NWI: Mathematics

Literature

- These lecture notes!
- Various books in the library with the title Linear Algebra I, or Analysis I. (And also Linear Algebra II, or Analysis II.)
- The lecture notes of some of the people who have given these lectures in the past. In particular:

  http://www.techfak.uni-bielefeld.de/fachschaft/skripten.html

  http://www.math.uni-bielefeld.de/ mspiess/Lehrealt.html

  http://www.math.uni-bielefeld.de/ froyshov/nwi2/index.html

  http://www.math.uni-bielefeld.de/ froyshov/nwi/index.html

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

To begin with, we have the "usual" and simple 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 then we have somewhat more unusual systems:

- The real numbers $\mathbb{R}$
- The complex numbers $\mathbb{C}$
- The residue classes modulo n, for $n \in \mathbb{N}$, namely $\mathbb{Z}/n\mathbb{Z}$

1.1 How computers deal with numbers

A computer deals with information by manipulating many tiny transistors, each of which either does, or does not, have a particular electrical potential applied to it. We can think of this as the transistor representing either the digit "0", or else the digit "1". In this way, natural numbers can be represented according to the following scheme.

$$\begin{array}{lcl} 1 & : & 1 \\ 2 & : & 10 \\ 3 & : & 11 \\ 4 & : & 100 \\ 5 & : & 101 \\ 6 & : & 110 \\ 7 & : & 111 \\ 8 & : & 1000 \end{array}$$
$$etc.$$

This is arithmetic to the base 2. The *binary system* of arithmetic. Earlier "computers": mechanical adding machines, often had toothed wheels for representing numbers. Thus some of the wheels

might have had 10 teeth, representing the numbers in our decimal system. Also persons counting numbers on the fingers of their two hands represent numbers in the decimal system. Mechanical clocks have toothed wheels with 60 teeth, representing the numbers to the base 60. This system continues the tradition of arithmetic used by the Babylonians in the ancient world.

All of these machines — clocks, mechanical calculators, fingers, computers — are *finite* systems. Therefore, since the set of natural numbers $\mathbb{N}$ is infinite, it follows that there is a limit to how much can be represented in each machine. For example, a mechanical clock generally goes through a cycle of 12 hours, then at midnight, or noon, it starts over again from the beginning.

Similarly, the standard unit of a computer, a byte, consists of 8 transistors, so it can only represent $2^8 = 256$ different numbers. The usual convention is to think of these as being the integers from 0 to 255.

How do we do arithmetic in the system of numbers from 0 to 255? Well, for example 2+2=4 is OK, since both the numbers 2 and 4 are represented in one byte. Namely, we have

$$00000010 + 00000010 = 00000100$$

in binary arithmetic. But how about 200+200? Unfortunately, 400 is not represented in the byte, since the number 400 is greater than 255. In fact, when the computer reaches 256, it simply cycles back to zero. Specifically, we have that 200 is 11001000 in the binary system. Therefore 200+200 is

$$11001000 + 11001000 = 1\underbrace{10010000}_{\text{8 places}}.$$

Unfortunately, there are 9 places (or bits) in 110010000. This is too much to fit into a byte. Therefore the last bit is simply removed, discarded. So we have

$$11001000 + 11001000 = 10010000.$$

Or, expressed in the more usual decimal notation, we have

$$200 + 200 = 144.$$

What we are doing here is *modular* arithmetic, modulo 256.

1.2 The system $\mathbb{Z}/n\mathbb{Z}$ for $n = 256$

We have seen that in the system of modular arithmetic modulo 256, we have the equation

$$200 + 200 = 144.$$

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

$$200 + 200 = 400 = 1 \times 256 + 144.$$

More generally, given any integer x, and any natural number n, 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. For example, in the C language, the whole number division of x by n is given by the instruction

```
a = x / n;
```

The remainder term is given by the instruction

```
b = x % n;
```

In mathematics, we use the notation

$$b = x \bmod n.$$

In particular then, we have the equation[1]

$$144 = 400 \bmod 256.$$

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 we 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 $\frac{1}{3}$ in our modular arithmetic modulo 256? That is, the problem is to find some number z with $0 \leq z < 256$, such that

$$1 = (z \times 3) \bmod 256.$$

The answer? It is $z = 171$, since $171 \times 3 = 513$, and $1 = 513 \bmod 256$.

On the other hand, what is $\frac{1}{2}$ modulo 256? That is, let z be such that

$$1 = (z \times 2) \bmod 256.$$

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 256. The reason for this is that for all z we always have $z \times 2$ being an even number, yet since 256 is also an even number, it must be that the equation $1 = y \bmod 256$ can only have a solution when y is an *odd* number.

So we see that arithmetic modulo 256 gives us problems when it comes to division. In fact, all of this discussion just shows that arithmetic in computers is much more complicated than we might at first have thought.

[1] Actually, for historical reasons, it is more correct to write "$b \equiv x \bmod n$" here. That is, we have an "equivalence relation". I will come back to this later. But for the moment, we should just think of the number "$x \bmod n$" as being the result of the arithmetical operation of finding the remainder after dividing x by n. This is always a number between 0 and $n - 1$.

1.3 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.*[2] *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 $n = 2$.)

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.*

[2] 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 $y \in Y$.

1.4 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

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

But rather than writing $x \sim y$, it is more usual to write

$$x \equiv y \bmod n$$

when describing this equivalence relation. One says that "x is congruent to y modulo 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 $x, y \in \mathbb{Z}$.

However, 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 $ax \equiv b \bmod n$?

1.5 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[3]

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

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 $b' < b$.

[3] 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.

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. □

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

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 I know that this is true??

Well, first of all, I 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 I 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 I already know that the formula is true for the case $n = 1$.

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

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

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

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

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.[4] The set of integers $\mathbb{Z}$ has addition, subtraction and multiplication, but division fails. However, in the 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.)

[4] 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 $\mathbb{N}$.

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 of the groups we will deal with in these lectures are definitely not Abelian. For example the matrix groups — which are used continuously when a computer calculates 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*

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

Some simple consequences of this 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
$$b = 1 \cdot b = (a^{-1}a)b = a^{-1}(ab) = a^{-1} \cdot 0 = 0.$$

□

Which groups and fields are important for these lectures?

The groups we will use:

- Of course fields are themselves groups under addition. So all fields — that is to say, all the number systems we will consider — are themselves groups.
- Linear algebra is concerned with vectors. A system of vectors is called a *vector space*. Each vector space is a group, with respect to *vector addition*.
- The set of *linear transformations* (rotations, inversions, changes of perspective) of a vector space are described using invertible, square matrices. These matrices form a non-Abelian group under matrix multiplication.
- When dealing with determinants of matrices, we will consider the group of *permutations* of n objects. This is also a non-Abelian group.

The fields we will use:

- We have already seen the two fields which give us the most basic number systems in mathematics: namely the rational numbers $\mathbb{Q}$ and the modular system $\mathbb{Z}/p\mathbb{Z}$, for prime numbers p.
- The real numbers $\mathbb{R}$ are constructed by "filling in the gaps" in $\mathbb{Q}$. This is the basis of real analysis, which will constitute half of these lectures.
- But after constructing $\mathbb{R}$, we see that something is still missing. Many polynomials, for example the polynomial $x^2 + 1$, have no roots in the system of numbers $\mathbb{R}$. To solve this problem, the system of complex numbers $\mathbb{C}$ will be constructed.

1.9 Analysis and Linear Algebra

Mathematics, as it is taught today in universities, always begins with two separate series of lectures. Namely Analysis and Linear Algebra. Only later, particularly when it comes to the subject of *functional analysis*, do we see that analysis and linear algebra are, in many ways, just two different aspects of the same thing. But, unfortunately (or fortunately??) you, as students of information technology, will probably leave pure mathematics before reaching that stage. In any case, I will continue these lectures by talking about analysis and linear algebra as if they were two entirely different subjects. In these lecture notes, they will be dealt with in two different chapters, which will be developed simultaneously.

Analysis

Analysis can be thought of as being the study of the real and the complex numbers. The idea of *functions* plays the important role. Which functions are *continuous*, or *differentiable*? How does *integration* work? How do we solve simple systems of differential equations? How should we define the basic functions, such as the *exponential* function, the *logarithm* function, the trigonometric functions, etc.? For this, we need to think about whether or not a given *infinite series* of numbers *converges*, or not.

Linear Algebra

This is concerned with geometry. How can a computer work out movements, perspective in 3-dimensional space, and then represent these on a 2-dimensional screen? What is a *basis* for a coordinate system? When is a set of vectors *linearly independent*? How are *linear mappings* of vector spaces represented by *matrices*? And then, taking a step away from geometry, how do we solve systems of linear equations using the methods of linear algebra? This last question is very important when it comes to the use of computers in economics.

Chapter 2

Analysis

2.1 Injections, Surjections, Bijections

The subject of mathematical analysis is mainly concerned 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.

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 nonempty 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.

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. 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

For example, 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.[2]

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.

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

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.)*

Theorem 2.4. *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 $\mathbb{R}$.

For let $(a_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $\mathbb{R}$. Let $A = \{q \in \mathbb{Q} : \forall N \in \mathbb{N}, \exists n \geq N, \text{ with } q < a_n\}$. Then let $B = \mathbb{Q} \setminus A$. If it happens to be the case that A has a largest element, say x_0, then that should be transferred over to B. That is, we change A to $A \setminus \{x_0\}$ and B to $B \cup \{x_0\}$. This gives a Dedekind cut of $\mathbb{Q}$, representing the real number $a \in \mathbb{R}$, say, and then 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 $N \in \mathbb{N}$, such that $|a - a_n| < \epsilon$, for all $n \geq N$.

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 $|p - q| < \epsilon/3$. Looking at the definition of the set A, we see that the number p must be such that for sufficiently large n, all the numbers a_n are less than p. On the other hand, given such an a_n which is greater than q, then it must be between q and p. That means that the distance between q and a_n must be less than $\epsilon/3$.

Since the sequence $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, there exists a number $N(\epsilon/3) \in \mathbb{N}$ such that for all n, $m \geq N(\epsilon/3)$, we have $|a_n - a_m| < \epsilon/3$. Then setting $N = N(\epsilon/3)$, and taking $m \geq N$ with $q < a_m$, we have for all $n \geq N$

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

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

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

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.

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

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

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

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. □

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 6.1.
- 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 $0 \leq x < 1$. 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

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

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[3] — 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

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

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

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

[3]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.

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 $(0,1)$.

Why are closed intervals important here?

After all, it's no use telling you to be careful without telling you what to be careful about! Why does our theorem 2.24 work for *closed* intervals $[a,b]$, where the endpoints a and b are included in the interval, yet it does not work for intervals which are not closed?

If you look at the proof of theorem 2.24, you will see that the property of the interval $[a,b]$ which we used was that every convergent sequence (in $\mathbb{R}$) of elements of $[a,b]$, converges to a point which is also in $[a,b]$. Or put another way, every Cauchy sequence in a closed interval converges to a point in that interval. Furthermore, the interval $[a,b]$ is bounded. In these lectures, at least at this stage of things, I will just skip over these ideas quickly, simply listing some of the things which are dealt with more thoroughly in a lecture for students of pure mathematics.

Metric spaces; open, closed, compact, complete subspaces. A quick sketch.

- A *metric space* M is a set, together with a *metric* — which is a "distance function",

  $$d : M \times M \to \mathbb{R},$$

  such that for all x, y, $z \in M$, we have

  1. $d(x,x) = 0$,
  2. $d(x,y) \geq 0$,
  3. $d(x,y) = d(y,x)$, and
  4. $d(x,z) \leq d(x,y) + d(y,z)$.

For example, $\mathbb{R}$ is a metric space, with the metric $d(x, y) = |x - y|$.

- Let M be a metric space, and let $U \subset M$ be a subset. We say that U is *open* in M if, for all $x \in U$, there exists some $\epsilon > 0$ such that $\{y \in M : d(y, x) < \epsilon\} \subset U$. On the other hand, a subset $A \subset M$ is called *closed*, if $M \setminus A$ is open in M.

- Let $K \subset M$ be some subset. An *open covering* of K is some collection $\{U_i : i \in I\}$ (where the variable i runs through some indexing set I) of open subsets of M, such that the union of all the subsets covers K. That is $K \subset \cup_{i \in I} U_i$.

  The subset $K \subset M$ is called *compact* if every open covering of K contains a finite subcovering, that is, some finite number of the U_i, which is itself a covering of K. If $K = M$, then we say that the whole metric space M is compact. (But even if $K \neq M$, still, the subset K itself is a compact metric space.)

- We then have the theorem which says that in any compact metric space, every Cauchy sequence converges to a point in that space. That is to say, every compact space is *complete*.

- Finally, we have the theorem (of Heine-Borel) which says that a subset $K \subset \mathbb{R}$ is compact if, and only if, it is closed and bounded.

So putting all of these ideas together, we can say that the important thing to think about in theorem 2.24 is that it deals with closed and bounded intervals of $\mathbb{R}$. That is, it deals with *compact* subsets of $\mathbb{R}$.

Open intervals, like (a, b), or half-open intervals like $(a, b]$, and so on, are *not* compact. They are not complete. But also the entire set of real numbers $\mathbb{R}$ is not compact, even though it is complete.

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)**.** *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

We have already seen that the equation $x^2 + 1 = 0$ has no solution within the system of real numbers $\mathbb{R}$. To solve this "problem", mathematicians have simply 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,

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

Anticipating the ideas of linear algebra somewhat, we see that $\mathbb{C}$ is really a 2-dimensional vector space over $\mathbb{R}$. Therefore it is natural to picture the numbers in $\mathbb{C}$ on the 2-dimensional plane, the horizontal axis representing $\mathbb{R}$, the real numbers, and the vertical axis representing the imaginary numbers $i\mathbb{R}$.

We have seen how important it is to think about the distance between two numbers in analysis. Therefore, 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

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

This means that

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

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

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

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

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

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

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 exactly the same, one need only replace the symbol for absolute values in the real numbers, namely $|\cdot|$, with the symbol in the complex numbers, $\|\cdot\|$. In particular, we see that a sequence $(z_n)_{n\in\mathbb{N}}$, with $z_n = a_n + ib_n$, for each n, converges if and only if both the sequences of real numbers $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ converge. In particular, 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.[4] 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 $\mathbb{C}$.

2.14 The trigonometric functions: sin and cos

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 sin and cosine functions are *defined* in terms of Euler's formula

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

Since $e^{i(-x)} = e^{-ix} = \overline{e^{ix}}$, it follows that

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

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,

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

[4] The exponential function is also everywhere differentiable, but in this lecture, we will not think about extending the idea of differentiation to complex numbers. The study of *complex analysis* (this is called "Funktionentheorie" in German) differs from *real analysis*, which is what we are mainly concerned with here.

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

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

The trigonometric functions can also 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, and 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. But first, we should think about how complex-valued functions of real variables can be differentiated.

Definition. *Let $U \subset \mathbb{R}$ be an open interval, and let $f : U \to \mathbb{C}$ be a complex-valued function. Then, as in the case of real-valued functions, the derivative of f at $x_0 \in U$ is defined to be the complex number $f'(x_0)$ if, for all $\epsilon > 0$, a $\delta > 0$ exists such that, for all $x \in U$ with $|x - x_0| < \delta$ and $x \neq x_0$, we have*

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

So this is exactly the same definition as before, but with the slight difference that we now use the absolute value function for complex numbers.

How does this work in practice? Given $f : U \to \mathbb{C}$, we can write

$$f(x) = f_r(x) + i f_i(x)$$

for each $x \in U$. Here, $f_r : U \to \mathbb{R}$ and $f_i : U \to \mathbb{R}$ are both simply real-valued functions. So we see that if f is differentiable at the point $x_0 \in U$, then it must be the case that both f_r and f_i are also differentiable at x_0, and we have

$$f'(x_0) = f_r'(x_0) + i f_i'(x_0).$$

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

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

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

□

2.15 The number π

Unfortunately, there is not really enough time in this lecture to deal properly with the trigonometric functions. So I will simply sketch out the ideas, without proof.

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*

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

Definition. *Let $f : [a, b] \to \mathbb{R}$ be a function, and let $P = \{[t_0, t_1], \ldots, [t_{n-1}, t_n]\}$ 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

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

- "$\Leftarrow$": 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. *the function fg is integrable.*[5]

[5]But, of course, we do not always have $\int fg = \int f \cdot \int g$. For example, $\int_{-1}^{+1} xdx = 0$, yet $\int_{-1}^{+1} x^2 dx = \frac{2}{3}$.

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.[6]

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.

$$g(x) = f\left(\frac{m(b-a)}{n}\right) + \frac{\epsilon}{2(b-a)}$$

and

$$h(x) = f\left(\frac{m(b-a)}{n}\right) - \frac{\epsilon}{2(b-a)},$$

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 (Intermediate 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*

$$\int_a^b f(x)g(x)dx = f(\xi)\int_a^b g(x)dx.$$

[6] If f is only defined on an open interval (a, b) then the integral may not exist, even if f is continuous. For example, $\lim_{\epsilon \to 0} \int_\epsilon^1 \frac{1}{x} dx = \infty$.

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 $f(\xi) = \mu$. □

2.17.5 Axiomatic characterization of the Riemann integral

Another way of thinking about integrals is to examine the set $\mathcal{C}_C(\mathbb{R})$ of continuous functions of compact support on $\mathbb{R}$. So a function $f : \mathbb{R} \to \mathbb{R}$ is an element of $\mathcal{C}_C(\mathbb{R})$ if there exists some bound $K > 0$, such that $f(x) = 0$ for all $x \in \mathbb{R}$ with $|x| > K$. As we have just seen, the integral of f can be taken, say over the interval $[-K, K]$. And, even expanding the integral to larger values of K, since f is just zero outside the interval, it doesn't change the value of the integral. Therefore, it makes sense to define this to be the integral over the whole set of real numbers, from $-\infty$ to $+\infty$. Therefore

$$\int_{-\infty}^{+\infty} f(x)dx$$

exists, for all $f \in \mathcal{C}_C(\mathbb{R})$. This gives us a function $int : \mathcal{C}_C(\mathbb{R}) \to \mathbb{R}$. As we have already seen, the function int has the properties:

1. $int(f+g) = int(f) + int(g)$,
2. $int(\lambda f) = \lambda\, int(f)$,
3. $f \leq g \Rightarrow int(f) \leq int(g)$,

for all $f, g \in \mathcal{C}_C(\mathbb{R})$ and $\lambda \in \mathbb{R}$. That is, int is a *linear and monotonic functional* on the set $\mathcal{C}_C(\mathbb{R})$

In addition, it is not difficult to see that the function int has the additional property of being *translation-invariant*. That is, given $f \in \mathcal{C}_C(\mathbb{R})$, and given some $a \in \mathbb{R}$, then we define the translated function $\tau_a f \in \mathcal{C}_C(\mathbb{R})$ to be $\tau_a f(x) = f(x-a)$, for all $x \in \mathbb{R}$. Then we have

$$\int_{-\infty}^{+\infty} f(x-a)dx = \int_{-\infty}^{+\infty} f(x)dx.$$

That is, $int(\tau_a f) = int(f)$, for all $f \in \mathcal{C}_C(\mathbb{R})$.

Given all this, then we have the following theorem, which I won't take the time to prove in this lecture.

Theorem 2.43. *Let $I : \mathcal{C}_C(\mathbb{R}) \to \mathbb{R}$ be any linear, monotonic, and translation-invariant functional. Then there exists a constant $c \in \mathbb{R}$, such that*

$$I(f) = c\int_{-\infty}^{+\infty} f(x)dx,$$

for all $f \in \mathcal{C}_C(\mathbb{R})$.

Remark. *The theorem is also true if we substitute the set of all possible finite step functions for the set $C_C(\mathbb{R})$. That is, let $\mathcal{T}_C(\mathbb{R})$ be the set of all step functions with finitely many steps, such that the intervals defining the steps are also finite. Let $I : \mathcal{T}_C(\mathbb{R}) \to \mathbb{R}$ be any linear, monotonic, and translation-invariant functional. Then there exists a constant $c \in \mathbb{R}$, such that*

$$I(f) = c \int_{-\infty}^{+\infty} f(x)dx,$$

for all $f \in \mathcal{T}_C(\mathbb{R})$.

2.18 The fundamental theorem of calculus

Theorem 2.44. *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. One need only observe that

$$F(x + h) - F(x) = \int_{x+h}^{x} f(x)dx.$$

□

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 2.45. *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 2.46. *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 2.47. *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 chain 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 $f : [a, b] \to \mathbb{R}$ and $g : [c, d] \to \mathbb{R}$ be differentiable functions, 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 then 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:

$$\int_c^d f(g(x))g'(x)dx = (F \circ g)(x)\Big|_c^d = F(g(d)) - F(g(c)) = \int_{g(c)}^{g(d)} f(x)dx.$$

2.19 Taylor series; Taylor formula

2.19.1 The Taylor formula

Theorem 2.48 (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 $f^{(k+1)}(x) = \big(f^{(k)}(x)\big)'$. 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 $n = 1$, 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 $n \geq 1$. 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

$$\begin{aligned} R_n(x) &= \frac{1}{(n-1)!}\int_{x_0}^{x}(x-t)^{n-1}f^{(n)}(t)dt \\ &= -\int_{x_0}^{x} f^{(n)}(t)\left(\frac{(x-t)^n}{n!}\right)' dt \\ &= -f^{(n)}(t)\frac{(x-t)^n}{n!}\Big|_{x_0}^{x} + \int_{x_0}^{x}\frac{(x-t)^n}{n!}f^{(n+1)}(t)dt \\ &= \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \int_{x_0}^{x}\frac{(x-t)^n}{n!}f^{(n+1)}(t)dt, \end{aligned}$$

which is just the next term in the Taylor formula, with the corresponding remainder term. □

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 intermediate value theorem for integrals to find some $\xi \in [a,b]$ 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

$$f(x) = f(x_0)+\frac{f'(x_0)}{1!}(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+\frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}.$$

2.19.2 The Taylor series

If f is infinitely differentiable[7] 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.

[7] 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.

2.19.3 Power series, Fourier series, etc.

Unfortunately, there is simply too little time in this semester to explain many of the important things you should know about the analysis of real functions.

Thinking about Taylor series leads naturally to the idea of examining more general power series of the form

$$\sum_{n=0}^{\infty} a_n x^n,$$

for possible values of $a_n \in \mathbb{R}$. We have the theorem that any continuous function $f : [a, b] \to \mathbb{R}$ can be arbitrarily well approximated by polynomials. That is, for each $\epsilon > 0$, there exists some polynomial $P(x) = a_n x^n + \cdots + a_1 x + a_0$, which is such that

$$|f(x) - P(x)| < \epsilon,$$

for all $x \in [a, b]$.

Then we have the theory of Fourier series. These are series of the form

$$\sum_{k=0}^{\infty} (a_k \cos(kx) + b_k \sin(kx)),$$

with a_k, $b_k \in \mathbb{R}$. One proves that each continuously differentiable periodic function with period 2π can be expressed as a Fourier series. In fact, by changing the length of the period, and generalizing the idea of a Fourier series to accommodate this idea, we see that as with power series, all continuously differentiable functions on closed intervals can be expressed as Fourier series. Then, noticing that the sine and cosine functions really come from the complex exponential function, we are lead to consider exponential sums. And so on.

These subjects — and many more which would be handled in a more extensive introduction to the theory of real analysis — are treated in many books in the library, and of course they can also be found in innumerable sites on the internet.

2.20 More general integrals

While it is all very nice to think of integrals as being simply ways of measuring the area "below the curve" of a real-valued function defined on a closed interval $[a, b] \subset \mathbb{R}$, in many practical situations, this is much too restrictive.

For example, these days many people seem to be extremely concerned about the average temperature of the Earth. So if we imagine that the surface of the Earth is a 2-sphere S^2, and the temperature at each point $x \in S^2$ on the sphere is some real number $t(x)$, then we obtain a function $t : S^2 \to \mathbb{R}$. Given this, then the average temperature would be given by

$$\frac{1}{vol(S^2)} \int_{S^2} t(x) dx.$$

But how do we do an integral over a 2-dimensional object? And even worse, how do we integrate over a sphere? It makes no sense to use our 1-dimensional Riemann integral here.

Another example comes up in probability theory. A probability space is an abstract concept, where the probabilities that different possible events might occur are assigned various values. Often the probability space is taken to be infinite, and these values are described in terms of a "probability measure". Calculating actual probabilities then involves finding integrals with respect to this measure.

2.20.1 Measure theory, general integrals: a brief sketch

Let X be some non-empty set. A *σ-algebra* Σ on X is a set of subsets of X such that:

- Σ is not empty,
- if $A \in \Sigma$, then also $X \setminus A \in \Sigma$, and
- the union of any countable number of sets in Σ is again in Σ.

