx} \frac{\partial}{\partial y'} f(x, \varphi(x), \varphi'(x)) = 0$$

must hold for a solution φ to our variational problem. (This follows from the so-called *Fundamental Lemma* of the variational calculus, which is an exercise.)

How do we evaluate the expression

$$\frac{d}{dx} \frac{\partial}{\partial y'} f(x, \varphi(x), \varphi'(x)) \, ?$$

For this we can think of $\frac{\partial}{\partial y'} f$ as defining a function of three variables, let's call it g for simplicity. Then we have

$$\frac{\partial}{\partial y'} f(x, \varphi(x), \varphi'(x)) = g(x, \varphi(x), \varphi'(x)).$$

We use the chain rule to obtain that

$$\begin{aligned} \frac{d}{dx} g(x, \varphi(x), \varphi'(x)) &= \frac{\partial}{\partial x} g(x, \varphi(x), \varphi'(x)) + \varphi'(x) \frac{\partial}{\partial y} g(x, \varphi(x), \varphi'(x)) \\ &\quad + \varphi''(x) \frac{\partial}{\partial y'} g(x, \varphi(x), \varphi'(x)). \end{aligned}$$

So the Euler-Lagrange equation becomes

$$\frac{\partial}{\partial y} f(x, y, y') - \frac{\partial^2}{\partial x \partial y'} f(x, y, y') - y' \frac{\partial^2}{\partial y \partial y'} f(x, y, y') - y'' \frac{\partial^2}{\partial y'^2} f(x, y, y') = 0.$$

This still looks rather complicated. Things become simpler if our function f does not depend explicitly upon x. In this case $\frac{\partial^2}{\partial x \partial y'} f(x, y, y') = 0$, and we can simply write $f(y, y')$, rather than $f(x, y, y')$. Therefore

$$\frac{\partial}{\partial y} f(y, y') - y' \frac{\partial^2}{\partial y \partial y'} f(y, y') - y'' \frac{\partial^2}{\partial y'^2} f(y, y') = 0.$$

Since

$$\frac{d}{dx} \Big(f(y, y') - y' \frac{\partial}{\partial y'} f(y, y') \Big) = y' \Big(\frac{\partial}{\partial y} f(y, y') - y' \frac{\partial^2}{\partial y \partial y'} f(y, y') - y'' \frac{\partial^2}{\partial y'^2} f(y, y') \Big) = 0,$$

it follows that

$$f(y, y') - y' \frac{\partial}{\partial y'} f(y, y') = k,$$

for some constant $k \in \mathbb{R}$. That is, substituting $\varphi(x)$ for y, we obtain the equation

$$f(\varphi(x), \varphi(x)') - \varphi(x)' \frac{\partial}{\partial y'} f(\varphi(x), \varphi(x)') = k.$$

(Of course we must maintain the notation $\frac{\partial}{\partial y'} f$ to indicate the partial derivative with respect to the second term in f.)

Examples

- We begin with the brachystochrone. Let $A = (x_1, y_1)$ and $B = (x_2, y_2)$. In order to simplify the notation, let us say that $x_1 = 0$ and $y_1 = 0$, and furthermore, y increases as we go *downwards*. According to the principles of classical physics, the velocity v of the particle will be

$$v = \sqrt{2gy},$$

where g is the gravitational constant.[12] But the velocity v is a function of time t. Let us consider the velocity v_x in the horizontal direction, and v_y in the vertical direction. Then we have $v = \sqrt{v_x^2 + v_y^2}$. Writing

$$v_x = \frac{dx}{dt} \quad \text{and} \quad v_y = \frac{dy}{dt},$$

we obtain

$$\frac{v_y}{v_x} = \frac{dy}{dx} = y'(x).$$

Thus

$$v = \sqrt{2gy} = v_x\sqrt{1 + y'^2} = \frac{dx}{dt}\sqrt{1 + y'^2}.$$

This leads to the equation

$$T = \int_0^T dt = \frac{1}{\sqrt{2g}} \int_0^{x_2} \frac{\sqrt{1 + y'^2}}{\sqrt{y}} dx,$$

where T is the time it takes for the particle to travel horizontally to x_2.

Therefore we can write

$$f(x, y, y') = \frac{\sqrt{1 + y'^2}}{\sqrt{y}},$$

and we see that x does not specifically occur in f. Using the equation

$$f(y, y') - y' \frac{\partial}{\partial y'} f(y, y') = k,$$

we obtain

$$\frac{1}{\sqrt{y(1 + y'^2)}} = k,$$

which finally gives us the differential equation

$$y' = \sqrt{\frac{1}{k^2 y} - 1}.$$

[12] If the mass of the particle is m then the change in the potential energy when falling the distance y is given by the product gmy. The kinetic energy which the particle then has is $\frac{1}{2}mv^2$, and since we assume that v is zero when y is zero, it follows that $\frac{1}{2}mv^2 = gmy$.

A substitution now shows that the solution is a cycloid:

$$x = r(\theta - \sin\theta), \quad y = r(1 - \cos\theta),$$

where θ can be considered to be a function of x, and r is some appropriate constant.

- As far as calculating the shape of a freely hanging telephone wire is concerned, the idea is that a solution must be such that the potential energy of the wire must be as small as possible. If the wire has a weight of m kilograms per meter, and if the wire follows the curve $\varphi : [a, b] \to \mathbb{R}$, then the potential energy is given by

$$\int_a^b mg\varphi(x)\sqrt{1 + (\varphi'(x))^2}dx,$$

where g is the gravitational constant. Setting both of the constants g and m to 1, we obtain the variational problem

$$F(\varphi) = \int_a^b \varphi(x)\sqrt{1 + (\varphi'(x))^2}dx.$$

But there is a further complication, owing to the fact that we assume the length of the wire to be fixed.[13] So let the length be L, a number greater than the distance between the two endpoints A and B. This gives us the further condition

$$\int_a^b \sqrt{1 + (\varphi'(x))^2}dx = L.$$

In order to solve this problem, we use the method of *Lagrange multipliers*.

The idea is that since L remains the same for all the possible functions φ which come into question, it must be that a solution will satisfy the variational problem given by the integral

$$\tilde{F}(\varphi) = \int_a^b (\varphi(x) + \lambda)\sqrt{1 + (\varphi'(x))^2}dx,$$

for some constant $\lambda \in \mathbb{R}$. That is, if $\tilde{\Gamma}(s) = \tilde{F}(\varphi + s\psi)$, for possible variations ψ, then we will have $\tilde{\Gamma}'(s) = 0$.

So here, the Euler-Lagrange equation is

$$f(y, y') - y'\frac{\partial}{\partial y'}f(y, y') = k,$$

with

$$f(y, y') = (y + \lambda)\sqrt{1 + y'^2}.$$

[13] Of course one could make things even more complicated by assuming that the weight of the wire varies along its length, and that it is elastic, like a rubber band. But for our present purposes, a fixed weight and a fixed length will be assumed.

Therefore

$$(y+\lambda)\sqrt{1+y'^2} - \frac{(y+\lambda)y'^2}{\sqrt{1+y'^2}} = k,$$

or

$$y + \lambda = k\sqrt{1+y'^2}.$$

The solution has the form

$$y = k\cosh\left(\frac{x-k_*}{k}\right) - \lambda,$$

where $k_* \in \mathbb{R}$ is another constant.