A *measure* on a σ-algebra on X is a real function $\mu : \Sigma \to [0, \infty]$, such that

- $\mu(\emptyset) = 0$, and
- if $A_1, A_2, \ldots$ is a countable sequence of disjoint (that is, $A_i \cap A_j = \emptyset$, if $i \neq j$) sets in Σ, then
$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

We say that the triple (X, Σ, μ) is a *measure space*, and the elements of Σ are called the *measurable sets* of X.

Given this, then it is — in principle — a simple business to define the idea of an integral on a measure space, using "step functions" where the steps are given by the measurable sets. There are some complications. For example, it is a good idea to have the measure being "σ-finite" (this is the case if X is a countable union of measurable sets of finite measure). We then imagine that the measurable sets in Σ are like the intervals ($[a, b]$, or $[a, b)$, and so forth), in $\mathbb{R}$. Then, given a real function $f : X \to \mathbb{R}$, we form upper and lower integrals to f, and if they are equal, then the function is itself integrable.

But, at least at this stage of things, such ideas are just too general for us.

2.21 Integrals in $\mathbb{R}^n$; Fubini's theorem

Let us consider the case of a function $f : G \to \mathbb{R}$, where $G \subset \mathbb{R}^n$. To make things as simple as possible, we can consider the case that G is an n-dimensional rectangle in $\mathbb{R}^n$. That is, for each $i = 1, \ldots, n$, we have some closed interval $[a_i, b_i] \subset \mathbb{R}$, and

$$G = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : a_i \leq x_i \leq b_i, \forall i = 1, \ldots, n\}.$$

In fact, to make things even more simple, let us just consider two intervals $[a, b]$ and $[c, d]$ in $\mathbb{R}$, and then we take G to be the rectangle $G = [a, b] \times [c, d] \subset \mathbb{R}^2$.

Let $f : G \to \mathbb{R}$ be some continuous function[8]. That means in particular that for all $x_0 \in [a, b]$, the function $f(x_0, \cdot) : [c, d] \to \mathbb{R}$ with $f(x_0, \cdot)(y) = f(x_0, y)$, for all $y \in [c, d]$, is continuous. Also for all $y_0 \in [c, d]$, the function $f(\cdot, y_0) : [a, b] \to \mathbb{R}$ is continuous.

Given this, then we obtain a new function $\phi : [a, b] \to \mathbb{R}$, by taking

$$\phi(x) = \int_c^d f(x, y)dy,$$

for each $x \in [a, b]$.

[8]The definition here is entirely analogous to our earlier definition for 1-dimensional functions: Let $u, v \in \mathbb{R}^n$. Then the distance between them is taken to be $\|u - v\|$. So if $f : G \to \mathbb{R}$ is some function from the subset $G \subset \mathbb{R}^n$ to $\mathbb{R}$, then we say that f is continuous at the point $x_0 \in G$ if for all $\epsilon > 0$, there exists a $\delta > 0$, such that for all $y \in G$ with $\|y - x_0\| < \delta$ we have $|f(y) - f(x_0)| < \epsilon$.

f is continuous if it is continuous at all points of G.

Theorem 2.49. *The function ϕ is continuous.*

Proof. Since the function $f(\cdot, y) : [a, b] \to \mathbb{R}$ is continuous, and since $[a, b]$ is a closed interval, it follows that $f(\cdot, y)$ must be uniformly continuous. Therefore, given an $\epsilon > 0$, there must exist some $\delta > 0$, such that for all x_1 and $x_2 \in [a, b]$ with $|x_1 - x_2| < \delta$, we have

$$|f(x_1, y) - f(x_2, y)| < \frac{\epsilon}{d - c}.$$

But then we have

$$\begin{aligned} |\phi(x_1) - \phi(x_2)| &= \left| \int_c^d f(x_1, y)dy - \int_c^d f(x_2, y)dy \right| \\ &\leq \int_c^d |f(x_1, y) - f(x_2, y)|dy \\ &< \int_c^d \frac{\epsilon}{d - c}dy = \epsilon. \end{aligned}$$

□

But since ϕ is continuous, it follows that the integral

$$\int_a^b \phi(x)dx = \int_a^b \left(\int_c^d f(x, y)dy \right) dx$$

must exist.

Remark. *If we take the measure space on $\mathbb{R}^2$ whose σ-algebra is the smallest one which contains all rectangles $[x_1, x_2] \times [y_1, y_2]$, and which is such that the measure μ is, in each case,*

$$\mu([x_1, x_2] \times [y_1, y_2]) = (x_2 - x_1) \cdot (y_2 - y_1),$$

then we will find that the integral of f is, in fact, equal to $\int_a^b \int_c^d f(x, y)dydx$.

This measure is called "Lebesgue measure".

2.21.1 Fubini's theorem

Theorem 2.50 (Fubini's theorem — in two dimensions)**.**

$$\int_a^b \left(\int_c^d f(x, y)dy \right) dx = \int_c^d \left(\int_a^b f(x, y)dx \right) dy$$

Proof. Let us take two partitions, namely

$$a = p_1 < p_2 < \cdots < p_n = b$$

of $[a, b]$, and

$$c = q_1 < q_2 < \cdots < q_m = d$$

of $[c, d]$. This gives us a partition of the rectangle $[a, b] \times [c, d]$ into smaller rectangles of the form $[p_i, p_{i+1}] \times [q_j, q_{j+1}]$. Let us call such a partition Δ.

Given Δ, then we can consider a kind of Riemann sum of f by taking the double sum

$$S_\Delta(f) = \sum_{i=1}^{n} \sum_{j=1}^{m} f(x_i, y_i) \cdot (p_i - p_{i-1})(q_j - q_{j-1}),$$

where $p_{i-1} \leq x_i < p_i$ and $q_{j-q} \leq y_j < q_j$, for each i and j.

Since we know that f is uniformly continuous, it follows that for each $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|f(x,y) - f(x',y')| < \frac{\epsilon}{(b-a)(d-c)}$$

whenever $\sqrt{(x-x')^2 + (y-y')^2} < \delta$. So choose the partition Δ to be sufficiently fine that for each rectangle $[p_i, p_{i+1}] \times [q_j, q_{j+1}]$ in Δ, the maximum distance between any two points is at most δ.

Now take the function $g : [a,b) \times [c,d) \to \mathbb{R}$ to be such that for each $v \in G$ we have $g(v)$ being the infimum of the values of $f(u)$, for all $u \in [p_i, p_{i+1}) \times [q_j, q_{j+1})$, which is the rectangle containing v. In this way, g is defined on the rectangle $[a,b) \times [c,d)$, and the values on the boundary of G are fixed by assuming that g is a continuous function defined everywhere in G. Therefore we have $g \leq f$, and so for the Riemann sum $S_\Delta(g) \leq S_\Delta(f)$.

Similarly, we define $h : G \to \mathbb{R}$ by taking $h(v)$ to be the supremum of the values of $f(u)$, for all $u \in [p_i, p_{i+1}) \times [q_j, q_{j+1})$, which is the rectangle containing v. Then we have $f \leq h$ and $S_\Delta(g) \leq S_\Delta(f)$.

If the lower and upper integrals are defined as in the one dimensional case of the Riemann integral, then we obtain

$$S_\Delta(g) \leq \int_* f \leq \int^* f \leq S_\Delta(h).$$

But

$$\begin{aligned} S_\Delta(h) - S_\Delta(g) &= \sum_{i=1}^{n}\sum_{j=1}^{m}(h(x_i,y_i) - g(x_i,y_i)) \cdot (p_i - p_{i-1})(q_j - q_{j-1}) \\ &\leq \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\epsilon}{(b-a)(d-c)} \cdot (p_i - p_{i-1})(q_j - q_{j-1}) \\ &= \frac{\epsilon}{(b-a)(d-c)} \cdot (b-a)(d-c) = \epsilon. \end{aligned}$$

Therefore, since ϵ could be taken to be an arbitrarily small positive number, we must have $\int_* f = \int^* f$. Then, looking at the way the double integral

$$\int_a^b \left(\int_c^d f(x,y)dy \right) dx$$

was defined in terms of one dimensional integrals; that is, using partitions of the intervals $[a,b]$ and $[c,d]$, we see that we must have

$$\int_* f = \int^* f = \int_a^b \left(\int_c^d f(x,y)dy \right) dx.$$

But the same argument shows that also

$$\int_* f = \int^* f = \int_c^d \left(\int_a^b f(x,y)dx \right) dy.$$

□

More generally, we can think about an n-dimensional rectangle G in $\mathbb{R}^n$, and a continuous function $f : G \to \mathbb{R}$. Then, by looking at typical points $(x_1, x_2, \ldots, x_n) \in G$, and their values $f(x_1, x_2, \ldots, x_n) \in \mathbb{R}$, we see that it is possible to treat the different x_i as being independent variables, and so we obtain the multiple integral

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_n}^{b_n} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n.$$

Then our proof of Fubini's theorem in the 2-dimensional case can be easily extended to show that this multiple integral is, in fact, the integral for f over G, and the order in which the variables in the multiple integration is taken makes no difference to the final result.

2.21.2 Axiomatic characterization of integrals in R^n

This is entirely analogous to the 1-dimensional case. We are again concerned with integrals of continuous functions of compact support. We take this to mean continuous functions $f : \mathbb{R}^n \to \mathbb{R}$, such that, for each such function, there is some finite n-dimensional rectangle G, with $f(\mathbf{x}) = 0$ if $\mathbf{x} \notin G$. Let us denote the set of all such functions by $\mathcal{C}_C(\mathbb{R}^n)$. Then we define

$$\int_{\mathbb{R}^n} f(\mathbf{x}) d^n x$$

to be simply the integral over G.

This integral is again linear and monotonic. Furthermore, it is translation-invariant with respect to arbitrary translations in n-dimensional space. That is, let $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$ be some arbitrary vector, and let $\tau_{\mathbf{v}} f(\mathbf{x}) = f(\mathbf{x} + \mathbf{v})$. Then we always have

$$\int_{\mathbb{R}^n} f(\mathbf{x}) d^n x = \int_{\mathbb{R}^n} \tau_{\mathbf{v}} f(\mathbf{x}) d^n x.$$

So the analogue of theorem 2.43 is

Theorem 2.51. *Let $I : \mathcal{C}_C(\mathbb{R}^n) \to \mathbb{R}$ be any linear, monotonic, and translation-invariant functional. Then there exists a constant $c \in \mathbb{R}$, such that*

$$I(f) = c \int_{\mathbb{R}^n} f(\mathbf{x}) d^n x,$$

for all $f \in \mathcal{C}_C(\mathbb{R}^n)$.

Again, a proof would take too much time to explain in this lecture. One reference for this is the book "Analysis 3", by Otto Forster.

2.22 Regions in $\mathbb{R}^n$; open sets, closed sets

The idea of always confining our thinking to rectangular regions in $\mathbb{R}^n$ is obviously rather awkward. In analysis, particularly when considering whether or not a function is differentiable, it is convenient to assume that it is defined on an "open" subset of $\mathbb{R}^n$.

Definition. *Let $G \subset \mathbb{R}^n$. Then G is called* open *if, for all $\mathbf{x} \in G$, there exists some positive number $\epsilon > 0$, such that $\{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{x}\| < \epsilon\} \subset G$. On the other hand, G is* closed *if $\mathbb{R}^n \setminus G$ is open.*

The set $B_\epsilon(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{x}\| < \epsilon\}$ is called the *open ball* with center $\mathbf{x}$ and radius ϵ. Thus we can say that a set is open in $\mathbb{R}^n$ if for each of its elements $\mathbf{x}$, there exists some positive number $\epsilon > 0$, such that the open ball around $\mathbf{x}$ with radius ϵ lies entirely within the set.

Examples

- In $\mathbb{R}$, every closed interval $[a, b]$ is closed; every open interval (a, b) is open. But $[a, b)$ is neither open, nor closed.
- In $\mathbb{R}^n$, the empty set $\emptyset$ and $\mathbb{R}^n$ itself are both open and closed.
- If $\mathbf{V} \subset \mathbb{R}^n$ is a subspace of dimension less than n, then it is closed.
- The $n-1$-sphere $S^{n-1} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}$ is closed in $\mathbb{R}^n$.

2.22.1 The topology of metric spaces

The definition can be generalized further to all metric spaces. That is, let M be a metric space, with metric $d : M \times M \to \mathbb{R}$. Then a subset $U \subset M$ is called *open*, if for all $x \in U$, there exists some $\epsilon > 0$, such that the open ball with radius ϵ around x, namely $B_\epsilon(x) = \{y \in M : d(y, x) < \epsilon\}$ is contained within U. The subset $V \subset M$ is called *closed* if $M \setminus V$ is open.

So given such a metric space as M, let $\mathcal{O}$ be the set of all possible open subsets of M. It is not difficult to see that $\mathcal{O}$ has the following properties.

1. Both $\emptyset$ and M are elements of $\mathcal{O}$.
2. Given two subsets U and $V \in \mathcal{O}$, then we also have $U \cap V \in \mathcal{O}$.
3. Let $\mathcal{I} \subset \mathcal{O}$ be any collection of open subsets of M. ($\mathcal{I}$ could be finite, or countably infinite, or even uncountably infinite!) Then the union of all the sets in $\mathcal{I}$ is also in $\mathcal{O}$. Namely $\cup_{U \in \mathcal{I}} U \in \mathcal{O}$.

These are the three properties which define a *topology*. Therefore every metric space can be taken to be a topological space. On the other hand, many topological spaces are not metric spaces.

So all this leads to further speculations in the realms of pure mathematics. In fact topology is one of the main branches of modern mathematics.

2.23 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.[9]

[9]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.

2.23.1 Partial derivatives commute if they are continuous

Theorem 2.52. *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 $\widetilde{\mu}$ between 0 and h, such that

$$\frac{k(h)-k(0)}{h} = k'(\widetilde{\mu}) = \partial_2 f(h,\widetilde{\mu}) - \partial_2 f(0,\widetilde{\mu}).$$

Arguing as before, we obtain a $\widetilde{\xi}$ between 0 and h with

$$F(h) = k(h) - k(0) = k'(\widetilde{\xi})h = (\partial_2 f(h,\widetilde{\mu}) - \partial_2 f(0,\widetilde{\mu}))h = \partial_1\partial_2 f(\widetilde{\xi},\widetilde{\mu}) \cdot h^2.$$

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 $\mathbf{x} \in G$.

2.24 Total derivatives

Let $U \subset \mathbb{R}^n$ be open, and let $f : U \to \mathbb{R}^m$ be a function. That is to say, for each $\mathbf{x} \in U$, $f(\mathbf{x}) \in \mathbb{R}^m$. Therefore we can write $f(\mathbf{x}) = (f_1(\mathbf{x}),\dots,f_m(\mathbf{x}))$, where $f_i : U \to \mathbb{R}$, 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 $j = 1,\dots,n$ and $i = 1,\dots,m$. 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 U is an open set, there exists some $\delta > 0$, such that $\mathbf{x} + \xi \in U$, for all $\xi \in \mathbb{R}^n$ with $\|\xi\| < \delta$.

Definition. *Let $f : U \to \mathbb{R}^m$ be a function, and take some point $\mathbf{x} \in U$. 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 $\varphi : B_\delta(\mathbf{x}) \to \mathbb{R}^m$ 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 2.53. *Let $U \subset \mathbb{R}^n$ be open, and let $f : U \to \mathbb{R}^m$ be a function. Assume that f is differentiable at the point $\mathbf{x} \in U$, 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

$$\partial_j f_i(\mathbf{x}) = \lim_{\substack{h\to 0\\ h\neq 0}} \frac{f(\mathbf{x} + h\mathbf{e}_j) - f(\mathbf{x})}{h} = \lim_{\substack{h\to 0\\ h\neq 0}} \frac{ha_{ij} + \varphi_i(h\mathbf{e}_j)}{h} = a_{ij}.$$

□

Theorem 2.54. *Again, $f : U \to \mathbb{R}^n$. This time assume that all partial derivatives $\partial_j f_i$ exist, and are continuous at some point $\mathbf{x} \in U$. 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 U. That is, $B_\delta(\mathbf{x}) \subset U$. Let $\xi = (\xi_1, \ldots, \xi_n) \in B_\delta(\mathbf{x})$. 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

$$\begin{aligned} f_i(\mathbf{x} + \xi) - f_i(\mathbf{x}) &= \sum_{k=1}^{n} (f_i(\mathbf{p}_k) - f_i(\mathbf{p}_{k-1})) \\ &= \sum_{k=1}^{n} \partial_k f_i(\mathbf{p}_{k-1} + \theta_k \xi_k \mathbf{e}_k)\xi_k \\ &= \sum_{k=1}^{n} \partial_k f_i(\mathbf{x})\xi_k + \varphi_k(\xi), \end{aligned}$$

where

$$\varphi_k(\xi) = \sum_{k=1}^{n} (\partial_k f_i(\mathbf{p}_{k-1} + \theta_k \xi_k \mathbf{e}_k) - \partial_k f_i(\mathbf{x}))\xi_k.$$

Then the fact that the function $\partial_k f_i$ is continuous at $\mathbf{x}$ means that we must have $\varphi_k(\xi) = o(\|\xi\|)$ for each k. 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

$$\begin{pmatrix} \varphi_1(\xi) \\ \vdots \\ \varphi_n(\xi) \end{pmatrix} = \varphi(\xi) = o(\|\xi\|).$$

□

2.25 Further results involving partial derivatives

There is no time to deal with the many important further results in the theory of higher-dimensional real analysis which you might need to know. The standard ideas are dealt with in any textbook with the title "Analysis II". So I will simply describe some of them without proof here.

2.25.1 The chain rule in higher dimensions

Let $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$ be open subsets, and let $g : U \to \mathbb{R}^m$ and $f : V \to \mathbb{R}^k$ be functions such that $g(U) \subset V$. Therefore, we can consider the combined function $f \circ g : U \to \mathbb{R}^k$, with $(f \circ g)(x) = f(g(x))$ for all $x \in U$. Now let x_0 be some point in U, and assume that g is totally differentiable at x, and furthermore, f is totally differentiable at $g(x)$. Thus the differential of g at x is the $m \times n$ matrix $Dg(x)$, and the differential of f at $g(x)$ is the $k \times m$ matrix $Df(g(x))$.

Then the chain rule says that also the function $f \circ g : U \to \mathbb{R}^k$ is totally differentiable at x, and the differential is the $k \times n$ matrix

$$Df(g(x)) \cdot Dg(x),$$

obtained from $Dg(x)$ and $Df(g(x))$ using matrix multiplication.

2.25.2 The directional derivative

This is a simple consequence of the chain rule. Let $U \subset \mathbb{R}^n$ be an open subset, and let $f : U \to \mathbb{R}$ 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 $\mathbf{x} \in U$ 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 2.55. $D_{\mathbf{v}} f(\mathbf{x}) = \langle \mathbf{v}, \operatorname{grad} f(\mathbf{x}) \rangle$. *That is, it is the scalar product of* $\mathbf{v}$ *with* $\operatorname{grad} f(x)$.

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 $\mathbf{y} \in U$. 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

$$\begin{aligned} D_{\mathbf{v}} f(\mathbf{x}) &= Df(g(0))Dg(0) \\ &= \Big(\partial_1 f(g(0)) \cdots \partial_n f(g(0))\Big) \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \\ &= \langle \mathbf{v}, \text{grad} f(\mathbf{x}) \rangle. \end{aligned}$$

□

2.25.3 The transformation formula for higher dimensional integrals

Let $G \subset \mathbb{R}^n$ be some bounded subset, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function which vanishes outside G. That is, $f(\mathbf{x}) = 0$ if $x \notin G$. Then, as we have already seen, it makes sense to define the integral

$$\int_{\mathbb{R}^n} f(\mathbf{x}) d^n \mathbf{x}$$

to be the integral of f, taken over a sufficiently large rectangular region containing G.

What happens if we now take some linear mapping $\phi : \mathbb{R}^n \to \mathbb{R}^n$, whose matrix is the $n \times n$ matrix A with respect to the canonical basis of $\mathbb{R}^n$? Then we can consider the function $f \circ \phi : \mathbb{R}^n \to \mathbb{R}$. Here, an arbitrary point $\mathbf{x} \in \mathbb{R}^n$ has the value

$$(f \circ \phi)(\mathbf{x}) = f(A\mathbf{x}).$$

If we now integrate the function $f \circ \phi$, rather than the original f, it is clear that the unit of measure (the small rectangular regions which we use for defining the integral in terms of step functions) will have changed in volume under the linear mapping. What is the change in volume? As we have seen in linear algebra, this change of volume is given by the absolute value of the determinant of the matrix A. This gives us the relation

$$\int_{\mathbb{R}^n} f(A\mathbf{x}) \big| det(A) \big| d^n \mathbf{x} = \int_{\mathbb{R}^n} f(\mathbf{y}) d^n \mathbf{y},$$

for linear mappings $\phi : \mathbb{R}^n \to \mathbb{R}^n$.

The more general transformation formula — which is the natural generalization of the substitution rule for simple 1-dimensional integrals — is then the following.

Let U and $V \subset \mathbb{R}^n$ be open subsets, and let $\phi : U \to V$ be a bijection, such that both ϕ and also ϕ^{-1} are continuously differentiable. Then, for all continuous functions f which vanish outside V, we have

$$\int_U f(\phi(\mathbf{x})) \big| det(D\phi(\mathbf{x})) \big| d^n \mathbf{x} = \int_V f(\mathbf{y}) d^n \mathbf{y}.$$

2.26 Uniformly convergent sequences of functions

Let us return to the simple situation of 1-dimensional functions. And we consider sequences of such functions. (All of this can be easily generalized to higher dimensions as well.)

Definition. *Let $[a,b] \subset \mathbb{R}$, and consider a sequence of functions $f_n : [a,b] \to \mathbb{R}$, where $n \in \mathbb{N}$. The sequence is called* pointwise convergent *if there exists some function $f : [a,b] \to \mathbb{R}$, such that for each $x \in [a,b]$, we have $\lim_{n\to\infty} f_n(x) = f(x)$.*

On the other hand, the sequence is called uniformly convergent *if there is a function $f : [a,b] \to \mathbb{R}$ such that for all $\epsilon > 0$, a number $N_\epsilon \in \mathbb{N}$ exists, with*

$$\big|f_n(x) - f(x)\big| < \epsilon,$$

for all $x \in [a,b]$ and for all $n \geq N_\epsilon$.

Theorem 2.56. *Let $f_n : [a,b] \to \mathbb{R}$ be a uniformly convergent sequence of continuous functions. Then $\lim_{n\to\infty} f_n$ is also continuous.*

Proof. Let $f_n \longrightarrow f$. Now choose any $x \in [a,b]$ and any $\epsilon > 0$. The problem is to show that there exists some $\delta > 0$, such that for all $y \in [a,b]$ with $|y - x| < \delta$, we must have $|f(y) - f(x)| < \epsilon$.

But since the sequence f_n is uniformly convergent, there exists some $N \in \mathbb{N}$ with

$$|f_n(z) - f(z)| < \frac{\epsilon}{3},$$

for all $n \geq N$. In particular, f_N is a continuous function, and so there exists some $\delta > 0$ such that for all $y \in [a,b]$ with $|y - x| < \delta$, we must have

$$|f_N(y) - f_N(x)| < \frac{\epsilon}{3}.$$

But then, for all such y with $|y - x| < \delta$, we have

$$|f(y) - f(x)| \leq |f(y) - f_N(y)| + |f_N(y) - f_N(x)| + |f_N(x) - f(x)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

□

When thinking about functions in an abstract way, it is often useful to remember that they form a vector space. Let $f : [a,b] \to \mathbb{R}$. Then we can define the *supremum norm* as follows

$$\|f\|_{Sup} = Sup\{|f(x)| : a \leq x \leq b\}.$$

(Of course, if the function is not bounded, we must admit the possibility that $\|f\|_{Sup} = \infty$, so it is — strictly speaking — not a norm.)

But given a uniformly convergent sequence of functions f_n, converging to the function $f : [a,b] \to \mathbb{R}$, we must have $\lim_{n\to\infty} \|f - f_n\|_{Sup} = 0$.

Theorem 2.57. *Let $f_n : [a,b] \to \mathbb{R}$ be a uniformly convergent sequence of continuous functions, converging to the function $f : [a,b] \to \mathbb{R}$. Then we have*

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

Proof. Since f is continuous, the integral on the left must exist. For each n, we have

$$\begin{aligned}\left|\int_a^b f(x)dx - \int_a^b f_n(x)dx\right| &\leq \int_a^b \Big|f(x) - f_n(x)\Big|dx \\ &\leq \int_a^b \|f - f_n\|_{Sup}dx \\ &= (b-a)\|f - f_n\|_{Sup} \underset{n\to\infty}{\longrightarrow} 0.\end{aligned}$$

□

Remark. *It is not really necessary for the functions f_n to be continuous in this theorem. It is sufficient to assume that they are integrable.*

On the other hand, the condition of uniform convergence is very necessary. For example the sequence of functions $f_n : [0,1] \to \mathbb{R}$ with

$$f_n(x) = \begin{cases} n^2 x, & 0 \leq x \leq 1/n, \\ n - n^2(x - 1/n), & 1/n \leq x \leq 2/n, \\ 0, & \textit{otherwise,} \end{cases}$$

is pointwise convergent to the constant function whose value is zero everywhere, yet the sequence of integrals does not converge to zero.

2.27 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?

2.27.1 Separation of variables

In fact 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 2.58. *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 $G'(x) = \frac{1}{g(x)} \neq 0$, for all $x \in I'$, 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 since both the functions F and H are given, then it follows that also $\varphi(x)$ is given, if we put in an arbitrary value for x.

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 $G(y_0) = 0$ as well. Thus

$$\varphi(x_0) = H(0) = H(G(y_0)) = y_0.$$

□

2.27.2 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

$$\varphi(x) = H(F(x)) = y_0 \exp\left(\int_{x_0}^{x} t\,dt\right) = y_0 \exp\left(\frac{x^2 - x_0^2}{2}\right).$$

2.27.3 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.

2.27.4 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 $\varphi(x_0) = 0$. 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

$$\psi(x) = \zeta(x)\varphi(x) = \exp\left(\int_{x_0}^{x} a(t)dt\right) \cdot \left(\int_{x_0}^{x} \frac{b(t)}{\exp\left(\int_{x_0}^{x} a(s)ds\right)}dt + y_0\right),$$

which satisfies the initial value $\psi(x_0) = y_0$.

2.28 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? Could it be that all of these elaborate equations just describe nonsense, based on a false assumption? To answer these questions, we need to give some thought to the general theory of differential equations.

2.28.1 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 $(x, \mathbf{y}_0) \in G$, then a solution to the differential equation $y' = f(x, y)$, 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 the differential equation. That is,

$$\varphi'(x) = \mathbf{f}(x, \varphi(x)),$$

for all $x \in I$.

This is the usual framework for a discussion of the theorem of Picard-Lindelöf. But for simplicity, I will simply consider a single equation; thus we have the open subset $G \subset \mathbb{R}^2$, and we are looking for solutions $\varphi : I \to \mathbb{R}$ for an equation of the form $y' = f(x, y)$, where $f : G \to \mathbb{R}$ is a continuous function.

2.28.2 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 $(x, \mathbf{y})$, $(x, \widetilde{\mathbf{y}}) \in G$, we have $\|\mathbf{f}(x, \mathbf{y}) - \mathbf{f}(x, \widetilde{\mathbf{y}})\| \leq \|\mathbf{y} - \widetilde{\mathbf{y}}\| L$.*

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.

However, for simplicity in the discussion here, I will 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 $G \subset \mathbb{R}^2$.

2.28.3 Uniqueness of solutions

Theorem 2.59. *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 $L \geq 0$. 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 $\varphi'(x) = f(x, y)$. 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 (2.25) 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 $\psi(x)$. □

2.28.4 The Banach fixed point theorem

Let $\mathbf{V}$ be a vector space over the field of real numbers $\mathbb{R}$. Assume that $\mathbf{V}$ is a normed vector space; that is, we have a norm function $\|\cdot\| : \mathbf{V} \to \mathbb{R}$. Assume furthermore that $\mathbf{V}$ is *complete* with respect to this norm; that is, every Cauchy sequence converges. For such a vector space, Banach's fixed point theorem is the following.

Theorem 2.60. *Let $\mathbf{V}$ be a complete, real, normed vector space, with norm $\|\cdot\|$. Let $f : \mathbf{V} \to \mathbf{V}$ be a mapping (not necessarily a linear mapping), with the property that there exists some constant $0 \leq K < 1$ such that $\|f(\mathbf{v}) - f(\mathbf{w})\| \leq K \cdot \|\mathbf{v} - \mathbf{w}\|$, for all $\mathbf{v}$, $\mathbf{w} \in \mathbf{V}$. Then there exists a unique fixed point $\mathbf{v}_* \in \mathbf{V}$, such that $f(\mathbf{v}_*) = \mathbf{v}_*$.*

Proof. Choose any vector $\mathbf{v} \in \mathbf{V}$ and call it $\mathbf{v}_0$. Then for each $n \in \mathbb{N}$, let $\mathbf{v}_n = f(\mathbf{v}_{n-1})$. We have

$$\|\mathbf{v}_{n+1} - \mathbf{v}_n\| = \|f(\mathbf{v}_n) - f(\mathbf{v}_{n-1})\| \leq K\|\mathbf{v}_n - \mathbf{v}_{n-1}\| \leq \cdots \leq K^n\|\mathbf{v}_1 - \mathbf{v}_0\| = K^n\|f(\mathbf{v}) - \mathbf{v}\|,$$

for each n. Therefore, given any two numbers m, $n \in \mathbb{N}$, say with $m \leq n$, we have[10]

$$\begin{aligned}
\|\mathbf{v}_n - \mathbf{v}_m\| &= \left\| \sum_{k=m}^{n-1} \mathbf{v}_{k+1} - \mathbf{v}_k \right\| \\
&\leq \sum_{k=m}^{n-1} \|\mathbf{v}_{k+1} - \mathbf{v}_k\| \\
&\leq \sum_{k=m}^{n-1} K^k \|f(\mathbf{v}) - \mathbf{v}\| \\
&= \|f(\mathbf{v}) - \mathbf{v}\| \cdot \left(\sum_{k=m}^{n-1} K^k \right) \\
&= K^m \cdot \|f(\mathbf{v}) - \mathbf{v}\| \cdot \left(\sum_{k=0}^{n-m-1} K^k \right) \\
&< K^m \cdot \|f(\mathbf{v}) - \mathbf{v}\| \cdot \left(\sum_{k=0}^{\infty} K^k \right) \\
&= K^m \cdot \left(\|f(\mathbf{v}) - \mathbf{v}\| \cdot \frac{1}{1-K} \right)
\end{aligned}$$

[10] We assume $K > 0$ here. If $K = 0$, the theorem is trivially true.

Therefore $(\mathbf{v}_n)_{n\in\mathbb{N}}$ must be a Cauchy sequence, and since $\mathbf{V}$ was assumed to be complete, the sequence must converge to some vector $\mathbf{v}_* \in \mathbf{V}$.

The fact that $f(\mathbf{v}_*) = \mathbf{v}_*$ follows, since f must be continuous (with respect to the norm metric), and so we have

$$\lim_{n\to\infty} f(\mathbf{v}_n) = f(\mathbf{v}_*).$$

The fixed point $\mathbf{v}_*$ is unique since, if we had some other fixed point $\mathbf{w}$ with $f(\mathbf{w}) = \mathbf{w}$, then we would have

$$\|\mathbf{v}_* - \mathbf{w}\| = \|f(\mathbf{v}_*) - f(\mathbf{w})\| \leq K \cdot \|\mathbf{v}_* - \mathbf{w}\|,$$

with $0 < K < 1$. But this must imply that $\|\mathbf{v}_* - \mathbf{w}\| = 0$, or $\mathbf{v}_* = \mathbf{w}$. □

As far as the theory of differential equations is concerned, the vector space we are interested in is the space of continuous functions $f : I \to \mathbb{R}$, with the supremum norm. But we have seen in theorem 2.56 that this space of continuous functions is, indeed, complete.[11]

2.28.5 Existence of solutions

Theorem 2.61. *Again, $G \subset \mathbb{R}^2$ open; $f : G \to \mathbb{R}$ continuous, satisfying a Lipschitz condition with constant $L \geq 0$. 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 $M \geq 0$, such that $|f(x, y)| \leq M$, for all $(x, y) \in S_{(x_0,y_0)}(\delta)$. (This was our exercise 9.3(a).) 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.[12] 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 $|x - x_0| < \epsilon \leq \delta$.

[11] The Banach fixed point theorem is more naturally formulated within the theory of metric spaces. Given a mapping $f : X \to X$, where X is a metric space with metric d, then there is a unique fixed point of the mapping if there exists a constant $0 \leq K < 1$ such that $d(f(x), f(y)) \leq Kd(x, y)$, for all $x, y \in X$.

[12] This procedure only gives us the values of $\varphi_n(x)$, for $x \geq x_0$. But again, it is a simple matter of integrating from x up to x_0, rather than from x_0 up to x, in order to obtain the values of $\varphi_n(x)$, for $x < x_0$.

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, using the argument in the proof of theorem 2.60, we see that the sequence of continuous functions φ_n 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 2.57, 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 $\mathbf{f}(t, \varphi_{n-1}(t))$ 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.

2.28.6 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

$$y' = f\left(\frac{y}{x}\right).$$

Obviously we are looking for a solution $\varphi : I \to \mathbb{R}$ with an interval $I \subset \mathbb{R}$, such that $0 \notin I$. Given this, then we have:

Theorem 2.62. *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

$$\begin{aligned}\varphi'(x) &= \psi'(x) \cdot x + \psi(x) \\ &= \psi(x) + \frac{(f(\psi(x)) - \psi(x)) \cdot x}{x} \\ &= f(\psi(x)) \\ &= f\left(\frac{\varphi(x)}{x}\right).\end{aligned}$$

□

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 $y = x \cdot z$.

2.29 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 $\varphi(x) = \cos(x)$.

2.30 Numerical methods for solving ordinary differential equations

2.30.1 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

$$\varphi(x) = \frac{1}{2}x^2.$$

What does Euler's method make of this problem if we take $\Delta x = 1$? We obtain the sequence of points

$$\begin{array}{c} (0,\ 0) \\ (1,\ 0) \\ (2,\ 1) \\ (3,\ 3) \\ (4,\ 6) \\ etc. \end{array}$$

But the correct solution $\varphi(x) = \frac{1}{2}x^2$ goes through the points

$$\begin{array}{c} (0,\ 0) \\ (1,\ \frac{1}{2}) \\ (2,\ 2) \\ (3,\ 4\frac{1}{2}) \\ (4,\ 8) \\ etc. \end{array}$$

So we see that Euler's method is not particularly good in this case.

2.30.2 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

$$\begin{array}{c} (0,\ 0) \\ (1,\ \frac{1}{2}) \\ (2,\ 2) \\ (3,\ 4\frac{1}{2}) \\ (4,\ 8) \\ etc. \end{array}$$

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 not exact.

A more complicated 4-step iteration, giving the standard Runga-Kutta method, together with the motivation behind these ideas, was discussed in the lecture.

2.31 The variational calculus: a quick 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. After all, if you think about it then you will soon realize that *any question at all* could be formulated in such a framework! 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.

For example, what is the shape of a telephone wire which hangs steadily, in equilibrium, under gravity between two posts? Assume that gravity is given by the constant g, and that the wire has the uniform weight σ per unit length (for example one kilogram per meter). We can imagine that the wire has a length L, and the two posts are located at the points a and b, measured along the real line $\mathbb{R}$. So we have $|b - a| < L$. Assume that the two ends of the wire are both at the height h, and that the height at points x between a and b is given by the function $\varphi : [a, b] \to \mathbb{R}$. Then, as explained in the lecture, the potential energy of the wire is

$$F(\varphi) = g\sigma \int_a^b \varphi(x)\sqrt{1 + (\varphi'(x))^2}dx.$$

The shape of the wire is therefore given by the function φ_0 which describes a curve of total length L, such that $F(\varphi_0)$ is a minimum.

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.

Being slightly more specific, 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 $\widetilde{\varphi} : I \to \mathbb{R}$, and compute the values of $F(\widetilde{\varphi})$, checking to see if they are always greater than, or equal to $F(\varphi)$. Writing

$$\psi = \widetilde{\varphi} - \varphi,$$

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

$$\Gamma(s) = F(\varphi + s\psi) = \int_a^b f\Big(x, \varphi(x) + s\psi(x), \varphi'(x) + s\psi'(x)\Big)dx.$$

Given that things are sufficiently differentiable, we then have that Γ is differentiable at zero; and if φ is a solution to our variational problem then it must be that

$$\Gamma'(0) = 0.$$

We then have

$$\begin{aligned}
\Gamma'(0) &= \frac{d}{ds}\Big|_{s=0} \int_a^b f\Big(x, \varphi(x) + s\psi(x), \varphi'(x) + s\psi'(x)\Big)dx \\
&= \int_a^b \frac{d}{ds}\Big|_{s=0} f\Big(x, \varphi(x) + s\psi(x), \varphi'(x) + s\psi'(x)\Big)dx \\
&= \int_a^b \left(\psi(x)\frac{\partial}{\partial y} f(x, \varphi(x), \varphi'(x)) + \psi'(x)\frac{\partial}{\partial y'} f(x, \varphi(x), \varphi'(x))\right) dx \\
&= \int_a^b \left(\psi(x)\frac{\partial}{\partial y} f(x, \varphi(x), \varphi'(x))\right) dt + \int_a^b \left(\psi'(x)\frac{\partial}{\partial y'} f(x, \varphi(x), \varphi'(x))\right) dx \\
&= \int_a^b \left(\psi(x)\frac{\partial}{\partial y} f(x, \varphi(x), \varphi'(x))\right) dt - \int_a^b \left(\psi(x)\frac{d}{dx}\frac{\partial}{\partial y'} f(x, \varphi(x), \varphi'(x))\right) dx \\
&= \int_a^b \psi(x)\left(\frac{\partial}{\partial y} f(x, \varphi(x), \varphi'(x)) - \frac{d}{dx}\frac{\partial}{\partial y'} f(x, \varphi(x), \varphi'(x))\right) dx \\
&= 0
\end{aligned}$$

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 we have uniform convergence of the fraction

$$\frac{g(x,s+h) - g(x,s)}{h}$$

  to

$$\frac{\partial}{\partial s} g(x,s)$$

  as $h \to 0$. Therefore the second equation is true by theorem 2.57.

- 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, as explained in the lecture.
- 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, as explained in the lecture.)

Putting this into the more usual form for writing differential equations, we have

$$\frac{\partial}{\partial y} f(x, y, y') = \frac{d}{dx} \frac{\partial}{\partial y'} f(x, y, y').$$

How do we evaluate the complicated looking expression

$$\frac{d}{dx} \frac{\partial}{\partial y'} f(x, \varphi(x), \varphi'(x)) \, ?$$

For this you should just 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

$$\frac{\partial}{\partial y} f(x, y, y') = \frac{\partial^2}{\partial x \partial y'} f(x, y, y') + y' \frac{\partial^2}{\partial y \partial y'} f(x, y, y') + y'' \frac{\partial^2}{\partial y'^2} f(x, y, y')$$

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

$$\frac{\partial}{\partial y} f(y, y') - y' \frac{\partial^2}{\partial y \partial y'} f(y, y') - y'' \frac{\partial^2}{\partial y'^2} f(y, y') = 0.$$

But we have

$$\frac{d}{dx} \Big(f(y, y') - y' \frac{\partial}{\partial y'} f(y, y') \Big) = y' \Big(\frac{\partial}{\partial y} f(y, y') - y' \frac{\partial^2}{\partial y \partial y'} f(y, y') - y'' \frac{\partial^2}{\partial y'^2} f(y, y') \Big) = 0.$$

Therefore

$$f(y, y') - y' \frac{\partial}{\partial y'} f(y, y') = k,$$

for some constant $k \in \mathbb{R}$.

An example

Let us return to the problem of calculating the shape of a freely hanging telephone wire. Forgetting the constants g and σ which don't affect the outcome, we have the variational problem

$$F(\varphi) = \int_a^b \varphi(x)\sqrt{1 + (\varphi'(x))^2}dx.$$

There is the further complication that the length of the wire is fixed. In fact the length is L. This gives us the further condition

$$\int_a^b \sqrt{1 + (\varphi'(x))^2}dx = L.$$

So in order to solve this problem, we need to use the method of *Lagrange multipliers* (discussed in the lecture). That means that our variational problem becomes

$$\widetilde{F}(\varphi) = \int_a^b (\varphi(x) + \lambda)\sqrt{1 + (\varphi'(x))^2}dx,$$

for some constant $\lambda \in \mathbb{R}$.

So here, the Euler-Lagrange equation is

$$f(y, y') - y' \frac{\partial}{\partial y'} f(y, y') = k,$$

with

$$f(y, y') = (y + \lambda)\sqrt{1 + y'^2}.$$

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.

Chapter 3

Linear Algebra

The subject of linear algebra is concerned — at least at the beginning — with geometry. We think of the normal geometry of physical space as being $\mathbb{R}^3$, that is, the Cartesian product of $\mathbb{R}$ with itself 3-times. In general, the notation X^n is used to represent the Cartesian product of a set X with itself n times. Thus

$$X^n = \underbrace{X \times X \times \cdots \times X}_{n \text{ times}} = \{(x_1, \ldots, x_n) : x_i \in X_i, \forall i\}.$$

So if physical space is $\mathbb{R}^3$, then an arbitrary point of space can be described by specifying a triple of real numbers (x, y, z). These are the coordinates of the point, measured with respect to a given coordinate system, consisting of the x-axis, the y-axis, and the z-axis.

From the viewpoint of linear algebra, we imagine that the point (x, y, z) is really a vector, which we can think about as being like an arrow, with its tail in the point $(0, 0, 0)$, that is, the middle point of the coordinate system, and the head of the arrow is at our point (x, y, z).

In order distinguish vectors from numbers, some books use the following notation. Let us say that a vector in $\mathbb{R}^3$ is given by the coordinates (v_x, v_y, v_z). Then one writes

$$\vec{v} = (v_x, v_y, v_z).$$

Thus, a vector is distinguished by the little arrow, floating above it. However, in these notes I will distinguish vectors from other possible objects by writing vectors in **bold face type**.

So let $\mathbf{p} = (p_x, p_y, p_z)$ and $\mathbf{q} = (q_x, q_y, q_z)$ be two points — or vectors — in our *vector space* $\mathbb{R}^3$. Then we can add the two vectors together to obtain a new vector.

$$\mathbf{p} + \mathbf{q} = (p_x + q_x, p_y + q_y, p_z + q_z).$$

In the vector space $\mathbb{R}^3$, we also have *scalar multiplication*. Given the vector $\mathbf{v} = (v_x, v_y, v_z)$, and given some real number a, then the *scalar product* of a with $\mathbf{v}$ is the vector

$$a\mathbf{v} = (av_x, av_y, av_z).$$

So the scalar product is a new vector which has the same *direction* as the original vector v, but its *length* is changed by the scalar a. For example, if $a = 2$, then the length is increased by the factor 2. On the other hand, if $a = 1/2$, then the length of the vector is halved.

3.1 Basic definitions

Definition. *A vector space* $\mathbf{V}$ *over a field* F *is an Abelian group under the operation of vector addition,*

$$+ : \mathbf{V} \times \mathbf{V} \to \mathbf{V}.$$

Furthermore, there is an operation of scalar multiplication, which is a mapping

$$\cdot : F \times \mathbf{V} \to \mathbf{V},$$

satisfying the following properties. For arbitrary $a, b \in F$ *and* $\mathbf{v}, \mathbf{w} \in V$*, we have*

- $a \cdot (\mathbf{v} + \mathbf{w}) = a \cdot \mathbf{v} + a \cdot \mathbf{w}$
- $(a + b) \cdot \mathbf{v} = a \cdot \mathbf{v} + b \cdot \mathbf{v}$
- $(ab) \cdot \mathbf{v} = a \cdot (b \cdot \mathbf{v})$
- $1 \cdot \mathbf{v} = \mathbf{v}$

Note here that the zero and the one of the field F *are simply called "0" and "1", as usual. Also the zero of the vector space, with respect to vector addition, is called* $\mathbf{0}$*. As is the case with multiplication in fields, in general we will simply write* $a\mathbf{v}$*, rather than* $a \cdot \mathbf{v}$*, to denote scalar multiplication.*

Examples

- Given any field F, then we can say that F is a vector space over itself. The vectors are just the elements of F. Vector addition is the addition in the field. Scalar multiplication is multiplication in the field.
- In $\mathbb{R}^n$, the vectors can be represented by n-tuples of real numbers $(x_1, \ldots, x_n)$, where $x_i \in \mathbb{R}$, for all $i = 1, \ldots, n$. Given two elements

  $$(x_1, \cdots, x_n) \quad \text{and} \quad (y_1, \ldots, y_n)$$

  in $\mathbb{R}^n$, then the vector sum is simply the new vector

  $$(x_1 + y_1, \cdots, x_n + y_n).$$

  Scalar multiplication is

  $$a \cdot (x_1, \cdots, x_n) = (a \cdot x_1, \cdots, a \cdot x_n).$$

  It is a trivial matter to verify that $\mathbb{R}^n$, with these operations, is a vector space over $\mathbb{R}$.
- Let $F(X, \mathbb{R})$ be the set of all functions $f : X \to \mathbb{R}$, from some set X to the set of real numbers $\mathbb{R}$. This is a vector space with vector addition

  $$(f + g)(x) = f(x) + g(x),$$

  for all $x \in X$, defining the new function $(f + g) \in F(X, \mathbb{R})$, for all $f, g \in F(X, \mathbb{R})$.

  Scalar multiplication is given by

  $$(a \cdot f)(x) = a \cdot f(x)$$

  for all $x \in X$.

3.2 Subspaces

Let $\mathbf{V}$ be a vector space over a field F and let $\mathbf{W} \subset \mathbf{V}$ be some subset. If $\mathbf{W}$ is itself a vector space over F, considered using the addition and scalar multiplication in $\mathbf{V}$, then we say that $\mathbf{W}$ is a *subspace* of $\mathbf{V}$. Analogously, a subset H of a group G, which is itself a group using the multiplication operation from G, is called a *subgroup* of G. Subfields are similarly defined.

Theorem 3.1. *Let $\mathbf{W} \subset \mathbf{V}$ be a subset of a vector space over the field F. Then*

$$\mathbf{W} \textit{ is a subspace of } \mathbf{V} \Leftrightarrow a \cdot \mathbf{v} + b \cdot \mathbf{w} \in \mathbf{W},$$

for all $\mathbf{v}, \mathbf{w} \in \mathbf{W}$ and $a, b \in F$.

Proof. The direction '$\Rightarrow$' is trivial.

For '$\Leftarrow$', begin by observing that $1 \cdot \mathbf{v} + 1 \cdot \mathbf{w} = \mathbf{v} + \mathbf{w} \in \mathbf{W}$, and $a \cdot \mathbf{v} + 0 \cdot \mathbf{w} = a \cdot \mathbf{v} \in \mathbf{W}$, for all $\mathbf{v}, \mathbf{w} \in \mathbf{W}$ and $a \in F$. Thus $\mathbf{W}$ is closed under vector addition and scalar multiplication.

Is $\mathbf{W}$ a group with respect to vector addition? We have $0 \cdot \mathbf{v} = \mathbf{0} \in \mathbf{W}$, for $\mathbf{v} \in \mathbf{W}$; therefore the neutral element $\mathbf{0}$ is contained in $\mathbf{W}$. For an arbitrary $\mathbf{v} \in \mathbf{W}$ we have

$$\begin{aligned} \mathbf{v} + (-1) \cdot \mathbf{v} &= 1 \cdot \mathbf{v} + (-1) \cdot \mathbf{v} \\ &= (1 + (-1)) \cdot \mathbf{v} \\ &= 0 \cdot \mathbf{v} \\ &= \mathbf{0}. \end{aligned}$$

Therefore $(-1) \cdot \mathbf{v}$ is the inverse element to $\mathbf{v}$ under addition, and so we can simply write $(-1) \cdot \mathbf{v} = -\mathbf{v}$.

The other axioms for a vector space can be easily checked. □

The method of this proof also shows that we have similar conditions for subsets of groups or fields to be subgroups, or subfields, respectively.

Theorem 3.2. *Let $H \subset G$ be a (non-empty) subset of the group G. Then H is a subgroup of $G \Leftrightarrow ab^{-1} \in H$, for all $a, b \in H$.*

Proof. The direction '$\Rightarrow$' is trivial. As for '$\Leftarrow$', let $a \in H$. Then $aa^{-1} = e \in H$. Thus the neutral element of the group multiplication is contained in H. Also $ea^{-1} = a^{-1} \in H$. Furthermore, for all $a, b \in H$, we have $a(b^{-1})^{-1} = ab \in H$. Thus H is closed under multiplication. The fact that the multiplication is associative ($a(bc) = (ab)c$, for all a, b and $c \in H$) follows since G itself is a group; thus the multiplication throughout G is associative. □

Theorem 3.3. *Let $\mathbf{U}, \mathbf{W} \subset \mathbf{V}$ be subspaces of the vector space $\mathbf{V}$ over the field F. Then $\mathbf{U} \cap \mathbf{W}$ is also a subspace.*

Proof. Let $\mathbf{v}, \mathbf{w} \in \mathbf{U} \cap \mathbf{W}$ be arbitrary vectors in the intersection, and let $a, b \in F$ be arbitrary elements of the field F. Then, since $\mathbf{U}$ is a subspace of $\mathbf{V}$, we have $a \cdot \mathbf{v} + b \cdot \mathbf{w} \in \mathbf{U}$. This follows from theorem 3.1. Similarly $a \cdot \mathbf{v} + b \cdot \mathbf{w} \in \mathbf{W}$. Thus it is in the intersection, and so theorem 3.1 shows that $\mathbf{U} \cap \mathbf{W}$ is a subspace. □

3.3 Linear independence and dimension

Definition. *Let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbf{V}$ be finitely many vectors in the vector space $\mathbf{V}$ over the field F. We say that the vectors are* linearly dependent *if there exists an equation of the form*

$$a_1 \cdot \mathbf{v}_1 + \cdots + a_n \cdot \mathbf{v}_n = \mathbf{0},$$

such that not all $a_i \in F$ are simply zero. If no such non-trivial equation exists, then the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbf{V}$ is said to be linearly independent.

This definition is undoubtedly the most important idea that there is in the theory of linear algebra!

Examples

- In $\mathbb{R}^2$ let $\mathbf{v}_1 = (1, 0)$, $\mathbf{v}_2 = (0, 1)$ and $\mathbf{v}_3 = (1, 1)$. Then the set $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is linearly *dependent*, since we have
$$\mathbf{v}_1 + \mathbf{v}_2 - \mathbf{v}_1 = \mathbf{0}.$$
On the other hand, the set $\{\mathbf{v}_1, \mathbf{v}_2\}$ is linearly *in*dependent.

- In $C_0([0, 1], \mathbb{R})$, let $f_1 : [0, 1] \to \mathbb{R}$ be given by $f_1(x) = 1$ for all $x \in [0, 1]$. Similarly, let f_2 be given by $f_2(x) = x$, and f_3 is $f_3(x) = 1 - x$. Then the set $\{f_1, f_2, f_3\}$ is linearly dependent.

Now take some vector space $\mathbf{V}$ over a field F, and let $S \subset \mathbf{V}$ be some subset of $\mathbf{V}$. (The set S can be finite or infinite here, although we will usually be dealing with finite sets.) Let $\mathbf{v}_1, \ldots, \mathbf{v}_n \subset S$ be some finite collection of vectors in S, and let $a_1, \ldots, a_n \in F$ be some arbitrary collection of elements of the field. Then the sum

$$a_1 \cdot \mathbf{v}_1 + \cdots + a_n \cdot \mathbf{v}_n$$

is a *linear combination* of the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ in S. The set of all possible linear combinations of vectors in S is denoted by $span(S)$, and it is called the linear span of S. One also writes $[S]$. S is the *generating set* of $[S]$. Therefore if $[S] = \mathbf{V}$, then we say that S is a generating set for V. If S is finite, and it generates $\mathbf{V}$, then we say that the vector space $\mathbf{V}$ is *finitely generated.*

Theorem 3.4. *Given $S \subset \mathbf{V}$, then $[S]$ is a subspace of $\mathbf{V}$.*

Proof. A simple consequence of theorem 3.1. □

Examples

- For any $n \in \mathbb{N}$, let
$$\begin{aligned} \mathbf{e}_1 &= (1, 0, \ldots, 0) \\ \mathbf{e}_2 &= (0, 1, \ldots, 0) \\ &\vdots \\ \mathbf{e}_n &= (0, 0, \ldots, 1) \end{aligned}$$
Then $S = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ is a generating set for $\mathbb{R}^n$.

- On the other hand, the vector space $C_0([0,1],\mathbb{R})$ is clearly *not* finitely generated.[1]

So let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbf{V}$ be a *finite* set. From now on in these discussions, we will assume that such sets are finite unless stated otherwise.

Theorem 3.5. *Let $\mathbf{w} = a_1\mathbf{v}_1 + \cdots a_n\mathbf{v}_n$ be some vector in $[S] \subset \mathbf{V}$, where $a_1, \ldots, a_n$ are arbitrarily given elements of the field F. We will say that this representation of* $\mathbf{w}$ *is* unique *if, given some other linear combination, $\mathbf{w} = b_1\mathbf{v}_1 + \cdots b_n\mathbf{v}_n$, then we must have $b_i = a_i$ for all $i = 1, \ldots, n$. Given this, then we have that the set S is linearly independent $\Leftrightarrow$ the representation of all vectors in the span of S as linear combinations of vectors in S is unique.*

Proof. '$\Leftarrow$' We certainly have $0 \cdot \mathbf{v}_1 + \cdots 0 \cdot \mathbf{v}_n = \mathbf{0}$. Since this representation of the zero vector is unique, it follows S is linearly independent.

'$\Rightarrow$' Can it be that S is linearly independent, and yet there exists a vector in the span of S which is not uniquely represented as a linear combination of the vectors in S? Assume that there exist elements $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ of the field F, where $a_j \neq b_j$, for at least one j between 1 and n, such that

$$a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n = b_1\mathbf{v}_1 + \cdots + b_n\mathbf{v}_n.$$

But then

$$(a_1 - b_1)\mathbf{v}_1 + \cdots + \underbrace{(a_j - b_j)}_{\neq 0}\mathbf{v}_i + \cdots + (a_n - b_n)\mathbf{v}_n = \mathbf{0}$$

shows that S cannot be a linearly independent set. □

Definition. *Assume that $S \subset \mathbf{V}$ is a finite, linearly independent subset with $[S] = \mathbf{V}$. Then S is called a* basis *for* $\mathbf{V}$.

Lemma. *Assume that $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbf{V}$ is linearly dependent. Then there exists some $j \in \{1, \ldots, n\}$, and elements $a_i \in F$, for $i \neq j$, such that*

$$\mathbf{v}_j = \sum_{i \neq j} a_i\mathbf{v}_i.$$

Proof. Since S is linearly dependent, there exists some non-trivial linear combination of the elements of S, summing to the zero vector,

$$\sum_{i=1}^{n} b_i\mathbf{v}_i = \mathbf{0},$$

such that $b_j \neq 0$, for at least one of the j. Take such a one. Then

$$b_j\mathbf{v}_j = -\sum_{i \neq j} b_i\mathbf{v}_i$$

and so

$$\mathbf{v}_j = \sum_{i \neq j} \left(-\frac{b_i}{b_j}\right)\mathbf{v}_j.$$

□

[1] In general such function spaces are not finitely generated. However in this lecture, we will mostly be concerned with finitely generated vector spaces.

Corollary. *Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbf{V}$ be linearly dependent, and let $\mathbf{v}_j$ be as in the lemma above. Let $S' = \{\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n\}$ be S, with the element $\mathbf{v}_j$ removed. Then $[S] = [S']$.*

Theorem 3.6. *Assume that the vector space $\mathbf{V}$ is finitely generated. Then there exists a basis for $\mathbf{V}$.*

Proof. Since $\mathbf{V}$ is finitely generated, there exists a finite generating set. Let S be such a finite generating set which has as few elements as possible. If S were linearly dependent, then we could remove some element, as in the lemma, leaving us with a still smaller generating set for $\mathbf{V}$. This is a contradiction. Therefore S must be a basis for $\mathbf{V}$. □

Theorem 3.7. *Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis for the vector space $\mathbf{V}$, and take some arbitrary non-zero vector $\mathbf{w} \in \mathbf{V}$. Then there exists some $j \in \{1, \ldots, n\}$, such that*

$$S' = \{\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}, \mathbf{w}, \mathbf{v}_{j+1}, \ldots, \mathbf{v}_n\}$$

is also a basis of $\mathbf{V}$.

Proof. Writing $\mathbf{w} = a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n$, we see that since $\mathbf{w} \neq \mathbf{0}$, at least one $a_j \neq 0$. Taking that j, we write

$$\mathbf{v}_j = a_j^{-1}\mathbf{w} + \sum_{i \neq j} \left(-\frac{a_i}{a_j}\right) \mathbf{v}_i.$$

We now prove that $[S'] = \mathbf{V}$. For this, let $\mathbf{u} \in \mathbf{V}$ be an arbitrary vector. Since S is a basis for $\mathbf{V}$, there exists a linear combination $\mathbf{u} = b_1\mathbf{v}_1 + \cdots b_n\mathbf{v}_n$. Then we have

$$\begin{aligned}
\mathbf{u} &= b_j\mathbf{v}_j + \sum_{i \neq j} b_i\mathbf{v}_i \\
&= b_j \left(a_j^{-1}\mathbf{w} + \sum_{i \neq j} \left(-\frac{a_i}{a_j}\right) \mathbf{v}_i \right) + \sum_{i \neq j} b_i\mathbf{v}_i \\
&= b_j a_j^{-1}\mathbf{w} + \sum_{i \neq j} \left(b_i - \frac{b_j a_i}{a_j} \right) \mathbf{v}_i
\end{aligned}$$

This shows that $[S'] = \mathbf{V}$.

In order to show that S' is linearly independent, assume that we have

$$\begin{aligned}
\mathbf{0} &= c\mathbf{w} + \sum_{i \neq j} c_i\mathbf{v}_i \\
&= c \left(\sum_{i=1}^{n} a_i\mathbf{v}_i \right) + \sum_{i \neq j} c_i\mathbf{v}_i \\
&= \sum_{i=1}^{n} (ca_i + c_i)\mathbf{v}_i \qquad (\text{with } c_j = 0),
\end{aligned}$$

for some c, and $c_i \in F$. Since the original set S was assumed to be linearly independent, we must have $ca_i + c_i = 0$, for all i. In particular, since $c_j = 0$, we have $ca_j = 0$. But the assumption was that $a_j \neq 0$. Therefore we must conclude that $c = 0$. It follows that also $c_i = 0$, for all $i \neq j$. Therefore, S' must be linearly independent. □

Theorem 3.8 (Steinitz Exchange Theorem)**.** *Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis of $\mathbf{V}$ and let $T = \{\mathbf{w}_1, \ldots, \mathbf{w}_m\} \subset \mathbf{V}$ be some linearly independent set of vectors in $\mathbf{V}$. Then we have $m \leq n$. By possibly re-ordering the elements of S, we may arrange things so that the set*

$$U = \{\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{v}_{m+1}, \ldots, \mathbf{v}_n\}$$

is a basis for $\mathbf{V}$.

Proof. Use induction over the number m. If $m = 0$ then $U = S$ and there is nothing to prove. Therefore assume $m \geq 1$ and furthermore, the theorem is true for the case $m - 1$. So consider the linearly independent set $T' = \{\mathbf{w}_1, \ldots, \mathbf{w}_{m-1}\}$. After an appropriate re-ordering of S, we have $U' = \{\mathbf{w}_1, \ldots, \mathbf{w}_{m-1}, \mathbf{v}_m, \ldots, \mathbf{v}_n\}$ being a basis for $\mathbf{V}$. Note that if we were to have $n < m$, then T' would itself be a basis for $\mathbf{V}$. Thus we could express $\mathbf{w}_m$ as a linear combination of the vectors in T'. That would imply that T was not linearly independent, contradicting our assumption. Therefore, $m \leq n$.

Now since U' is a basis for $\mathbf{V}$, we can express $\mathbf{w}_m$ as a linear combination

$$\mathbf{w}_m = a_1\mathbf{w}_1 + \cdots + a_{m-1}\mathbf{w}_{m-1} + a_m\mathbf{v}_m + \cdots + a_n\mathbf{v}_n.$$

If we had all the coefficients of the vectors from S being zero, namely

$$a_m = a_{m+1} = \cdots = a_n = 0,$$

then we would have $\mathbf{w}_m$ being expressed as a linear combination of the other vectors in T. Therefore T would be linearly dependent, which is not true. Thus one of the $a_j \neq 0$, for $j \geq m$. Using theorem 3.7, we may exchange $\mathbf{w}_m$ for the vector $\mathbf{v}_j$ in U', thus giving us the basis U. □

Theorem 3.9 (Extension Theorem)**.** *Assume that the vector space $\mathbf{V}$ is finitely generated and that we have a linearly independent subset $S \subset \mathbf{V}$. Then there exists a basis B of $\mathbf{V}$ with $S \subset B$.*

Proof. If $[S] = \mathbf{V}$ then we simply take $B = S$. Otherwise, start with some given basis $A \subset \mathbf{V}$ and apply theorem 3.8 successively. □

Theorem 3.10. *Let $\mathbf{U}$ be a subspace of the (finitely generated) vector space $\mathbf{V}$. Then $\mathbf{U}$ is also finitely generated, and each possible basis for $\mathbf{U}$ has no more elements than any basis for $\mathbf{V}$.*

Proof. Assume there is a basis B of $\mathbf{V}$ containing n vectors. Then, according to theorem 3.8, there cannot exist more than n linearly independent vectors in $\mathbf{U}$. Therefore $\mathbf{U}$ must be finitely generated, such that any basis for $\mathbf{U}$ has at most n elements. □

Theorem 3.11. *Assume the vector space $\mathbf{V}$ has a basis consisting of n elements. Then every basis of $\mathbf{V}$ also has precisely n elements.*

Proof. This follows directly from theorem 3.10, since any basis generates $\mathbf{V}$, which is a subspace of itself. □

Definition. *The number of vectors in a basis of the vector space $\mathbf{V}$ is called the* dimension *of $\mathbf{V}$, written $dim(\mathbf{V})$.*

Definition. *Let $\mathbf{V}$ be a vector space with subspaces $\mathbf{X}, \mathbf{Y} \subset \mathbf{V}$. The subspace $\mathbf{X} + \mathbf{Y} = [X \cup Y]$ is called the* sum *of $\mathbf{X}$ and $\mathbf{Y}$. If $\mathbf{X} \cap \mathbf{Y} = \{\mathbf{0}\}$, then it is the* direct sum*, written $\mathbf{X} \oplus \mathbf{Y}$.*

Theorem 3.12 (A Dimension Formula)**.** *Let* $\mathbf{V}$ *be a finite dimensional vector space with subspaces* $\mathbf{X}, \mathbf{Y} \subset \mathbf{V}$*. Then we have*

$$dim(\mathbf{X} + \mathbf{Y}) = dim(\mathbf{X}) + dim(\mathbf{Y}) - dim(\mathbf{X} \cap \mathbf{Y}).$$

Corollary. $dim(\mathbf{X} \oplus \mathbf{Y}) = dim(\mathbf{X}) + dim(\mathbf{Y})$.

Proof of Theorem 3.12. Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis of $\mathbf{X} \cap \mathbf{Y}$. According to theorem 3.9, there exist extensions $T = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and $U = \{\mathbf{y}_1, \ldots, \mathbf{y}_r\}$, such that $S \cup T$ is a basis for $\mathbf{X}$ and $S \cup U$ is a basis for $\mathbf{Y}$. We will now show that, in fact, $S \cup T \cup U$ is a basis for $\mathbf{X} + \mathbf{Y}$.

To begin with, it is clear that $\mathbf{X} + \mathbf{Y} = [S \cup T \cup U]$. Is the set $S \cup T \cup U$ linearly independent? Let

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^{n} a_i \mathbf{v}_i + \sum_{j=1}^{m} b_j \mathbf{x}_j + \sum_{k=1}^{r} c_k \mathbf{y}_k \\ &= \mathbf{v} + \mathbf{x} + \mathbf{y}, \quad \text{say.} \end{aligned}$$

Then we have $\mathbf{y} = -\mathbf{v} - \mathbf{x}$. Thus $\mathbf{y} \in \mathbf{X}$. But clearly we also have, $\mathbf{y} \in \mathbf{Y}$. Therefore $\mathbf{y} \in \mathbf{X} \cap \mathbf{Y}$. Thus $\mathbf{y}$ can be expressed as a linear combination of vectors in S alone, and since $S \cup U$ is is a basis for Y, we must have $c_k = 0$ for $k = 1, \ldots, r$. Similarly, looking at the vector $\mathbf{x}$ and applying the same argument, we conclude that all the b_j are zero. But then all the a_i must also be zero since the set S is linearly independent.

Putting this all together, we see that the $dim(\mathbf{X}) = n+m$, $dim(\mathbf{Y}) = n+r$ and $dim(\mathbf{X} \cap \mathbf{Y}) = n$. This gives the dimension formula. □

Theorem 3.13. *Let* $\mathbf{V}$ *be a finite dimensional vector space, and let* $\mathbf{X} \subset \mathbf{V}$ *be a subspace. Then there exists another subspace* $\mathbf{Y} \subset \mathbf{V}$*, such that* $\mathbf{V} = \mathbf{X} \oplus \mathbf{Y}$.

Proof. Take a basis S of $\mathbf{X}$. If $[S] = \mathbf{V}$ then we are finished. Otherwise, use the extension theorem (theorem 3.9) to find a basis B of $\mathbf{V}$, with $S \subset B$. Then $\mathbf{Y} = [B \setminus S]$ satisfies the condition of the theorem. □

3.4 Linear mappings

Definition. *Let* $\mathbf{V}$ *and* $\mathbf{W}$ *be vector spaces, both over the field* F*. Let* $f : \mathbf{V} \to \mathbf{W}$ *be a mapping from the vector space* $\mathbf{V}$ *to the vector space* $\mathbf{W}$*. The mapping* f *is called a linear mapping if*

$$f(a\mathbf{u} + b\mathbf{v}) = af(\mathbf{u}) + bf(\mathbf{v})$$

for all $a, b \in F$ *and all* $\mathbf{u}, \mathbf{v} \in \mathbf{V}$.

By choosing a and b to be either 0 or 1, we immediately see that a linear mapping always has both $f(a\mathbf{v}) = af(\mathbf{v})$ and $f(\mathbf{u} + \mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v})$, for all $a \in F$ and for all $\mathbf{u}$ and $\mathbf{v} \in \mathbf{V}$. Also it is obvious that $f(\mathbf{0}) = \mathbf{0}$ always.

Definition. *Let* $f : \mathbf{V} \to \mathbf{W}$ *be a linear mapping. The* kernel *of the mapping, denoted by* $ker(f)$*, is the set of vectors in* $\mathbf{V}$ *which are mapped by* f *into the zero vector in* $\mathbf{W}$.

Theorem 3.14. *If* $ker(f) = \{\mathbf{0}\}$*, that is, if the zero vector in* $\mathbf{V}$ *is the only vector which is mapped into the zero vector in* $\mathbf{W}$ *under* f*, then* f *is an injection (monomorphism). The converse is of course trivial.*

Proof. That is, we must show that if $\mathbf{u}$ and $\mathbf{v}$ are two vectors in $\mathbf{V}$ with the property that $f(\mathbf{u}) = f(\mathbf{v})$, then we must have $\mathbf{u} = \mathbf{v}$. But

$$f(\mathbf{u}) = f(\mathbf{v}) \quad \Rightarrow \quad \mathbf{0} = f(\mathbf{u}) - f(\mathbf{v}) = f(\mathbf{u} - \mathbf{v}).$$

Thus the vector $\mathbf{u} - \mathbf{v}$ is mapped by f to the zero vector. Therefore we must have $\mathbf{u} - \mathbf{v} = \mathbf{0}$, or $\mathbf{u} = \mathbf{v}$.

Conversely, since $f(\mathbf{0}) = \mathbf{0}$ always holds, and since f is an injection, we must have $ker(f) = \{\mathbf{0}\}$. □

Theorem 3.15. *Let $f : \mathbf{V} \to \mathbf{W}$ be a linear mapping and let $A = \{\mathbf{w}_1, \ldots, \mathbf{w}_m\} \subset \mathbf{W}$ be linearly independent. Assume that m vectors are given in $\mathbf{V}$, so that they form a set $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\} \subset \mathbf{V}$ with $f(\mathbf{v}_i) = \mathbf{w}_i$, for all i. Then the set B is also linearly independent.*

Proof. Let $a_1, \ldots, a_m \in F$ be given such that $a_1\mathbf{v}_1 + \cdots + a_m\mathbf{v}_m = \mathbf{0}$. But then

$$\mathbf{0} = f(\mathbf{0}) = f(a_1\mathbf{v}_1 + \cdots + a_m\mathbf{v}_m) = a_1 f(\mathbf{v}_1) + \cdots + a_m f(\mathbf{v}_m) = a_1\mathbf{w}_1 + \cdots + a_m\mathbf{w}_m.$$

Since A is linearly independent, it follows that all the a_i's must be zero. But that implies that the set B is linearly independent. □

Remark. *If $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\} \subset \mathbf{V}$ is linearly independent, and $f : \mathbf{V} \to \mathbf{W}$ is linear, still, it does not necessarily follow that $\{f(\mathbf{v}_1), \ldots, f(\mathbf{v}_m)\}$ is linearly independent in $\mathbf{W}$. On the other hand, if f is an* injection*, then $\{f(\mathbf{v}_1), \ldots, f(\mathbf{v}_m)\}$ is linearly independent. This follows since, if $a_1 f(\mathbf{v}_1) + \cdots + a_m f(\mathbf{v}_m) = \mathbf{0}$, then we have*

$$\mathbf{0} = a_1 f(\mathbf{v}_1) + \cdots + a_m f(\mathbf{v}_m) = f(a_1\mathbf{v}_1 + \cdots + a_m\mathbf{v}_m) = f(\mathbf{0}).$$

But since f is an injection, we must have $a_1\mathbf{v}_1 + \cdots + a_m\mathbf{v}_m = \mathbf{0}$. Thus $a_i = 0$ for all i.

On the other hand, what is the condition for $f : \mathbf{V} \to \mathbf{W}$ to be a surjection (epimorphism)? That is, $f(\mathbf{V}) = \mathbf{W}$. Or put another way, for every $\mathbf{w} \in \mathbf{W}$, can we find some vector $\mathbf{v} \in \mathbf{V}$ with $f(\mathbf{v}) = \mathbf{w}$? One way to think of this is to consider a basis $B \subset \mathbf{W}$. For each $\mathbf{w} \in B$, we take

$$f^{-1}(\mathbf{w}) = \{\mathbf{v} \in \mathbf{V} : f(\mathbf{v}) = \mathbf{w}\}.$$

Then f is a surjection if $f^{-1}(\mathbf{w}) \neq \emptyset$, for all $\mathbf{w} \in B$.

Definition. *A linear mapping which is a bijection (that is, an injection and a surjection) is called an* isomorphism*. Often one writes $\mathbf{V} \cong \mathbf{W}$ to say that there exists an isomorphism from $\mathbf{V}$ to $\mathbf{W}$.*

Theorem 3.16. *Let $f : \mathbf{V} \to \mathbf{W}$ be an isomorphism. Then the inverse mapping $f^{-1} : \mathbf{W} \to \mathbf{V}$ is also a linear mapping.*

Proof. To see this, let $a, b \in F$ and $\mathbf{x}, \mathbf{y} \in \mathbf{W}$ be arbitrary. Let $f^{-1}(\mathbf{x}) = \mathbf{u} \in \mathbf{V}$ and $f^{-1}(\mathbf{y}) = \mathbf{v} \in \mathbf{V}$, say. Then

$$f(a\mathbf{u} + b\mathbf{v}) = (f(af^{-1}(\mathbf{x}) + bf^{-1}(\mathbf{v})) = af(f^{-1}(\mathbf{x})) + bf(f^{-1}(\mathbf{v})) = a\mathbf{x} + b\mathbf{y}.$$

Therefore, since f is a bijection, we must have

$$f^{-1}(a\mathbf{x} + b\mathbf{y}) = a\mathbf{u} + b\mathbf{v} = af^{-1}(\mathbf{x}) + bf^{-1}(\mathbf{y}).$$

□

Theorem 3.17. *Let* $\mathbf{V}$ *and* $\mathbf{W}$ *be finite dimensional vector spaces over a field* F*, and let* $f : \mathbf{V} \to \mathbf{W}$ *be a linear mapping. Let* $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ *be a basis for* $\mathbf{V}$*. Then* f *is uniquely determined by the* n *vectors* $\{f(\mathbf{v}_1), \ldots, f(\mathbf{v}_n)\}$ *in* $\mathbf{W}$*.*

Proof. Let $\mathbf{v} \in \mathbf{V}$ be an arbitrary vector in $\mathbf{V}$. Since B is a basis for $\mathbf{V}$, we can uniquely write

$$\mathbf{v} = a_1\mathbf{v}_1 + \cdots a_n\mathbf{v}_n,$$

with $a_i \in F$, for each i. Then, since the mapping f is linear, we have

$$\begin{aligned} f(\mathbf{v}) &= f(a_1\mathbf{v}_1 + \cdots a_n\mathbf{v}_n) \\ &= f(a_1\mathbf{v}_1) + \cdots + f(a_n\mathbf{v}_n) \\ &= a_1 f(\mathbf{v}_1) + \cdots + a_n f(\mathbf{v}_n). \end{aligned}$$

Therefore we see that if the values of $f(\mathbf{v}_1)$,..., $f(\mathbf{v}_n)$ are given, then the value of $f(\mathbf{v})$ is uniquely determined, for each $\mathbf{v} \in \mathbf{V}$.

On the other hand, let $A = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ be a set of n *arbitrarily given* vectors in $\mathbf{W}$. Then let a mapping $f : \mathbf{V} \to \mathbf{W}$ be *defined* by the rule

$$f(\mathbf{v}) = a_1\mathbf{u}_1 + \cdots a_n\mathbf{u}_n,$$

for each arbitrarily given vector $\mathbf{v} \in \mathbf{V}$, where $\mathbf{v} = a_1\mathbf{v}_1 + \cdots a_n\mathbf{v}_n$. Clearly the mapping is uniquely determined, since $\mathbf{v}$ is uniquely determined as a linear combination of the basis vectors B. It is a trivial matter to verify that the mapping which is so defined is also linear. We have $f(\mathbf{v}_i) = \mathbf{u}_i$ for all the basis vectors $\mathbf{v}_i \in B$. □

Theorem 3.18. *Let* $\mathbf{V}$ *and* $\mathbf{W}$ *be two finite dimensional vector spaces over a field* F*. Then we have* $\mathbf{V} \cong \mathbf{W} \Leftrightarrow dim(\mathbf{V}) = dim(\mathbf{W})$*.*

Proof. "⇒" Let $f : \mathbf{V} \to \mathbf{W}$ be an isomorphism, and let $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbf{V}$ be a basis for $\mathbf{V}$. Then, as shown in our Remark above, we have $A = \{f(\mathbf{v}_1), \ldots, f(\mathbf{v}_n)\} \subset \mathbf{W}$ being linearly independent. Furthermore, since B is a basis of $\mathbf{V}$, we have $[B] = \mathbf{V}$. Thus $[A] = \mathbf{W}$ also. Therefore A is a basis of $\mathbf{W}$, and it contains precisely n elements; thus $dim(\mathbf{V}) = dim(\mathbf{W})$.

"⇐" Take $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbf{V}$ to again be a basis of $\mathbf{V}$ and let $A = \{\mathbf{w}_1, \ldots, \mathbf{w}_n\} \subset \mathbf{W}$ be some basis of $\mathbf{W}$ (with n elements). Now define the mapping $f : \mathbf{V} \to \mathbf{W}$ by the rule $f(\mathbf{v}_i) = \mathbf{w}_i$, for all i. By theorem 3.17 we see that a linear mapping f is thus uniquely determined. Since A and B are both bases, it follows that f must be a bijection. □

This immediately gives us a complete classification of all finite-dimensional vector spaces. For let $\mathbf{V}$ be a vector space of dimension n over the field F. Then clearly F^n is also a vector space of dimension n over F. The canonical basis is the set of vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, where

$$\mathbf{e}_i = (0, \cdots, 0, \underbrace{1}_{i\text{-th Position}}, 0, \ldots, 0\}$$

for each i. Therefore, when thinking about $\mathbf{V}$, we can think that it is "really" just F^n. On the other hand, the central idea in the theory of linear algebra is that we can look at things using different possible bases (or "frames of reference"). The space F^n seems to have a preferred, fixed frame of reference, namely the canonical basis. Thus it is better to think about an abstract $\mathbf{V}$, with various possible bases.

Examples

For these examples, we will consider the 2-dimensional real vector space $\mathbb{R}^2$, together with its canonical basis $B = \{\mathbf{e}_1, \mathbf{e}_2\} = \{(1,0),(0,1)\}$.

- $f_1 : \mathbb{R}^2 \to \mathbb{R}^2$ with $f_1(\mathbf{e}_1) = (-1,0)$ and $f_1(\mathbf{e}_2) = (0,1)$. This is a *reflection* of the 2-dimensional plane into itself, with the axis of reflection being the second coordinate axis; that is the set of points $(x_1, x_2) \in \mathbb{R}^2$ with $x_1 = 0$.

- $f_2 : \mathbb{R}^2 \to \mathbb{R}^2$ with $f_2(\mathbf{e}_1) = \mathbf{e}_2$ and $f_1(\mathbf{e}_2) = \mathbf{e}_1$. This is a reflection of the 2-dimensional plane into itself, with the axis of reflection being the diagonal axis $x_1 = x_2$.

- $f_3 : \mathbb{R}^2 \to \mathbb{R}^2$ with $f_3(\mathbf{e}_1) = (\cos\phi, \sin\phi)$ and $f_3(\mathbf{e}_2) = (-\sin\phi, \cos\phi)$, for some real number $\phi \in \mathbb{R}$. This is a *rotation* of the plane about its middle point, through an angle of ϕ.[2] For let $\mathbf{v} = (x_1, x_2)$ be some arbitrary point of the plane $\mathbb{R}^2$. Then we have

$$\begin{aligned} f_3(\mathbf{v}) &= x_1 f_3(\mathbf{e}_1) + x_2 f(\mathbf{e}_2) \\ &= x_1(\cos\phi, \sin\phi) + x_2(-\sin\phi, \cos\phi) \\ &= (x_1\cos\phi - x_2\sin\phi, x_1\sin\phi + x_2\cos\phi). \end{aligned}$$

  Looking at this from the point of view of geometry, the question is, what happens to the vector $\mathbf{v}$ when it is rotated through the angle ϕ while preserving its length? Perhaps the best way to look at this is to think about $\mathbf{v}$ in *polar coordinates*. That is, given any two real numbers x_1 and x_2 then, assuming that they are not both zero, we find two *unique* real numbers $r \geq 0$ and $\theta \in [0, 2\pi)$, such that

$$x_1 = r\cos\theta \quad \text{and} \quad x_2 = r\sin\theta,$$

  where $r = \sqrt{x_1^2 + x_2^2}$. Then $\mathbf{v} = (r\cos\theta, r\sin\theta)$. So a rotation of $\mathbf{v}$ through the angle ϕ must bring it to the new vector $(r\cos(\phi+\theta), r\sin(\phi+\theta))$ which, if we remember the formulas for cosines and sines of sums, turns out to be

$$(r(\cos(\theta)\cos(\phi) - \sin(\theta)\sin(\phi)), r(\sin(\theta)\cos(\phi) - \cos(\theta)\sin(\phi)).$$

  But then, remembering that $x_1 = r\cos\theta$ and $x_2 = r\sin\theta$, we see that the rotation brings the vector $\mathbf{v}$ into the new vector

$$(x_1\cos\phi - x_2\sin\phi, x_1\sin\phi + x_2\cos\phi),$$

  which was precisely the specification for $f_3(\mathbf{v})$.

3.5 Linear mappings and matrices

This last example of a linear mapping of $\mathbb{R}^2$ into itself — which should have been simple to describe — has brought with it long lines of lists of coordinates which are difficult to think about.

[2] In analysis, we learn about the formulas of trigonometry. In particular we have

$$\begin{aligned} \cos(\theta+\phi) &= \cos(\theta)\cos(\phi) - \sin(\theta)\sin(\phi), \\ \sin(\theta+\phi) &= \sin(\theta)\cos(\phi) - \cos(\theta)\sin(\phi). \end{aligned}$$

Taking $\theta = \pi/2$, we note that $\cos(\phi + \pi/2) = -\sin(\phi)$ and $\sin(\phi + \pi/2) = \cos(\phi)$.

In three and more dimensions, things become even worse! Thus it is obvious that we need a more sensible system for describing these linear mappings. The usual system is to use *matrices*.

Now, the most obvious problem with our previous notation for vectors was that the lists of the coordinates $(x_1, \cdots, x_n)$ run over the page, leaving hardly any room left over to describe symbolically what we want to do with the vector. The solution to this problem is to write vectors not as *horizontal* lists, but rather as *vertical* lists. We say that the horizontal lists are *row vectors*, and the vertical lists are *column vectors*. This is a great improvement! So whereas before, we wrote

$$\mathbf{v} = (x_1, \cdots, x_n),$$

now we will write

$$\mathbf{v} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

It is true that we use up lots of vertical space on the page in this way, but since the rest of the writing is horizontal, we can afford to waste this vertical space. In addition, we have a very nice system for writing down the coordinates of the vectors after they have been mapped by a linear mapping.

To illustrate this system, consider the rotation of the plane through the angle ϕ, which was described in the last section. In terms of row vectors, we have (x_1, x_2) being rotated into the new vector $(x_1 \cos\phi - x_2 \sin\phi, x_1 \sin\phi + x_2 \cos\phi)$. But if we change into the column vector notation, we have

$$\mathbf{v} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

being rotated to

$$\begin{pmatrix} x_1 \cos\phi - x_2 \sin\phi \\ x_1 \sin\phi + x_2 \cos\phi \end{pmatrix}.$$

But then, remembering how we *multiplied* matrices, we see that this is just

$$\begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \cos\phi - x_2 \sin\phi \\ x_1 \sin\phi + x_2 \cos\phi \end{pmatrix}.$$

So we can say that the 2×2 matrix $A = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$ represents the mapping $f_3 : \mathbb{R}^2 \to \mathbb{R}^2$, and the 2×1 matrix $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ represents the vector $\mathbf{v}$. Thus we have

$$A \cdot \mathbf{v} = f(\mathbf{v}).$$

That is, matrix multiplication gives the result of the linear mapping.

Expressing $f : \mathbf{V} \to \mathbf{W}$ in terms of bases for both $\mathbf{V}$ and $\mathbf{W}$

The example we have been thinking about up till now (a rotation of $\mathbb{R}^2$) is a linear mapping of $\mathbb{R}^2$ into itself. More generally, we have linear mappings from a vector space $\mathbf{V}$ to a *different* vector space $\mathbf{W}$ (although, of course, both $\mathbf{V}$ and $\mathbf{W}$ are vector spaces over the same field F).

So let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis for $\mathbf{V}$ and let $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ be a basis for $\mathbf{W}$. Finally, let $f : \mathbf{V} \to \mathbf{W}$ be a linear mapping. An arbitrary vector $\mathbf{v} \in \mathbf{V}$ can be expressed in terms of the basis for $\mathbf{V}$ as

$$\mathbf{v} = a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n = \sum_{j=1}^{n} a_j\mathbf{v}_j.$$

The question is now, what is $f(\mathbf{v})$? As we have seen, $f(\mathbf{v})$ can be expressed in terms of the images $f(\mathbf{v}_j)$ of the basis vectors of $\mathbf{V}$. Namely

$$f(\mathbf{v}) = \sum_{j=1}^{n} a_j f(\mathbf{v}_j).$$

But then, each of these vectors $f(\mathbf{v}_j)$ in $\mathbf{W}$ can be expressed in terms of the basis vectors in $\mathbf{W}$, say

$$f(\mathbf{v}_j) = \sum_{i=1}^{m} c_{ij}\mathbf{w}_i,$$

for appropriate choices of the "numbers" $c_{ij} \in F$. Therefore, putting this all together, we have

$$f(\mathbf{v}) = \sum_{j=1}^{n} a_j f(\mathbf{v}_j) = \sum_{j=1}^{n}\sum_{i=1}^{m} a_j c_{ij}\mathbf{w}_i.$$

In the matrix notation, using column vectors relative to the two bases $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$, we can write this as

$$f(\mathbf{v}) = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mn} \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} a_j c_{1j} \\ \vdots \\ \sum_{j=1}^{n} a_j c_{mj} \end{pmatrix}.$$

When looking at this $m \times n$ matrix which represents the linear mapping $f : \mathbf{V} \to \mathbf{W}$, we can imagine that the matrix consists of n columns. The i-th column is then

$$\mathbf{u}_i = \begin{pmatrix} c_{1i} \\ \vdots \\ c_{mi} \end{pmatrix} \in \mathbf{W}.$$

That is, it represents a vector in $\mathbf{W}$, namely the vector $\mathbf{u}_i = c_{1i}\mathbf{w}_1 + \cdots + c_{mi}\mathbf{w}_m$. But what is this vector $\mathbf{u}_i$? In the matrix notation, we have

$$\mathbf{v}_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbf{V},$$

where the single non-zero element of this column matrix is a 1 in the i-th position from the top. But then we have

$$f(\mathbf{v}_i) = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mn} \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} c_{1i} \\ \vdots \\ c_{mi} \end{pmatrix} = \mathbf{u}_i.$$

Therefore *the columns of the matrix representing the linear mapping $f : \mathbf{V} \to \mathbf{W}$ are the images of the basis vectors of $\mathbf{V}$.*

Two linear mappings, one after the other

Things become more interesting when we think about the following situation. Let $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{X}$ be vector spaces over a common field F. Assume that $f : \mathbf{V} \to \mathbf{W}$ and $g : \mathbf{W} \to \mathbf{X}$ are linear. Then the composition $f \circ g : \mathbf{V} \to \mathbf{X}$, given by

$$f \circ g(\mathbf{v}) = g(f(\mathbf{v}))$$

for all $\mathbf{v} \in \mathbf{V}$ is clearly a linear mapping. One can write this as

$$\mathbf{V} \xrightarrow{f} \mathbf{W} \xrightarrow{g} \mathbf{X}.$$

Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis for $\mathbf{V}$, $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ be a basis for $\mathbf{W}$, and $\{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ be a basis for $\mathbf{X}$. Assume that the linear mapping f is given by the matrix

$$A = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mn} \end{pmatrix},$$

and the linear mapping g is given by the matrix

$$B = \begin{pmatrix} d_{11} & \cdots & d_{1m} \\ \vdots & \ddots & \vdots \\ d_{r1} & \cdots & d_{rm} \end{pmatrix}.$$

Then, if $\mathbf{v} = \sum_{j=1}^{n} a_j v_j$ is some arbitrary vector in $\mathbf{V}$, we have

$$\begin{aligned} f \circ g(\mathbf{v}) &= g\left(f\left(\sum_{j=1}^{n} a_j \mathbf{v}_j\right)\right) \\ &= g\left(\sum_{j=1}^{n} a_j f(\mathbf{v}_j)\right) \\ &= g\left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_j c_{ij} \mathbf{w}_i\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} a_j c_{ij} g(\mathbf{w}_i) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{r} a_j c_{ij} d_{ki} \mathbf{x}_k. \end{aligned}$$

There are so many summations here! How can we keep track of everything? The answer is to use the matrix notation. The composition of linear mappings is then simply represented by matrix *multiplication*. That is, if

$$\mathbf{v} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix},$$

then we have

$$f \circ g(\mathbf{v}) = g(f(\mathbf{v})) = \begin{pmatrix} d_{11} & \cdots & d_{1m} \\ \vdots & \ddots & \vdots \\ d_{r1} & \cdots & d_{rm} \end{pmatrix} \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mn} \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = BA\mathbf{v}.$$

So this is the reason we have defined matrix multiplication in this way.[3]

3.6 Matrix transformations

Matrices are used to describe linear mappings $f : \mathbf{V} \to \mathbf{W}$ with respect to particular bases of $\mathbf{V}$ and $\mathbf{W}$. But clearly, if we choose *different* bases than the ones we had been thinking about before, then we will have a different matrix for describing the *same* linear mapping. Later on in these lectures we will see how changing the bases changes the matrix, but for now, it is time to think about various systematic ways of changing matrices — in a purely abstract way.

Elementary column operations

We begin with the elementary column operations. Let us denote the set of all $n \times m$ matrices of elements from the field F by $M(m \times n, F)$. Thus, if

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \in M(m \times n, F)$$

then it contains n columns which, as we have seen, are the images of the basis vectors of the linear mapping which is being represented by the matrix. So The first elementary column operation is to exchange column i with column j, for $i \neq j$. We can write

$$\begin{pmatrix} a_{11} & \cdots & a_{1i} & \cdots & a_{1j} & \cdots & a_{1m} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{m1} & \cdots & a_{mi} & \cdots & a_{mj} & \cdots & a_{mm} \end{pmatrix} \xrightarrow{S_{ij}} \begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1i} & \cdots & a_{1m} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{m1} & \cdots & a_{mj} & \cdots & a_{mi} & \cdots & a_{mm} \end{pmatrix}$$

So this column operation is denoted by S_{ij}. It can be thought of as being a mapping $S_{ij} : M(m \times n, F) \to M(m \times n, F)$.

Another way to imagine this is to say that S is the set of column vectors in the matrix A considered as an ordered list. Thus $S \subset F^m$. Then S_{ij} is the same set of column vectors, but with the positions of the i-th and the j-th vectors interchanged. But obviously, as a subset of F^n, the order of the vectors makes no difference. Therefore we can say that the span of S is the same as the span of S_{ij}. That is $[S] = [S_{ij}]$.

The second elementary column operation, denoted $S_i(a)$, is that we form the scalar product of the element $a \neq 0$ in F with the i-th vector in S. So the i-th vector

$$\begin{pmatrix} a_{1i} \\ \vdots \\ a_{mi} \end{pmatrix}$$

[3] Recall that if $A = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mn} \end{pmatrix}$ is an $m \times n$ matrix and $B = \begin{pmatrix} d_{11} & \cdots & d_{1m} \\ \vdots & \ddots & \vdots \\ d_{r1} & \cdots & d_{rm} \end{pmatrix}$ is an $r \times m$ matrix, then the product BA is an $r \times n$ matrix whose kj-th element is $\sum_{i=1}^{m} d_{ki}c_{ij}$.

is changed to

$$a\begin{pmatrix} a_{1i} \\ \vdots \\ a_{mi} \end{pmatrix} = \begin{pmatrix} aa_{1i} \\ \vdots \\ aa_{mi} \end{pmatrix}.$$

All the other column vectors in the matrix remain unchanged.

The third elementary column operation, denoted $S_{ij}(c)$ is that we take the j-th column (where $j \neq i$) and multiply it with $c \neq 0$, then add it to the i-th column. Therefore the i-th column is changed to

$$\begin{pmatrix} a_{1i} + ca_{1j} \\ \vdots \\ a_{mi} + ca_{mj} \end{pmatrix}.$$

All the other columns — *including the j-th column* — remain unchanged.

Theorem 3.19. *$[S] = [S_{ij}] = [S_i(a)] = [S_{ij}(c)]$, where $i \neq j$ and $a \neq 0 \neq c$.*

Proof. Let us say that $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset F^m$. That is, $\mathbf{v}_i$ is the i-th column vector of the matrix A, for each i. We have already seen that $[S] = [S_{ij}]$ is trivially true. But also, say $\mathbf{v} = x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n$ is some arbitrary vector in $[S]$. Then, since $a \neq 0$, we can write

$$\mathbf{v} = x_1\mathbf{v}_1 + \cdots + a^{-1}x_i(a\mathbf{v}_i) + \cdots + x_n\mathbf{v}_n.$$

Therefore $[S] \subset [S_i(a)]$. The other inclusion, $[S_i(a)] \subset [S]$ is also quite trivial so that we have $[S] = [S_i(a)]$.

Similarly we can write

$$\begin{aligned} \mathbf{v} &= x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n \\ &= x_1\mathbf{v}_1 + \cdots + x_i(\mathbf{v}_i + c\mathbf{v}_j) + \cdots + (x_j - x_ic)\mathbf{v}_j + \cdots + x_n\mathbf{v}_n. \end{aligned}$$

Therefore $[S] \subset [S_{ij}(c)]$, and again, the other inclusion is similar. □

Let us call $[S]$ the *column space* (Spaltenraum), which is a subspace of F^m. Then we see that the column space remains invariant under the three types of elementary column operations. In particular, the *dimension* of the column space remains invariant.

Elementary row operations

Again, looking at the $m \times n$ matrix A in a purely abstract way, we can say that it is made up of m *row vectors*, which are just the rows of the matrix. Let us call them $\mathbf{w}_1, \ldots, \mathbf{w}_m \in F^n$. That is,

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = \begin{pmatrix} \mathbf{w}_1 & = & (a_{11} & \cdots & a_{1n}) \\ & \vdots & & & \\ \mathbf{w}_m & = & (a_{m1} & \cdots & a_{mn}) \end{pmatrix}.$$

Again, we define the three elementary row operations analogously to the way we defined the elementary column operations. Clearly we have the same results. Namely if $R = \{w_1, \ldots, w_m\}$ are the original rows, in their proper order, then we have $[R] = [R_{ij}] = [R_i(a)] = [R_{ij}(c)]$.

But it is perhaps easier to think about the row operations when changing a matrix into a form which is easier to think about. We would like to change the matrix into a step form (*Zeilenstufenform*).

Definition. *The $m \times n$ matrix A is in* step form *if there exists some r with $0 \leq r \leq m$ and indices $1 \leq j_1 < j_2 < \cdots < j_r \leq m$ with $a_{ij_i} = 1$ for all $i = 1, \ldots, r$ and $a_{st} = 0$ for all s, t with $t < j_s$ or $s > j_r$. That is:*

$$A = \begin{pmatrix} \cdots & 1 & a_{1j_1+1} & & \cdots & \cdots & & & a_{1n} \\ 0 & & 1 & a_{2j_2+1} & & \cdots & & & a_{2n} \\ 0 & & 0 & 1 & a_{3j_3+1} & \cdots & & & a_{2n} \\ & & & & \ddots & & & & \vdots \\ 0 & & \cdots & & 0 & 1 & a_{rj_r+1} & \cdots & a_{rn} \\ 0 & & & & 0 & 0 & 0 & & 0 \\ \vdots & & & & \vdots & & & & \vdots \end{pmatrix}.$$

Theorem 3.20. *By means of a finite sequence of elementary row operations, every matrix can be transformed into a matrix in step form.*

Proof. Induction on m, the number of rows in the matrix. We use the technique of "Gaussian elimination", which is simply the usual way anyone would go about solving a system of linear equations. This will be dealt with in the next section. The induction step in this proof, which uses a number of simple ideas which are easy to write on the blackboard, but overly tedious to compose here in TEX, will be described in the lecture. □

Now it is obvious that the *row space* (Zeilenraum), that is $[R] \subset F^n$, has the dimension r, and in fact the non-zero row vectors of a matrix in step form provide us with a basis for the row space. But then, looking at the *column* vectors of this matrix in step form, we see that the columns j_1, j_2, and so on up to j_r are all linearly independent, and they generate the column space. (This is discussed more fully in the lecture!)

Definition. *Given an $m \times n$ matrix, the dimension of the column space is called the column rank; similarly the dimension of the row space is the row rank.*

So, using theorem 3.20, we conclude that:

Theorem 3.21. *For any matrix A, the column rank is equal to the row rank. This common dimension is simply called the* rank *— written $Rank(A)$ — of the matrix.*

Definition. *Let A be a quadratic $n \times n$ matrix. Then A is called* regular *if $Rank(A) = n$, otherwise A is called* singular.

Theorem 3.22. *The $n \times n$ matrix A is regular $\Leftrightarrow$ the linear mapping $f : F^n \to F^n$, represented by the matrix A with respect to the canonical basis of F^n is an isomorphism.*

Proof. '$\Rightarrow$' If A is regular, then the rank of A — namely the dimension of the column space $[S]$ — is n. Since the dimension of F^n is n, we must therefore have $[S] = F^n$. The linear mapping $f : F^n \to F^n$ is then both an injection (since S must be linearly independent) and also a surjection.

'$\Leftarrow$' Since the set of column vectors S is the set of images of the canonical basis vectors of F^n under f, they must be linearly independent. There are n column vectors; thus the rank of A is n. □

3.7 Systems of linear equations

We now take a small diversion from our idea of linear algebra as being a method of describing *geometry*, and instead we will consider simple linear equations. In particular, we consider a system of m equations in n unknowns.

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

We can also think about this as being a vector equation. That is, if

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix},$$

and $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in F^n$ and $\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in F^m$, then our system of linear equations is just the single vector equation

$$A \cdot \mathbf{x} = \mathbf{b}.$$

But what is the most obvious way to solve this system of equations? It is a simple matter to write down an algorithm, as follows. The numbers a_{ij} and b_k are given (as elements of F), and the problem is to find the numbers x_l.

1. Let $i := 1$ and $j := 1$.
2. if $a_{ij} = 0$ then if $a_{kj} = 0$ for all $i < k \leq m$, set $j := j + 1$. Otherwise find the smallest index $k > i$ such that $a_{kj} \neq 0$ and exchange the i-th equation with the k-th equation.
3. Multiply both sides of the (possibly new) i-th equation by a_{ij}^{-1}. Then for each $i < k \leq m$, subtract a_{kj} times the i-th equation from the k-th equation. Therefore, at this stage, after this operation has been carried out, we will have $a_{kj} = 0$, for all $k > i$.
4. Set $i := i + 1$. If $i \leq n$ then return to step 2.

So at this stage, we have transformed the system of linear equations into a system in step form.

The next thing is to solve the system of equations in step form. The problem is that perhaps there is no solution, or perhaps there are many solutions. The easiest way to decide which case we have is to reorder the variables — that is the various x_i — so that the steps start in the upper left-hand corner, and they are all one unit wide. That is, things then look like this:

$$\begin{array}{ccccccccccccc} x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \cdots & + & \cdots & + & a_{1n}x_n & = & b_1 \\ & & x_2 & + & a_{23}x_3 & + & a_{24}x_4 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ & & & & x_3 & + & a_{34}x_4 & + & \cdots & + & a_{3n}x_n & = & b_3 \\ & & & & & & & & \vdots & & & & \\ & & & & & & x_k & + & \cdots & + & a_{kn}x_k & = & b_k \\ & & & & & & & & & & 0 & = & b_{k+1} \\ & & & & & & & & & & & \vdots & \\ & & & & & & & & & & 0 & = & b_m \end{array}$$

(Note that this reordering of the variables is like our first elementary column operation for matrices.)

So now we observe that:

- If $b_l \neq 0$ for some $k+1 \leq l \leq m$, then the system of equations has *no solution*.
- Otherwise, if $k = n$ then the system has precisely one single solution. It is obtained by working backwards through the equations. Namely, the last equation is simply $x_n = b_n$, so that is clear. But then, substitute b_n for x_n in the $n-1$-st equation, and we then have $x_{n-1} = b_{n-1} - a_{n-1\,n}b_n$. By this method, we progress back to the first equation and obtain values for all the x_j, for $1 \leq j \leq n$.
- Otherwise, $k < n$. In this case we can assign *arbitrary* values to the variables $x_{k+1}, \ldots, x_n$, and then that fixes the value of x_k. But then, as before, we progressively obtain the values of x_{k-1}, x_{k-2} and so on, back to x_1.

This algorithm for finding solutions of systems of linear equations is called "Gaussian Elimination".

All of this can be looked at in terms of our matrix notation. Let us call the following $m \times n+1$ matrix the augmented matrix for our system of linear equations:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{pmatrix}.$$

Then by means of elementary row and column operations, the matrix is transformed into the new matrix which is in *simple* step form

$$A' = \begin{pmatrix} 1 & a'_{12} & \cdots & \cdot & a'_{1\,k+1} & \cdots & a'_{1n} & b'_1 \\ 0 & 1 & a'_{23} & \cdot & a'_{2\,k+1} & \cdots & a'_{2n} & b'_2 \\ \vdots & & \ddots & \cdot & \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 1 & a'_{k\,k+1} & \cdots & a'_{kn} & b'_k \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & b'_{k+1} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & b'_m \end{pmatrix}.$$

Finding the eigenvectors of linear mappings

Definition. *Let* $\mathbf{V}$ *be a vector space over a field* F*, and let* $f : \mathbf{V} \to \mathbf{V}$ *be a linear mapping of* $\mathbf{V}$ *into itself. An* eigenvector *of* f *is a non-zero vector* $\mathbf{v} \in \mathbf{V}$ *(so we have* $\mathbf{v} \neq \mathbf{0}$*) such that there exists some* $\lambda \in F$ *with* $f(\mathbf{v}) = \lambda\mathbf{v}$*. The scalar* λ *is then called the* eigenvalue *associated with this eigenvector.*

So if f is represented by the $n \times n$ matrix A (with respect to some given basis of $\mathbf{V}$), then the problem of finding eigenvectors and eigenvalues is simply the problem of solving the equation

$$A\mathbf{v} = \lambda\mathbf{v}.$$

But here *both* λ and $\mathbf{v}$ are variables. So how should we go about things? Well, as we will see, it is necessary to look at the *characteristic polynomial* of the matrix, in order to find an eigenvalue

λ. Then, once an eigenvalue is found, we can consider it to be a constant in our system of linear equations. And they become the *homogeneous*[4] system

$$\begin{array}{ccccccccc} (a_{11}-\lambda)x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & 0 \\ a_{21}x_1 & + & (a_{22}-\lambda)x_2 & + & \cdots & + & a_{2n}x_n & = & 0 \\ & & & & & & & \vdots & \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \cdots & + & (a_{nn}-\lambda)x_n & = & 0 \end{array}$$

which can be easily solved to give us the (or one of the) eigenvector(s) whose eigenvalue is λ.

Now the $n \times n$ *identity* matrix is

$$E = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Thus we see that an eigenvalue is any scalar $\lambda \in F$ such that the vector equation

$$(A - \lambda E)\mathbf{v} = \mathbf{0}$$

has a solution vector $\mathbf{v} \in \mathbf{V}$, such that $\mathbf{v} \neq \mathbf{0}$.[5]

3.8 Invertible matrices

Let $f : \mathbf{V} \to \mathbf{W}$ be a linear mapping, and let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbf{V}$ and $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\} \subset \mathbf{W}$ be bases for $\mathbf{V}$ and $\mathbf{W}$, respectively. Then, as we have seen, the mapping f can be uniquely described by specifying the values of $f(\mathbf{v}_j)$, for each $j = 1, \ldots, n$. We have

$$f(\mathbf{v}_j) = \sum_{i=1}^{m} a_{ij}\mathbf{w}_i,$$

And the resulting matrix $A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$ is the matrix describing f with respect to these *given bases*.

A particular case

This is the case that $\mathbf{V} = \mathbf{W}$. So we have the linear mapping $f : \mathbf{V} \to \mathbf{V}$. But now, we only need a *single* basis for $\mathbf{V}$. That is, $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbf{V}$ is the only basis we need. Thus the matrix for f with respect to this single basis is determined by the specifications

$$f(\mathbf{v}_j) = \sum_{i=1}^{m} a_{ij}\mathbf{v}_i.$$

[4] That is, all the b_i are zero. Thus a homogeneous system with matrix A has the form $A\mathbf{v} = \mathbf{0}$.

[5] Given any solution vector $\mathbf{v}$, then clearly we can multiply it with any scalar $\kappa \in F$, and we have

$$(A - \lambda E)(\kappa\mathbf{v}) = \kappa(A - \lambda E)\mathbf{v} = \kappa\mathbf{0} = \mathbf{0}.$$

Therefore, as long as $\kappa \neq 0$, we can say that $\kappa\mathbf{v}$ is also an eigenvector whose eigenvalue is λ.

A trivial example

For example, one particular case is that we have the identity mapping

$$f = id : \mathbf{V} \to \mathbf{V}.$$

Thus $f(\mathbf{v}) = \mathbf{v}$, for all $\mathbf{v} \in \mathbf{V}$. In this case it is obvious that the matrix of the mapping is the $n \times n$ identity matrix I_n.

Regular matrices

Let us now assume that A is some regular $n \times n$ matrix. As we have seen in theorem 3.22, there is an *isomorphism* $f : \mathbf{V} \to \mathbf{V}$, such that A is the matrix representing f with respect to the given basis of $\mathbf{V}$. According to theorem 3.16, the inverse mapping f^{-1} is also linear, and we have $f^{-1} \circ f = id$. So let f^{-1} be represented by the matrix B (again with respect to the same basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$). Then we must have the matrix equation

$$B \cdot A = I_n.$$

Or, put another way, in the multiplication system of matrix algebra we must have $B = A^{-1}$. That is, the matrix A is *invertible*.

Theorem 3.23. *Every regular matrix is invertible.*

Definition. *The set of all regular $n \times n$ matrices over the field F is denoted $GL(n, F)$.*

Theorem 3.24. *$GL(n, F)$ is a group under matrix multiplication. The identity element is the identity matrix.*

Proof. We have already seen that matrix multiplication is associative. The fact that the identity element in $GL(n, F)$ is the identity matrix is clear. By definition, all members of $GL(n, F)$ have an inverse. It only remains to see that $GL(n, F)$ is closed under matrix multiplication. So let $A, C \in GL(n, F)$. Then there exist A^{-1}, $C^{-1} \in GL(n, F)$, and we have that $C^{-1} \cdot A^{-1}$ is itself an $n \times n$ matrix. But then

$$\left(C^{-1}A^{-1}\right) AC = C^{-1}\left(A^{-1}A\right) C = C^{-1}I_nC = C^{-1}C = I_n.$$

Therefore, according to the definition of $GL(n, F)$, we must also have $AC \in GL(n, F)$. □

Simplifying matrices using multiplication with regular matrices

Theorem 3.25. *Let A be an $m \times n$ matrix. Then there exist regular matrices $C \in GL(m, F)$ and $D \in GL(n, F)$ such that the matrix $A' = CAD^{-1}$ consists simply of zeros, except possibly for a block in the upper lefthand corner, which is an identity matrix. That is*

$$A' = \left(\begin{array}{ccc|ccc} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ \hline 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{array}\right)$$

(Note that A' is also an $m \times n$ matrix. That is, it is not necessarily square.)

Proof. A is the representation of a linear mapping $f : \mathbf{V} \to \mathbf{W}$, with respect to bases $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ of $\mathbf{V}$ and $\mathbf{W}$, respectively. The idea of the proof is to now find *new* bases $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbf{V}$ and $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\} \subset \mathbf{W}$, such that the matrix of f with respect to these new bases is as simple as possible.

So to begin with, let us look at $ker(f) \subset \mathbf{V}$. It is a subspace of $\mathbf{V}$, so its dimension is at most n. In general, it might be less than n, so let us write $dim(ker(f)) = n - p$, for some integer $0 \leq p \leq n$. Therefore we choose a basis for $ker(f)$, and we call it

$$\{\mathbf{x}_{p+1}, \ldots, \mathbf{x}_n\} \subset ker(f) \subset \mathbf{V}.$$

Using the extension theorem (theorem 3.9), we extend this to a basis

$$\{\mathbf{x}_1, \ldots, \mathbf{x}_p, \mathbf{x}_{p+1}, \ldots, \mathbf{x}_n\}$$

for $\mathbf{V}$.

Now at this stage, we look at the images of the vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ under f in $\mathbf{W}$. We find that the set $\{f(\mathbf{x}_1), \ldots, f(\mathbf{x}_p)\} \subset \mathbf{W}$ is linearly independent. To see this, let us assume that we have the vector equation

$$\mathbf{0} = \sum_{i=1}^{p} a_i f(\mathbf{x}_i) = f\left(\sum_{i=1}^{p} a_i \mathbf{x}_i\right)$$

for some choice of the scalars a_i. But that means that $\sum_{i=1}^{p} a_i \mathbf{x}_i \in ker(f)$. However $\{\mathbf{x}_{p+1}, \ldots, \mathbf{x}_n\}$ is a basis for $ker(f)$. Thus we have

$$\sum_{i=1}^{p} a_i \mathbf{x}_i = \sum_{j=p+1}^{n} b_j \mathbf{x}_j$$

for appropriate choices of scalars b_j. But $\{\mathbf{x}_1, \ldots, \mathbf{x}_p, \mathbf{x}_{p+1}, \ldots, \mathbf{x}_n\}$ is a basis for $\mathbf{V}$. Thus it is itself linearly independent and therefore we must have $a_i = 0$ and $b_j = 0$ for all possible i and j. In particular, since the a_i's are all zero, we must have the set $\{f(\mathbf{x}_1), \ldots, f(\mathbf{x}_p)\} \subset \mathbf{W}$ being linearly independent.

To simplify the notation, let us call $f(\mathbf{x}_i) = \mathbf{y}_i$, for each $i = 1, \ldots, p$. Then we can again use the extension theorem to find a basis

$$\{\mathbf{y}_1, \ldots, \mathbf{y}_p, \mathbf{y}_{p+1}, \ldots, \mathbf{y}_m\}$$

of $\mathbf{W}$.

So now we define the isomorphism $g : \mathbf{V} \to \mathbf{V}$ by the rule

$$g(\mathbf{x}_i) = \mathbf{v}_i, \quad \text{for all} \quad i = 1, \ldots, n.$$

Similarly the isomorphism $h : \mathbf{W} \to \mathbf{W}$ is defined by the rule

$$h(\mathbf{y}_j) = \mathbf{w}_j, \quad \text{for all} \quad j = 1, \ldots, m.$$

Let D be the matrix representing the mapping g with respect to the basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of $\mathbf{V}$, and also let C be the matrix representing the mapping h with respect to the basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ of $\mathbf{W}$.

Let us now look at the mapping

$$h \cdot f \cdot g^{-1} : \mathbf{V} \to \mathbf{W}.$$

For the basis vector $\mathbf{v}_i \in \mathbf{V}$, we have

$$hfg^{-1}(\mathbf{v}_i) = hf(\mathbf{x}_i) = \begin{cases} h(\mathbf{y}_i) = \mathbf{w}_i, & \text{for } i \leq p \\ h(\mathbf{0}) = \mathbf{0}, & \text{otherwise.} \end{cases}$$

This mapping must therefore be represented by a matrix in our simple form, consisting of only zeros, except possibly for a block in the upper lefthand corner which is an identity matrix. Furthermore, the rule that the composition of linear mappings is represented by the product of the respective matrices leads to the conclusion that the matrix $A' = CAD^{-1}$ must be of the desired form. □

3.9 Similar matrices; changing bases

Definition. *Let A and A' be $n \times n$ matrices. If a matrix $C \in GL(n, F)$ exists, such that $A' = C^{-1}AC$ then we say that the matrices A and A' are* similar.

Theorem 3.26. *Let $f : \mathbf{V} \to \mathbf{V}$ be a linear mapping and let $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$, $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be two bases for $\mathbf{V}$. Assume that A is the matrix for f with respect to the basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and furthermore A' is the matrix for f with respect to the basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$. Let $\mathbf{u}_i = \sum_{j=1}^{n} c_{ji}\mathbf{v}_j$ for all i, and*

$$C = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix}$$

Then we have $A' = C^{-1}AC$.

Proof. From the definition of A', we have

$$f(\mathbf{u}_i) = \sum_{j=1}^{n} a'_{ji}\mathbf{u}_j$$

for all $i = 1, \ldots, n$. On the other hand we have

$$\begin{aligned} f(\mathbf{u}_i) &= f\left(\sum_{j=1}^{n} c_{ji}\mathbf{v}_j\right) \\ &= \sum_{j=1}^{n} c_{ji} f(\mathbf{v}_j) \\ &= \sum_{j=1}^{n} c_{ji}\left(\sum_{k=1}^{n} a_{kj}\mathbf{v}_k\right) \\ &= \sum_{k=1}^{n}\left(\sum_{j=1}^{n} c_{ji}a_{kj}\right)\mathbf{v}_k \\ &= \sum_{k=1}^{n}\left(\sum_{j=1}^{n} a_{kj}c_{ji}\right)\left(\sum_{l=1}^{n} c^*_{lk}\mathbf{u}_l\right) \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\left(c^*_{lk}a_{kj}c_{ji}\right)\mathbf{u}_l. \end{aligned}$$

Here, the inverse matrix C^{-1} is denoted by

$$C^{-1} = \begin{pmatrix} c_{11}^* & \cdots & c_{1n}^* \\ \vdots & \ddots & \vdots \\ c_{n1}^* & \cdots & c_{nn}^* \end{pmatrix}.$$

Therefore we have $A' = C^{-1}AC$. □

Note that we have written here $\mathbf{v}_k = \sum_{l=1}^n c_{lk}^* \mathbf{u}_l$, and then we have said that the resulting matrix (which we call C^*) is, in fact, C^{-1}. To see that this is true, we begin with the definition of C itself. We have

$$\mathbf{u}_l = \sum_{j=1}^n c_{jl} \mathbf{v}_j.$$

Therefore

$$\mathbf{v}_k = \sum_{l=1}^n \sum_{j=1}^n c_{jl} c_{lk}^* \mathbf{v}_j.$$

That is, $CC^* = I_n$, and therefore $C^* = C^{-1}$.

Which mapping does the matrix C represent? From the equations $\mathbf{u}_i = \sum_{j=1}^n c_{ji} \mathbf{v}_j$ we see that it represents a mapping $g : \mathbf{V} \to \mathbf{V}$ such that $g(\mathbf{v}_i) = \mathbf{u}_i$ for all i, expressed in terms of the original basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. So we see that a *similarity transformation*, taking a square matrix A to a similar matrix $A' = C^{-1}AC$ is always associated with a *change of basis* for the vector space V.

Much of the theory of linear algebra is concerned with finding a *simple* basis (with respect to a given linear mapping of the vector space into itself), such that the matrix of the mapping with respect to this simpler basis is itself simple — for example diagonal, or at least triagonal.

3.10 Eigenvalues, eigenspaces, matrices which can be diagonalized

Definition. *Let $f : \mathbf{V} \to \mathbf{V}$ be a linear mapping of an n-dimensional vector space into itself. A subspace $\mathbf{U} \subset \mathbf{V}$ is called* invariant *with respect to f if $f(\mathbf{U}) \subset \mathbf{U}$. That is, $f(\mathbf{u}) \in \mathbf{U}$ for all $\mathbf{u} \in \mathbf{U}$.*

Theorem 3.27. *Assume that the r dimensional subspace $\mathbf{U} \subset \mathbf{V}$ is invariant with respect to $f : \mathbf{V} \to \mathbf{V}$. Let A be the matrix representing f with respect to a given basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of $\mathbf{V}$. Then A is similar to a matrix A' which has the following form*

$$A' = \left(\begin{array}{ccc|ccc} a'_{11} & \ldots & a'_{1r} & a'_{1(r+1)} & \ldots & a'_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a'_{r1} & \ldots & a'_{rr} & a'_{r(r+1)} & \ldots & a'_{rn} \\ \hline & & & a'_{(r+1)(r+1)} & \ldots & a'_{(r+1)n} \\ & 0 & & \vdots & & \vdots \\ & & & a'_{n(r+1)} & \ldots & a'_{nn} \end{array}\right)$$

Proof. Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ be a basis for the subspace $\mathbf{U}$. Then extend this to a basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \ldots, \mathbf{u}_n\}$ of $\mathbf{V}$. The matrix of f with respect to this new basis has the desired form. □

Definition. *Let* $\mathbf{U}_1, \ldots, \mathbf{U}_p \subset \mathbf{V}$ *be subspaces. We say that* $\mathbf{V}$ *is the* direct sum *of these subspaces if* $\mathbf{V} = \mathbf{U}_1 + \cdots + \mathbf{U}_p$*, and furthermore if* $\mathbf{v} = \mathbf{u}_1 + \cdots + \mathbf{u}_p$ *such that* $\mathbf{u}_i \in \mathbf{U}_i$*, for each* i*, then this expression for* $\mathbf{v}$ *is* unique*. In other words, if* $\mathbf{v} = \mathbf{u}_1 + \cdots + \mathbf{u}_p = \mathbf{u}'_1 + \cdots + \mathbf{u}'_p$ *with* $\mathbf{u}'_i \in \mathbf{U}_i$ *for each* i*, then* $\mathbf{u}_i = \mathbf{u}'_i$*, for each* i*. In this case, one writes* $\mathbf{V} = \mathbf{U}_1 \oplus \cdots \oplus \mathbf{U}_p$

This immediately gives the following result:

Theorem 3.28. *Let* $f : \mathbf{V} \to \mathbf{V}$ *be such that there exist subspaces* $\mathbf{U}_i \subset \mathbf{V}$*, for* $i = 1, \ldots, p$*, such that* $\mathbf{V} = \mathbf{U}_1 \oplus \cdots \oplus \mathbf{U}_p$ *and also* f *is invariant with respect to each* $\mathbf{U}_i$*. Then there exists a basis of* $\mathbf{V}$ *such that the matrix of* f *with respect to this basis has the following* block form*.*

$$A = \begin{pmatrix} A_1 & 0 & \ldots & & 0 \\ 0 & A_2 & 0 & & \\ \vdots & 0 & \ddots & 0 & \vdots \\ & & 0 & A_{p-1} & 0 \\ 0 & & \ldots & 0 & A_p \end{pmatrix}$$

where each block A_i *is a square matrix, representing the restriction of* f *to the subspace* $\mathbf{U}_i$*.*

Proof. Choose the basis to be a union of bases for each of the $\mathbf{U}_i$. □

A special case is when the invariant subspace is an eigenspace.

Definition. *Assume that* $\lambda \in F$ *is an eigenvalue of the mapping* $f : \mathbf{V} \to \mathbf{V}$*. The set* $\{\mathbf{v} \in \mathbf{V} : f(\mathbf{v}) = \lambda\mathbf{v}\}$ *is called the* eigenspace *of* λ *with respect to the mapping* f*. That is, the eigenspace is the set of all eigenvectors (and with the zero vector* $\mathbf{0}$ *included) with eigenvalue* λ*.*

Theorem 3.29. *Each eigenspace is a subspace of* $\mathbf{V}$*.*

Proof. Let $\mathbf{u}, \mathbf{w} \in \mathbf{V}$ be in the eigenspace of λ. Let $a, b \in F$ be arbitrary scalars. Then we have

$$f(a\mathbf{u} + b\mathbf{w}) = af(\mathbf{u}) + bf(\mathbf{w}) = a\lambda\mathbf{u} + b\lambda\mathbf{w} = \lambda(a\mathbf{u} + b\mathbf{w}).$$

□

Obviously if λ_1 and λ_2 are two different ($\lambda_1 \neq \lambda_2$) eigenvalues, then the only common element of the eigenspaces is the zero vector $\mathbf{0}$. Thus if every vector in $\mathbf{V}$ is an eigenvector, then we have the situation of theorem 3.28. One very particular case is that we have n different eigenvalues, where n is the dimension of $\mathbf{V}$.

Theorem 3.30. *Let* $\lambda_1, \ldots, \lambda_n$ *be eigenvalues of the linear mapping* $f : \mathbf{V} \to \mathbf{V}$*, where* $\lambda_i \neq \lambda_j$ *for* $i \neq j$*. Let* $\mathbf{v}_1, \ldots, \mathbf{v}_n$ *be eigenvectors to these eigenvalues. That is,* $\mathbf{v}_i \neq \mathbf{0}$ *and* $f(\mathbf{v}_i) = \lambda_i\mathbf{v}_i$*, for each* $i = 1, \ldots, n$*. Then the set* $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ *is linearly independent.*

Proof. Assume to the contrary that there exist $a_1, \ldots, a_n$, not all zero, with

$$a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n = \mathbf{0}.$$

Assume further that as few of the a_i as possible are non-zero. Let a_p be the first non-zero scalar. That is, $a_i = 0$ for $i < p$, and $a_p \neq 0$. Obviously some other a_k is non-zero, for some $k \neq p$, for otherwise we would have the equation $\mathbf{0} = a_p\mathbf{v}_p$, which would imply that $\mathbf{v}_p = \mathbf{0}$, contrary to the assumption that $\mathbf{v}_p$ is an eigenvector. Therefore we have

$$\mathbf{0} = f(\mathbf{0}) = f\left(\sum_{i=1}^{n} a_i\mathbf{v}_i\right) = \sum_{i=1}^{n} a_i f(\mathbf{v}_i) = \sum_{i=1}^{n} a_i\lambda_i\mathbf{v}_i.$$

Also

$$\mathbf{0} = \lambda_p \mathbf{0} = \lambda_p \left(\sum_{i=1}^{n} a_i \mathbf{v}_i \right).$$

Therefore

$$\mathbf{0} = \mathbf{0} - \mathbf{0} = \lambda_p \left(\sum_{i=1}^{n} a_i \mathbf{v}_i \right) - \sum_{i=1}^{n} a_i \lambda_i \mathbf{v}_i = \sum_{i=1}^{n} a_i (\lambda_p - \lambda_i) \mathbf{v}_i.$$

But, remembering that $\lambda_i \neq \lambda_j$ for $i \neq j$, we see that the scalar term for $\mathbf{v}_p$ is zero, yet all other non-zero scalar terms remain non-zero. Thus we have found a new sum with fewer non-zero scalars than in the original sum with the a_is. This is a contradiction. □

Therefore, in this particular case, the given set of eigenvectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ form a basis for $\mathbf{V}$. With respect to this basis, the matrix of the mapping is *diagonal*, with the diagonal elements being the eigenvalues.

$$A = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

3.11 The elementary matrices

These are $n \times n$ matrices which we denote by S_{ij}, $S_i(a)$, and $S_{ij}(c)$. They are such that when any $n \times n$ matrix A is multiplied on the *right* by such an S, then the given elementary *column* operation is performed on the matrix A. Furthermore, if the matrix A is multiplied on the *left* by such an elementary matrix, then the given *row* operation on the matrix is performed.[6] It is a simple matter to verify that the following matrices are the ones we are looking for.

$$S_{ij} = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & 0 & \\ & & 1 & & & & & & \\ & & & 0 & \overset{i-\text{th row}}{\longrightarrow} & & 1 & & \\ & & & & 1 & & & & \\ & & & \downarrow & & \ddots & & \uparrow & \\ & & & & & & 1 & & \\ & & & 1 & & \underset{j-\text{th row}}{\longleftarrow} & & 0 & \\ & & & & & & & & 1 \\ & & 0 & & & & & & \ddots \\ & & & & & & & & & 1 \end{pmatrix}$$

Here, everything is zero except for the two elements at the positions ij and ji, which have the value 1. Also the diagonal elements are all 1 except for the elements at ii and jj, which are zero.

[6] But note that using the matrix $S_{ij}(c)$, we find that the product $S_{ij}(c) \cdot A$ gives the row operation on the matrix A, where the row j is multiplied with the scalar c, then added to the row i. On the other hand, the column operation where the j-th column is multiplied with the scalar c and added to the i-th column, is given by the matrix product $A \cdot S_{ji}(c)$.

Then we have

$$S_i(a) = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & 0 & \\ & & 1 & & & & \\ & & & a & & & \\ & & & & 1 & & \\ & 0 & & & & \ddots & \\ & & & & & & 1 \end{pmatrix}$$

That is, $S_i(a)$ is a diagonal matrix, all of whose diagonal elements are 1 except for the single element at the position ii, which has the value a.

Finally we have

$$S_{ij}(c) = \begin{pmatrix} 1 & & & & & & & 0 \\ & \ddots & & & & & & \\ & & 1 & \overset{i-\text{th row}}{\longrightarrow} & & c & & \\ & & & 1 & & & & \\ & & & & \ddots & \uparrow & j-\text{th column} & \\ & & & & & 1 & & \\ & & & & & & 1 & \\ & 0 & & & & & & \ddots \\ & & & & & & & 1 \end{pmatrix}$$

So this is again just the $n \times n$ identity matrix, but this time we have replaced the zero in the ij-th position with the scalar c. It is an elementary exercise to see that:

Theorem 3.31. *Each of the $n \times n$ elementary matrices are regular.*

And thus we can prove that these elementary matrices *generate* the group $GL(n, F)$. Furthermore, for every elementary matrix, the inverse matrix is again elementary.

Theorem 3.32. *Every matrix in $GL(n, F)$ can be represented as a product of elementary matrices.*

Proof. Let $A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \in GL(n, F)$ be some arbitrary regular matrix. We have already seen that A can be transformed into a matrix in step form by means of elementary row operations. That is, there is some sequence of elementary matrices: $S_1, \ldots, S_p$, such that the product

$$A^* = S_p \cdots S_1 A$$

is an $n \times n$ matrix in step form. However, since A was a *regular* matrix, the number of steps must be equal to n. That is, A^* must be a *triangular* matrix whose diagonal elements are all equal to 1.

$$A^* = \begin{pmatrix} 1 & a^*_{12} & a^*_{13} & \star & \cdots & \star & a^*_{1n} \\ 0 & 1 & a^*_{23} & \star & \cdots & \star & a^*_{2n} \\ 0 & 0 & 1 & \star & \cdots & \star & a^*_{3n} \\ \vdots & \vdots & & \ddots & \star & \star & \vdots \\ 0 & & \cdots & 0 & 1 & a^*_{(n-2)(n-1)} & a^*_{(n-2)n} \\ 0 & & & \cdots & 0 & 1 & a^*_{(n-1)n} \\ 0 & & & \cdots & & 0 & 1 \end{pmatrix}$$

But now it is obvious that the elements above the diagonal can all be reduced to zero by elementary row operations of type $S_{ij}(c)$. These row operations can again be realized by multiplication of A^* on the right by some further set of elementary matrices: S_{p+1}, ..., S_q. This gives us the matrix equation

$$S_q \cdots S_{p+1} S_p \cdots S_1 A = I_n$$

or

$$A = S_1^{-1} \cdot \cdots \cdot S_p^{-1} S_{p+1}^{-1} \cdots S_q^{-1}.$$

Since the inverse of each elementary matrix is itself elementary, we have thus expressed A as a product of elementary matrices. □

This proof also shows how we can go about programming a computer to calculate the inverse of an invertible matrix. Namely, through the process of Gauss elimination, we convert the given matrix into the identity matrix I_n. During this process, we keep multiplying together the elementary matrices which represent the respective row operations. In the end, we obtain the inverse matrix

$$A^{-1} = S_q \cdots S_{p+1} S_p \cdots S_1.$$

We also note that this is the method which can be used to obtain the value of the determinant function for the matrix. But first we must find out what the definition of determinants of matrices is!

3.12 The determinant

Let $M(n \times n, F)$ be the set of all $n \times n$ matrices of elements of the field F.

Definition. *A mapping $det : M(n \times n, F) \to F$ is called a* determinant function *if it satisfies the following four conditions.*

1. *$det(I_n) = 1$, where I_n is the identity matrix.*
2. *If $A \in M(n \times n, F)$ is changed to the matrix A' by multiplying all the elements in a single row with the scalar $a \in F$, then $det(A') = a \cdot det(A)$. (This is our row operation $S_i(a)$.)*
3. *Let $A = (\mathbf{u}_1, \ldots, \mathbf{u}_k, \ldots, \mathbf{u}_n)$, where $\mathbf{u}_k$ is the k-th row of the matrix, for each k. Assume that for some particular $i \in \{1, \ldots, n\}$ we have $\mathbf{u}_i = \mathbf{u}_i' + \mathbf{u}_i''$. Let $A' = (\mathbf{u}_1, \ldots, \mathbf{u}_i', \ldots \mathbf{u}_n)$ and $A'' = (\mathbf{u}_1, \ldots, \mathbf{u}_i'', \ldots \mathbf{u}_n)$. Then[7] $det(A) = det(A') + det(A'')$.*
4. *If A' is obtained from A by adding one row to a different row, then $det(A') = det(A)$. (This is our row operation $S_{ij}(1)$.)*

Simple consequences of this definition

Let $A \in M(n \times n, F)$ be an arbitrary $n \times n$ matrix, and let us say that A is transformed into the new matrix A' by an elementary row operation. Then we have:

1. If A' is obtained by multiplying row i by the scalar $a \in F$, then $det(A') = a \cdot det(A)$. This is completely obvious! It is just part of the definition of "determinants".

[7] Actually, it is not necessary to assume that this *additive* property is part of the definition of the determinant. In fact, it can be proved to follow as a consequence of the other properties.

2. Therefore, if A' is obtained from A by multiplying a row with -1 then we have $det(A') = -det(A)$.

3. Also, it follows that a matrix containing a row consisting of zeros must have zero as its determinant.

4. If A has two identical rows, then its determinant must also be zero. For can we multiply one of these rows with -1, then add it to the other row, obtaining a matrix with a zero row.

5. If A' is obtained by exchanging rows i and j, then $det(A') = -det(A)$. This is a bit more difficult to see. Again, let us say that $A = (\mathbf{u}_1, \ldots, \mathbf{u}_i, \ldots, \mathbf{u}_j, \ldots \mathbf{u}_n)$, where $\mathbf{u}_k$ is the k-th row of the matrix, for each k. Then we can write

$$\begin{aligned} det(A) &= det(\mathbf{u}_1, \ldots, \mathbf{u}_i, \ldots, \mathbf{u}_j, \ldots \mathbf{u}_n) \\ &= det(\mathbf{u}_1, \ldots, \mathbf{u}_i + \mathbf{u}_j, \ldots, \mathbf{u}_j, \ldots, \mathbf{u}_n) \\ &= -det(\mathbf{u}_1, \ldots, -(\mathbf{u}_i + \mathbf{u}_j), \ldots, \mathbf{u}_j, \ldots, \mathbf{u}_n) \\ &= -det(\mathbf{u}_1, \ldots, -(\mathbf{u}_i + \mathbf{u}_j), \ldots, \mathbf{u}_j - (\mathbf{u}_i + \mathbf{u}_j), \ldots, \mathbf{u}_n) \\ &= det(\mathbf{u}_1, \ldots, \mathbf{u}_i + \mathbf{u}_j, \ldots, -\mathbf{u}_i, \ldots, \mathbf{u}_n) \\ &= det(\mathbf{u}_1, \ldots, (\mathbf{u}_i + \mathbf{u}_j) - \mathbf{u}_i, \ldots, -\mathbf{u}_i, \ldots, \mathbf{u}_n) \\ &= det(\mathbf{u}_1, \ldots, \mathbf{u}_j, \ldots, -\mathbf{u}_i, \ldots, \mathbf{u}_n) \\ &= -det(\mathbf{u}_1, \ldots, \mathbf{u}_j, \ldots, \mathbf{u}_i, \ldots, \mathbf{u}_n) \end{aligned}$$

   (This is the elementary row operation S_{ij}.)

6. If A' is obtained from A by an elementary row operation of the form $S_{ij}(c)$, then $det(A') = det(A)$. For we have:

$$\begin{aligned} det(A) &= det(\mathbf{u}_1, \ldots, \mathbf{u}_i, \ldots, \mathbf{u}_j, \ldots, \mathbf{u}_n) \\ &= c^{-1} det(\mathbf{u}_1, \ldots, \mathbf{u}_i, \ldots, c\mathbf{u}_j, \ldots, \mathbf{u}_n) \\ &= c^{-1} det(\mathbf{u}_1, \ldots, \mathbf{u}_i + c\mathbf{u}_j, \ldots, c\mathbf{u}_j, \ldots, \mathbf{u}_n) \\ &= det(\mathbf{u}_1, \ldots, \mathbf{u}_i + c\mathbf{u}_j, \ldots, \mathbf{u}_j, \ldots, \mathbf{u}_n) \end{aligned}$$

7. We have thus shown that if $A = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ is such that the n row vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ are linearly *dependent*, then it follows that $det(A) = 0$.

8. Let $A = (\mathbf{u}_1, \ldots, \mathbf{u}'_i + \mathbf{u}''_i, \ldots, \mathbf{u}_n)$. Furthermore, let $A' = (\mathbf{u}_1, \ldots, \mathbf{u}'_i, \ldots, \mathbf{u}_n)$ and $A'' = (\mathbf{u}_1, \ldots, \mathbf{u}''_i, \ldots, \mathbf{u}_n)$. The set $\{\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}, \mathbf{u}'_i, \mathbf{u}''_i, \mathbf{u}_{i+1}, \ldots, \mathbf{u}_n\}$ consists of $n+1$ row vectors. Since these vectors are all contained in the n-dimensional vector space F^n, they must be linearly dependent. Therefore there is an equation of the form

$$c_1\mathbf{u}_1 + \cdots + c_{i-1}\mathbf{u}_{i-1} + c'\mathbf{u}'_i + c''\mathbf{u}''_i + c_{i+1}\mathbf{u}_{+1} + \cdots + c_n\mathbf{u}_n = \mathbf{0},$$

   where at least one of the coefficients c', c'', or c_k is not equal to zero.

   If we have $c' = c'' = 0$, then we must have $det(A) = det(A') = det(A'') = 0$, since the set of row vectors of each of these matrices is linearly dependent. Therefore, in this case, we have

$$det(A) = det(A') + det(A'').$$

So assume now that at least one of the coefficients c' or c'' is not zero. Say $c' \neq 0$. But then (possibly by multiplying with c'^{-1}) we may assume that $c' = 1$. This gives us the equation

$$\mathbf{u}'_i = -c_1\mathbf{u}_1 - \cdots - c_{i-1}\mathbf{u}_{i-1} - c''\mathbf{u}''_i - c_{i+1}\mathbf{u}_{+1} - \cdots - c_n\mathbf{u}_n.$$

Therefore, if $A^{\#} = (\mathbf{u}_1, \ldots, -c''\mathbf{u}''_i, \ldots, \mathbf{u}_n)$, then, using (6), we have $det(A') = det(A^{\#})$. But using (1), we see that $det(A^{\#}) = -c''det(A'')$. Finally, substituting $-c''\mathbf{u}''_i$ for $\mathbf{u}'_i$ in the matrix A, we obtain the matrix

$$A^* = (\mathbf{u}_1, \ldots, -c''\mathbf{u}''_i +'' \mathbf{u}''_i, \ldots, \mathbf{u}_n) = (\mathbf{u}_1, \ldots, (1-c'')\mathbf{u}''_i, \ldots, \mathbf{u}_n).$$

Using (6) and (1), we have

$$\begin{aligned} det(A) &= det(A^*) \\ &= (1-c'')det(A'') \\ &= -c''det(A'') + det(A'') \\ &= det(A^{\#}) + det(A'') \\ &= det(A') + det(A''). \end{aligned}$$

This shows that the additivity property (number 3 in the definition of the determinant) follows from the other properties 1, 2, and 4.

Therefore we see that each elementary row operation has a well-defined effect on the determinant of the matrix. This gives us the following algorithm for calculating the determinant of an arbitrary matrix in $M(n \times n, F)$.

How to find the determinant of a matrix

Given: An arbitrary matrix $A \in M(n \times n, F)$.
Find: $det(A)$.

Method:

1. Using elementary row operations, transform A into a matrix in step form, keeping track of the changes in the determinant at each stage.

2. If the bottom line of the matrix we obtain only consists of zeros, then the determinant is zero, and thus the determinant of the original matrix was zero.

3. Otherwise, the matrix has been transformed into an upper triangular matrix, all of whose diagonal elements are 1. But now we can transform this matrix into the identity matrix I_n by elementary row operations of the type $S_{ij}(c)$. Since we know that $det(I_n)$ must be 1, we then find a unique value for the determinant of the original matrix A. In particular, in this case $det(A) \neq 0$.

Note that in both this algorithm, as well as in the algorithm for finding the inverse of a regular matrix, the method of Gaussian elimination was used. Thus we can combine both ideas into a single algorithm, suitable for practical calculations in a computer, which yields both the matrix inverse (if it exists), and the determinant. This algorithm also proves the following theorem.

Theorem 3.33. *There is only one determinant function and it is uniquely given by our algorithm. Furthermore, a matrix $A \in M(n \times n, F)$ is regular if and only if $det(A) \neq 0$.*

In particular, using these methods it is easy to see that the following theorem is true.

Theorem 3.34. *Let A, $B \in M(n \times n, F)$. Then we have $det(A \cdot B) = det(A) \cdot det(B)$.*

Proof. If either A or B is singular, then $A \cdot B$ is singular. This can be seen by thinking about the linear mappings $\mathbf{V} \to \mathbf{V}$ which A and B represent. At least one of these mappings is singular. Thus the dimension of the image is less than n, so the dimension of the image of the composition of the two mappings must also be less than n. Therefore $A \cdot B$ must be singular. That means, on the one hand, that $det(A \cdot B) = 0$. And on the other hand, that either $det(A) = 0$ or else $det(B) = 0$. Either way, the theorem is true in this case.

If both A and B are regular, then they are both in $GL(n, F)$. Therefore, as we have seen, they can be written as products of elementary matrices. It suffices then to prove the theorem in the special case that A is an elementary matrix. But this is just part of the "simple consequences" of the definition of the determinant. □

Remembering that A is regular if and only if $A \in GL(n, F)$, we have:

Corollary. *If $A \in GL(n, F)$ then $det(A^{-1}) = (det(A))^{-1}$.*

In particular, if $det(A) = 1$ then we also have $det(A^{-1}) = 1$. The set of all such matrices must then form a group.

Another simple corollary is the following.

Corollary. *Assume that the matrix A is in block form, so that the linear mapping which it represents splits into a direct sum of invariant subspaces (see theorem 3.28). Then $det(A)$ is the product of the determinants of the blocks.*

Proof. If

$$A = \begin{pmatrix} A_1 & 0 & \dots & & 0 \\ 0 & A_2 & 0 & & \\ \vdots & 0 & \ddots & 0 & \vdots \\ & & 0 & A_{p-1} & 0 \\ 0 & & \dots & 0 & A_p \end{pmatrix}$$

then for each $i = 1, \dots, p$ let

$$A_i^* = \begin{pmatrix} 1 & 0 & \dots & & 0 \\ 0 & \ddots & 0 & & \\ \vdots & 0 & A_i & 0 & \vdots \\ & & 0 & \ddots & 0 \\ 0 & & \dots & 0 & 1 \end{pmatrix}.$$

That is, for the matrix A_i^*, all the blocks except the i-th block are replaced with identity-matrix blocks. Then $A = A_1^* \cdots A_p^*$, and it is easy to see that $det(A_i^*) = det(A_i)$ for each i. □

Definition. *The* special linear group *of order n is defined to be the set*

$$SL(n, F) = \{A \in GL(n, F) : det(A) = 1\}.$$

Theorem 3.35. *Let $A' = C^{-1}AC$. Then $det(A') = det(A)$.*

Proof. This follows, since $det(C^{-1}) = (det(C))^{-1}$. □

3.13 Leibniz formula

Definition. *A* permutation *of the numbers $\{1, \ldots, n\}$ is a bijection*

$$\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}.$$

The set of all permutations of the numbers $\{1, \ldots, n\}$ is denoted S_n. In fact, S_n is a group: the symmetric group *of order n. Given a permutation $\sigma \in S_n$, we will say that a pair of numbers (i, j), with $i, j \in \{1, \ldots, n\}$ is a "reversed pair" if $i < j$, yet $\sigma(i) > \sigma(j)$. Let $s(\sigma)$ be the total number of reversed pairs in σ. Then the* sign *of sigma is defined to be the number*

$$sign(\sigma) = (-1)^{s(\sigma)}.$$

Theorem 3.36 (Leibniz)**.** *Let the elements in the matrix A be a_{ij}, for i, j between 1 and n. Then we have*

$$det(A) = \sum_{\sigma \in S_n} sign(\sigma) \prod_{i=1}^{n} a_{\sigma(i)i}.$$

As a consequence of this formula, the following theorems can be proved:

Theorem 3.37. *Let A be a diagonal matrix.*

$$A = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & & \lambda_n \end{pmatrix}$$

Then $det(A) = \lambda_1 \lambda_2 \cdots \lambda_n$.

Theorem 3.38. *Let A be a triangular matrix.*

$$\begin{pmatrix} a_{11} & a_{12} & \star & \cdots & \star \\ 0 & a_{22} & \star & \cdots & \star \\ 0 & 0 & \ddots & & \vdots \\ \vdots & & 0 & a_{(n-1)(n-1)} & a_{(n-1)n} \\ 0 & \cdots & 0 & 0 & a_{nn} \end{pmatrix}$$

Then $det(A) = a_{11} a_{22} \cdots a_{nn}$.

Leibniz formula also gives:

Definition. *Let $A \in M(n \times n, F)$. The* transpose *A^t of A is the matrix consisting of elements a^t_{ij} such that for all i and j we have $a^t_{ij} = a_{ji}$, where a_{ji} are the elements of the original matrix A.*

Theorem 3.39. $det(A^t) = det(A)$.

3.13.1 Special rules for 2×2 and 3×3 matrices

Let $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$. Then Leibniz formula reduces to the simple formula

$$det(A) = a_{11} a_{22} - a_{12} a_{21}.$$

For 3×3 matrices, the formula is a little more complicated. Let $A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$. Then we have

$$det(A) = a_{11} a_{22} a_{33} + a_{12} a_{23} a_{33} + a_{13} a_{21} a_{32} - a_{11} a_{23} a_{32} - a_{12} a_{21} a_{33} - a_{11} a_{23} a_{32}.$$

3.13.2 A proof of Leibniz formula

Let the rows of the $n \times n$ identity matrix be $\epsilon_1, \ldots, \epsilon_n$. Thus

$$\epsilon_1 = (1\ 0\ 0 \cdots 0), \quad \epsilon_2 = (0\ 1\ 0 \cdots 0), \ldots, \quad \epsilon_n = (0\ 0\ 0 \cdots 1).$$

Therefore, given that the i-th row in a matrix is

$$\xi_i = (a_{i1}\ a_{i2}\ \cdots\ a_{in}),$$

then we have

$$\xi_i = \sum_{j=1}^{n} a_{ij}\epsilon_j.$$

So let the matrix A be represented by its rows,

$$A = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix}.$$

Then we can write

$$\begin{aligned} det(A) &= det \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} \\ &= \sum_{j_1=1}^{n} a_{1j_1} det \begin{pmatrix} \epsilon_{j_1} \\ \xi_2 \vdots \\ \xi_n \end{pmatrix} \\ &= \sum_{j_1=1}^{n} a_{1j_1} \sum_{j_2=1}^{n} a_{2j_2} det \begin{pmatrix} \epsilon_{j_1} \\ \epsilon_{j_2} \\ \xi_3 \\ \vdots \\ \xi_n \end{pmatrix} \\ &= \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \cdots \sum_{j_n=1}^{n} a_{1j_1} \cdots a_{nj_n} det \begin{pmatrix} \epsilon_{j_1} \\ \vdots \\ \epsilon_{j_n} \end{pmatrix}. \end{aligned}$$

But what is $det \begin{pmatrix} \epsilon_{j_1} \\ \vdots \\ \epsilon_{j_n} \end{pmatrix}$? To begin with, observe that if $\epsilon_{j_k} = \epsilon_{j_l}$ for some $j_k \neq j_l$, then two rows are identical, and therefore the determinant is zero. Thus we need only the sum over all possible *permutations* $(j_1, j_2, \ldots, j_n)$ of the numbers $(1, 2, \ldots, n)$. Then, given such a permutation, we have the matrix $\begin{pmatrix} \epsilon_{j_1} \\ \vdots \\ \epsilon_{j_n} \end{pmatrix}$. This can be transformed back into the identity matrix $\begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$ by means of successively exchanging pairs of rows. Each time this is done, the determinant changes sign (from +1 to -1, or from -1 to +1). Finally, of course, we know that the determinant of the identity matrix is 1. Therefore we obtain Leibniz formula

$$det(A) = \sum_{\sigma \in S_n} sign(\sigma) \prod_{i=1}^{n} a_{i\sigma(i)}.$$

3.14 The characteristic polynomial

Let $f : \mathbf{V} \to \mathbf{V}$ be a linear mapping, and let $\mathbf{v}$ be an eigenvector of f with $f(\mathbf{v}) = \lambda\mathbf{v}$. That means that $(f - \lambda id)(\mathbf{v}) = \mathbf{0}$; therefore the mapping $(f - \lambda id) : \mathbf{V} \to \mathbf{V}$ is singular. Now consider the matrix A, representing f with respect to some particular basis of $\mathbf{V}$. Since λI_n is the matrix representing the mapping λid, we must have that the difference $A - \lambda I_n$ is a singular matrix. In particular, we have $det(A - \lambda I_n) = 0$.

Another way of looking at this is to take a "variable" x, and then calculate (for example, using the Leibniz formula) the polynomial in x

$$P(x) = det(A - xI_n).$$

This polynomial is called the *characteristic polynomial* for the matrix A. Therefore we have the theorem:

Theorem 3.40. *The zeros of the characteristic polynomial of A are the eigenvalues of the linear mapping $f : \mathbf{V} \to \mathbf{V}$ which A represents.*

Obviously the degree of the polynomial is n for an $n \times n$ matrix A. So let us write the characteristic polynomial in the standard form

$$P(x) = c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0.$$

The coefficients $c_0, \ldots, c_n$ are all elements of our field F.

Now the matrix A represents the mapping f with respect to a particular choice of basis for the vector space $\mathbf{V}$. With respect to some other basis, f is represented by some other matrix A', which is similar to A. That is, there exists some $C \in GL(n, F)$ with $A' = C^{-1}AC$. But we have

$$\begin{aligned} det(A' - xI_n) &= det(C^{-1}AC - xC^{-1}I_nC) \\ &= det(C^{-1}(A - xI_n)C) \\ &= det(C^{-1})det(A - xI_n)det(C) \\ &= det(A - xI_n) \\ &= P(x). \end{aligned}$$

Therefore we have:

Theorem 3.41. *The characteristic polynomial is invariant under a change of basis; that is, under a similarity transformation of the matrix.*

In particular, each of the coefficients c_i of the characteristic polynomial $P(x) = c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0$ remains unchanged after a similarity transformation of the matrix A.

What is the coefficient c_n? Looking at the Leibniz formula, we see that the term x^n can only occur in the product

$$(a_{11} - x)(a_{22} - x)\cdots(a_{nn} - x) = (-1)^n x^n + (-1)^{n-1}(a_{11} + a_{22} + \cdots + a_{nn})x^{n-1} + \cdots .$$

Therefore $c_n = 1$ if n is even, and $c_n = -1$ if n is odd. This is not particularly interesting.

So let us go one term lower and look at the coefficient c_{n-1}. Where does x^{n-1} occur in the Leibniz formula? Well, as we have just seen, there certainly is the term

$$(-1)^{n-1}(a_{11} + a_{22} + \cdots + a_{nn})x^{n-1},$$

which comes from the product of the diagonal elements in the matrix $A - xI_n$. Do any other terms also involve the power x^{n-1}? Let us look at Leibniz formula more carefully in this situation. We have

$$\begin{aligned} det(A - xI_n) &= (a_{11} - x)(a_{22} - x)\cdots(a_{nn} - x) \\ &\quad + \sum_{\substack{\sigma \in S_n \\ \sigma \neq id}} sign(\sigma) \prod_{i=1}^{n} \left(a_{\sigma(i)i} - x\delta_{\sigma(i)i}\right) \end{aligned}$$

Here, $\delta_{ij} = 1$ if $i = j$. Otherwise, $\delta_{ij} = 0$. Now if σ is a *non-trivial* permutation — not just the identity mapping — then obviously we must have two *different* numbers i_1 and i_2, with $\sigma(i_1) \neq i_1$ and also $\sigma(i_2) \neq i_2$. Therefore we see that these further terms in the sum can only contribute at most $n - 2$ powers of x. So we conclude that the $(n-1)$-st coefficient is

$$c_{n-1} = (-1)^{n-1}(a_{11} + a_{22} + \cdots + a_{nn}).$$

Definition. *Let* $A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$ *be an* $n \times n$ *matrix. The* trace *of* A *(in German: the "*Spur*" of* A*) is the sum of the diagonal elements:*

$$tr(A) = a_{11} + a_{22} + \cdots + a_{nn}.$$

Theorem 3.42. $tr(A)$ *remains unchanged under a similarity transformation.*

An example

Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a rotation through the angle θ. Then, with respect to the canonical basis of $\mathbb{R}^2$, the matrix of f is

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Therefore the characteristic polynomial of A is

$$\begin{aligned} det\left[\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} - x\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right] &= det\begin{pmatrix} \cos\theta - x & -\sin\theta \\ \sin\theta & \cos\theta - x \end{pmatrix} \\ &= x^2 - 2x\cos\theta + 1. \end{aligned}$$

That is to say, if $\lambda \in \mathbb{R}$ is an eigenvalue of f, then λ must be a zero of the characteristic polynomial. That is,

$$\lambda^2 - 2\lambda\cos\theta + 1 = 0.$$

But, looking at the well-known formula for the roots of quadratic polynomials, we see that such a λ can only exist if $|\cos\theta| = 1$. That is, $\theta = 0$ or π. This reflects the obvious geometric fact that a rotation through any angle other than 0 or π rotates any vector away from its original axis. In any case, the two possible values of θ give the two possible eigenvalues for f, namely $+1$ and -1.

3.15 Scalar products, norms, etc.

From now on, we will assume that all vector spaces considered are either *Euclidean*, or else *unitary* vector spaces. That is, we assume that they are defined either over the field of real numbers $\mathbb{R}$ —

so that the vector space is of the form $\mathbb{R}^n$, for some $n \in \mathbb{N}$ — or else over the field of complex numbers, giving $\mathbb{C}^n$. Within these vector spaces, it makes sense to define the idea of a "distance" between two vectors. More particularly, we have the idea of the "length" of a given vector.

So let $\mathbf{V}$ be some finite dimensional vector space over $\mathbb{R}$, or $\mathbb{C}$. Let $\mathbf{v} \in \mathbf{V}$ be some vector in $\mathbf{V}$. Then, since $\mathbf{V} \cong \mathbb{R}^n$, or $\mathbb{C}^n$, we can write $\mathbf{v} = \sum_{j=1}^{n} a_j \mathbf{e}_j$, where $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is the canonical basis for $\mathbb{R}^n$ or $\mathbb{C}^n$, and $a_j \in \mathbb{R}$ or $\mathbb{C}$, respectively, for all j. Then the *length* of $\mathbf{v}$ is defined to be the non-negative real number

$$\|\mathbf{v}\| = \sqrt{|a_1|^2 + \cdots + |a_n|^2}.$$

But in mathematics, we wish to extend this concept beyond the normal idea of simply defining lengths in the normal Euclidean space. Infinite dimensional spaces, or even abstract spaces of functions should also be dealt with.

Definition. *Let $F = \mathbb{R}$ or $\mathbb{C}$ and let $\mathbf{V}$, $\mathbf{W}$ be two vector spaces over F. A* bilinear form *is a mapping $s : \mathbf{V} \times \mathbf{W} \to F$ satisfying the following conditions with respect to arbitrary elements $\mathbf{v}$, $\mathbf{v}_1$ and $\mathbf{v}_2 \in \mathbf{V}$, $\mathbf{w}$, $\mathbf{w}_1$ and $\mathbf{w}_2 \in \mathbf{W}$, and $a \in F$.*

1. $s(\mathbf{v}_1 + \mathbf{v}_2, \mathbf{w}) = s(\mathbf{v}_1, \mathbf{w}) + s(\mathbf{v}_2, \mathbf{w})$,
2. $s(a\mathbf{v}, \mathbf{w}) = as(\mathbf{v}, \mathbf{w})$,
3. $s(\mathbf{v}, \mathbf{w}_1 + \mathbf{w}_2) = s(\mathbf{v}, \mathbf{w}_1) + s(\mathbf{v}, \mathbf{w}_2)$ *and*
4. $s(\mathbf{v}, a\mathbf{w}) = as(\mathbf{v}, \mathbf{w})$.

If $\mathbf{V} = \mathbf{W}$, then we say that a bilinear form $s : \mathbf{V} \times \mathbf{V} \to F$ is symmetric*, if we always have $s(\mathbf{v}_1, \mathbf{v}_2) = s(\mathbf{v}_2, \mathbf{v}_1)$. Also the form is called* positive definite *if $s(\mathbf{v}, \mathbf{v}) > 0$ for all $\mathbf{v} \neq \mathbf{0}$.*

On the other hand, if $F = \mathbb{C}$ and $f : \mathbf{V} \to \mathbf{W}$ is such that we always have

1. $f(\mathbf{v}_1 + \mathbf{v}_2) = f(\mathbf{v}_1) + f(\mathbf{v}_2)$ *and*
2. $f(a\mathbf{v}) = \overline{a} f(\mathbf{v})$

Then f is a semi-linear *(not a linear) mapping. (Note: if $F = \mathbb{R}$ then semi-linear is the same as linear.)*

A mapping $s : \mathbf{V} \times \mathbf{W} \to F$ such that

1. *The mapping given by $s(\cdot, \mathbf{w}) : \mathbf{V} \to F$, where $\mathbf{v} \to s(\mathbf{v}, \mathbf{w})$ is semi-linear for all $\mathbf{w} \in \mathbf{W}$, whereas*
2. *The mapping given by $s(\mathbf{v}, \cdot) : \mathbf{W} \to F$, where $\mathbf{w} \to s(\mathbf{v}, \mathbf{w})$ is linear for all $\mathbf{v} \in \mathbf{V}$*

is called a sesqui-linear *form.*

In the case $\mathbf{V} = \mathbf{W}$, we say that the sesqui-linear form is Hermitian *(or* Euclidean*, if we only have $F = \mathbb{R}$), if we always have $s(\mathbf{v}_1, \mathbf{v}_2) = \overline{s(\mathbf{v}_2, \mathbf{v}_1)}$. (Therefore, if $F = \mathbb{R}$, an Hermitian form is symmetric.)*

Finally, a scalar product *is a positive definite Hermitian form $s : \mathbf{V} \times \mathbf{V} \to F$. Normally, one writes $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle$, rather than $s(\mathbf{v}_1, \mathbf{v}_2)$.*

Well, these are a lot of new words. To be more concrete, we have the *inner products*, which are examples of scalar products.

Inner products

Let $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{C}^n$. Thus, we are considering these vectors as column vectors, defined with respect to the canonical basis of $\mathbb{C}^n$. Then define (using matrix multiplication)

$$\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\mathbf{u}}^t \mathbf{v} = (\overline{u}_1 \quad \overline{u}_2 \quad \cdots \quad \overline{u}_n) \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \sum_{j=1}^{n} \overline{u}_j v_j.$$

It is easy to check that this gives a scalar product on $\mathbb{C}^n$. This particular scalar product is called the *inner product*.

Remark. *One often writes* $\mathbf{u} \cdot \mathbf{v}$ *for the inner product. Thus, considering it to be a scalar product, we just have* $\mathbf{u} \cdot \mathbf{v} = \langle \mathbf{u}, \mathbf{v} \rangle$.

Definition. *A real vector space (that is, over the field of the real numbers* $\mathbb{R}$*), together with a scalar product is called a* Euclidean *vector space. A complex vector space with scalar product is called a* unitary *vector space.*

Now, the basic reason for making all these definitions is that we want to define the length — that is the norm — of the vectors in $\mathbf{V}$. Given a scalar product, then the norm of $\mathbf{v} \in \mathbf{V}$ — with respect to this scalar product — is the non-negative real number

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}.$$

More generally, one defines a norm-function on a vector space in the following way.

Definition. *Let* $\mathbf{V}$ *be a vector space over* $\mathbb{C}$ *(and thus we automatically also include the case* $\mathbb{R} \subset \mathbb{C}$ *as well). A function* $\| \cdot \| : \mathbf{V} \to \mathbb{R}$ *is called a* norm *on* $\mathbf{V}$ *if it satisfies the following conditions.*

1. $\|a\mathbf{v}\| = |a|\|\mathbf{v}\|$ *for all* $\mathbf{v} \in \mathbf{V}$ *and for all* $a \in \mathbb{C}$*,*
2. $\|\mathbf{v}_1 + \mathbf{v}_2\| \leq \|\mathbf{v}_1\| + \|\mathbf{v}_2\|$ *for all* $\mathbf{v}_1, \mathbf{v}_2 \in \mathbf{V}$ *(the* triangle inequality*), and*
3. $\|\mathbf{v}\| = 0 \Leftrightarrow \mathbf{v} = \mathbf{0}$.

Theorem 3.43 (Cauchy-Schwarz inequality)**.** *Let* $\mathbf{V}$ *be a Euclidean or a unitary vector space, and let* $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$ *for all* $\mathbf{v} \in \mathbf{V}$*. Then we have*

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \cdot \|\mathbf{v}\|$$

for all $\mathbf{u}$ *and* $\mathbf{v} \in \mathbf{V}$*. Furthermore, the equality* $|\langle \mathbf{u}, \mathbf{v} \rangle| = \|\mathbf{u}\| \cdot \|\mathbf{v}\|$ *holds if, and only if, the set* $\{\mathbf{u}, \mathbf{v}\}$ *is linearly* dependent*.*

Proof. It suffices to show that $|\langle \mathbf{u}, \mathbf{v} \rangle|^2 \leq \langle \mathbf{u}, \mathbf{u} \rangle \langle \mathbf{v}, \mathbf{v} \rangle$. Now, if $\mathbf{v} = \mathbf{0}$, then — using the properties of the scalar product — we have both $\langle \mathbf{u}, \mathbf{v} \rangle = 0$ and $\langle \mathbf{v}, \mathbf{v} \rangle = 0$. Therefore the theorem is true in this case, and we may assume that $\mathbf{v} \neq \mathbf{0}$. Thus $\langle \mathbf{v}, \mathbf{v} \rangle > 0$. Let

$$a = \frac{\overline{\langle \mathbf{u}, \mathbf{v} \rangle}}{\langle \mathbf{v}, \mathbf{v} \rangle} \in \mathbb{C}.$$

Then we have

$$\begin{aligned} 0 &\leq \langle \mathbf{u} - a\mathbf{v}, \mathbf{u} - a\mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} - a\mathbf{v} \rangle + \langle -a\mathbf{v}, \mathbf{u} - a\mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, -a\mathbf{v} \rangle + \langle -a\mathbf{v}, \mathbf{u} \rangle + \langle -a\mathbf{v}, -a\mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle - \underbrace{a\langle \mathbf{u}, \mathbf{v} \rangle}_{\frac{\overline{\langle \mathbf{u},\mathbf{v} \rangle}\langle \mathbf{u},\mathbf{v} \rangle}{\langle \mathbf{v},\mathbf{v} \rangle}} - \underbrace{\overline{a}\overline{\langle \mathbf{u}, \mathbf{v} \rangle}}_{\frac{\langle \mathbf{u},\mathbf{v} \rangle\overline{\langle \mathbf{u},\mathbf{v} \rangle}}{\langle \mathbf{v},\mathbf{v} \rangle}} + \underbrace{a\overline{a}\langle \mathbf{v}, \mathbf{v} \rangle}_{\frac{\langle \mathbf{u},\mathbf{v} \rangle\overline{\langle \mathbf{u},\mathbf{v} \rangle}}{\langle \mathbf{v},\mathbf{v} \rangle}}. \end{aligned}$$

Therefore,

$$0 \leq \langle \mathbf{u}, \mathbf{u} \rangle \langle \mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle \overline{\langle \mathbf{u}, \mathbf{v} \rangle}.$$

But

$$\langle \mathbf{u}, \mathbf{v} \rangle \overline{\langle \mathbf{u}, \mathbf{v} \rangle} = |\langle \mathbf{u}, \mathbf{v} \rangle|^2,$$

which gives the Cauchy-Schwarz inequality. When do we have equality?

If $\mathbf{v} = \mathbf{0}$ then, as we have already seen, the equality $|\langle \mathbf{u}, \mathbf{v} \rangle| = \|\mathbf{u}\| \cdot \|\mathbf{v}\|$ is trivially true. On the other hand, when $\mathbf{v} \neq \mathbf{0}$, then equality holds when $\langle \mathbf{u} - a\mathbf{v}, \mathbf{u} - a\mathbf{v} \rangle = 0$. But since the scalar product is positive definite, this holds when $\mathbf{u} - a\mathbf{v} = \mathbf{0}$. So in this case as well, $\{\mathbf{u}, \mathbf{v}\}$ is linearly dependent. □

Theorem 3.44. *Let $\mathbf{V}$ be a vector space with scalar product, and define the non-negative function $\| \cdot \| : \mathbf{V} \to \mathbb{R}$ by $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$. Then $\| \cdot \|$ is a norm function on $\mathbf{V}$.*

Proof. The first and third properties in our definition of norms are obviously satisfied. As far as the triangle inequality is concerned, begin by observing that for arbitrary complex numbers $z = x + yi \in \mathbb{C}$ we have

$$z + \overline{z} = (x + yi) + (x - yi) = 2x \leq 2|x| \leq 2|z|.$$

Therefore, let $\mathbf{u}$ and $\mathbf{v} \in \mathbf{V}$ be chosen arbitrarily. Then we have

$$\begin{aligned} \|\mathbf{u} + \mathbf{v}\|^2 &= \langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, \mathbf{v} \rangle + \overline{\langle \mathbf{u}, \mathbf{v} \rangle} + \langle \mathbf{v}, \mathbf{v} \rangle \\ &\leq \langle \mathbf{u}, \mathbf{u} \rangle + 2|\langle \mathbf{u}, \mathbf{v} \rangle| + \langle \mathbf{v}, \mathbf{v} \rangle \\ &\leq \langle \mathbf{u}, \mathbf{u} \rangle + 2\|\mathbf{u}\| \cdot \|\mathbf{v}\| + \langle \mathbf{v}, \mathbf{v} \rangle \quad \text{(Cauchy-Schwarz inequality)} \\ &= \|\mathbf{u}\|^2 + 2\|\mathbf{u}\| \cdot \|\mathbf{v}\| + \|\mathbf{v}\|^2 \\ &= (\|\mathbf{u}\| + \|\mathbf{v}\|)^2. \end{aligned}$$

Therefore $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$. □

3.16 Orthonormal bases

Our vector space $\mathbf{V}$ is now assumed to be either Euclidean, or else unitary — that is, it is defined over either the real numbers $\mathbb{R}$, or else the complex numbers $\mathbb{C}$. In either case we have a scalar product $\langle \cdot, \cdot \rangle : \mathbf{V} \times \mathbf{V} \to F$ (here, $F = \mathbb{R}$ or $\mathbb{C}$).

As always, we assume that $\mathbf{V}$ is finite dimensional, and thus it has a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. Thinking about the *canonical* basis for $\mathbb{R}^n$ or $\mathbb{C}^n$, and the *inner product* as our scalar product, we see that it would be nice if we had

- $\langle \mathbf{v}_j, \mathbf{v}_j \rangle = 1$, for all j (that is, the basis vectors are *normalized*), and furthermore
- $\langle \mathbf{v}_j, \mathbf{v}_k \rangle = 0$, for all $j \neq k$ (that is, the basis vectors are an *orthogonal* set in $\mathbf{V}$).[8]

That is to say, $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an *orthonormal basis* of $\mathbf{V}$. Unfortunately, most bases are not orthonormal. But this doesn't really matter. For, starting from any given basis, we can successively alter the vectors in it, gradually changing it into an orthonormal basis. This process is often called the *Gram-Schmidt orthonormalization process*. But first, to show you why orthonormal bases are good, we have the following theorem.

Theorem 3.45. *Let $\mathbf{V}$ have the orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, and let $\mathbf{x} \in \mathbf{V}$ be arbitrary. Then*

$$\mathbf{x} = \sum_{j=1}^{n} \langle \mathbf{v}_j, \mathbf{x} \rangle \mathbf{v}_j.$$

That is, the coefficients of $\mathbf{x}$, with respect to the orthonormal basis, are simply the scalar products with the respective basis vectors.

Proof. This follows simply because if $\mathbf{x} = \sum_{j=1}^{n} a_j \mathbf{v}_j$, then we have for each k,

$$\langle \mathbf{v}_k, \mathbf{x} \rangle = \langle \mathbf{v}_k, \sum_{j=1}^{n} a_j \mathbf{v}_j \rangle = \sum_{j=1}^{n} a_j \langle \mathbf{v}_k, \mathbf{v}_j \rangle = a_k.$$

□

So now to the Gram-Schmidt process. To begin with, if a non-zero vector $\mathbf{v} \in \mathbf{V}$ is not normalized — that is, its norm is not one — then it is easy to multiply it by a scalar, changing it into a vector with norm one. For we have $\langle \mathbf{v}, \mathbf{v} \rangle > 0$. Therefore $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} > 0$ and we have

$$\left\| \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\| = \sqrt{\left\langle \frac{\mathbf{v}}{\|\mathbf{v}\|}, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle} = \sqrt{\frac{\langle \mathbf{v}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle}} = \frac{\|\mathbf{v}\|}{\|\mathbf{v}\|} = 1.$$

In other words, we simply multiply the vector by the inverse of its norm.

Theorem 3.46. *Every finite dimensional vector space $\mathbf{V}$ which has a scalar product has an orthonormal basis.*

Proof. The proof proceeds by constructing an orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ from a given, arbitrary basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. To describe the construction, we use induction on the dimension, n. If $n = 1$ then there is almost nothing to prove. Any non-zero vector is a basis for $\mathbf{V}$, and as we have seen, it can be normalized by dividing by the norm. (That is, scalar multiplication with the inverse of the norm.)

So now assume that $n \geq 2$, and furthermore assume that the Gram-Schmidt process can be constructed for any $n - 1$ dimensional space. Let $\mathbf{U} \subset \mathbf{V}$ be the subspace spanned by the first

[8] Note that *any* orthogonal set of non-zero vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ in $\mathbf{V}$ is linearly independent. This follows because if

$$\mathbf{0} = \sum_{j=1}^{m} a_j \mathbf{u}_j$$

then

$$0 = \langle \mathbf{u}_k, \mathbf{0} \rangle = \langle \mathbf{u}_k, \sum_{j=1}^{m} a_j \mathbf{u}_j \rangle = \sum_{j=1}^{m} a_j \langle \mathbf{u}_k, \mathbf{u}_j \rangle = a_k \langle \mathbf{u}_k, \mathbf{u}_k \rangle$$

since $\langle \mathbf{u}_k, \mathbf{u}_j \rangle = 0$ if $j \neq k$, and otherwise it is not zero. Thus, we must have $a_k = 0$. This is true for *all* the a_k.

$n-1$ basis vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}\}$. Since $\mathbf{U}$ is only $n-1$ dimensional, our assumption is that there exists an orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}\}$ for $\mathbf{U}$. Clearly[9], adding in $\mathbf{v}_n$ gives a new basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}, \mathbf{v}_n\}$ for $\mathbf{V}$. Unfortunately, this last vector, $\mathbf{v}_n$, might disturb the nice orthonormal character of the other vectors. Therefore, we *replace* $\mathbf{v}_n$ with the new vector[10]

$$\mathbf{u}_n^* = \mathbf{v}_n - \sum_{j=1}^{n-1} \langle \mathbf{u}_j, \mathbf{v}_n \rangle \mathbf{u}_j.$$

Thus the new set $\{\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}, \mathbf{u}_n^*\}$ is a basis of $\mathbf{V}$. Also, for $k < n$, we have

$$\begin{aligned} \langle \mathbf{u}_k, \mathbf{u}_n^* \rangle &= \left\langle \mathbf{u}_k, \mathbf{v}_n - \sum_{j=1}^{n-1} \langle \mathbf{u}_j, \mathbf{v}_n \rangle \mathbf{u}_j \right\rangle \\ &= \langle \mathbf{u}_k, \mathbf{v}_n \rangle - \sum_{j=1}^{n-1} \langle \mathbf{u}_j, \mathbf{v}_n \rangle \langle \mathbf{u}_k, \mathbf{u}_j \rangle \\ &= \langle \mathbf{u}_k, \mathbf{v}_n \rangle - \langle \mathbf{u}_k, \mathbf{v}_n \rangle = 0. \end{aligned}$$

Thus the basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}, \mathbf{u}_n^*\}$ is orthogonal. Perhaps $\mathbf{u}_n^*$ is not normalized, but as we have seen, this can be easily changed by taking the normalized vector

$$\mathbf{u}_n = \frac{\mathbf{u}_n^*}{\|\mathbf{u}_n^*\|}.$$

□

3.17 Orthogonal, unitary and self-adjoint linear mappings

Definition. *A finite dimensional vector space is called a* Euclidean *vector space if it is defined over the real numbers $\mathbb{R}$; if it is defined over the complex numbers $\mathbb{C}$, then it is called a* unitary *vector space. Thus a Euclidean vector space is isomorphic with $\mathbb{R}^n$, for some $n \in \mathbb{N}$; a unitary vector space is isomorphic with some $\mathbb{C}^n$.*

But this definition, using the word "unitary", should not be confused with the the definition of unitary *mappings*. Both orthogonal and unitary mappings use the fact that both $\mathbb{R}^n$ and $\mathbb{C}^n$ have a natural scalar product, thus giving us an idea of the length of a vector. So these mappings are such that the lengths of vectors are not changed under the mappings. Orthogonal and Unitary mappings can be thought of as *length-preserving* mappings of $\mathbb{R}^n$, and $\mathbb{C}^n$ respectively, into themselves.

Definition. *Let $\mathbf{V}$ be a Euclidean vector space. The linear mapping $f : \mathbf{V} \to \mathbf{V}$ is an* orthogonal *mapping if*

$$\langle \mathbf{v}, \mathbf{w} \rangle = \langle f(\mathbf{v}), f(\mathbf{w}) \rangle,$$

for all $\mathbf{v}$, $\mathbf{w} \in \mathbf{V}$. Similarly, if $\mathbf{V}$ is a unitary vector space, then linear mapping $f : \mathbf{V} \to \mathbf{V}$ is called a unitary *mapping if*

$$\langle \mathbf{v}, \mathbf{w} \rangle = \langle f(\mathbf{v}), f(\mathbf{w}) \rangle,$$

for all $\mathbf{v}$, $\mathbf{w} \in \mathbf{V}$.

[9] Since both $\{\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}\}$ and $\{\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}\}$ are bases for $\mathbf{U}$, we can write each $\mathbf{v}_j$ as a linear combination of the $\mathbf{u}_k$'s. Therefore $\{\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}, \mathbf{v}_n\}$ spans $\mathbf{V}$, and since the dimension is n, it must be a basis.

[10] A linearly independent set remains linearly independent if one of the vectors has some linear combination of the *other* vectors added on to it.

Obviously, since we have defined the "length" of a vector $\mathbf{v}$ to be $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$, it follows that under an orthogonal or unitary mapping f, we must have $\|\mathbf{v}\| = \|f(\mathbf{v})\|$, for all $\mathbf{v} \in \mathbf{V}$. But we also have the following theorem.

Theorem 3.47. *Let $\mathbf{V}$ be a Euclidean or a unitary vector space, and let $f : \mathbf{V} \to \mathbf{V}$ be such that $\|\mathbf{v}\| = \|f(\mathbf{v})\|$, for all $\mathbf{v} \in \mathbf{V}$. Then f is Euclidean, or unitary respectively.*

Proof. I will give the proof in the case that $\mathbf{V}$ is a Euclidean vector space. The proof in the unitary case is left as an exercise.

Let $\mathbf{u}$, $\mathbf{v} \in \mathbf{V}$ be given. Then we have

$$\langle f(\mathbf{u}+\mathbf{v}), f(\mathbf{u}+\mathbf{v}) \rangle = \langle \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v} \rangle,$$

and, of course both

$$\langle f(\mathbf{u}), f(\mathbf{u}) \rangle = \langle \mathbf{u}, \mathbf{u} \rangle$$

and

$$\langle f(\mathbf{v}), f(\mathbf{v}) \rangle = \langle \mathbf{v}, \mathbf{v} \rangle.$$

But

$$\langle f(\mathbf{u}+\mathbf{v}), f(\mathbf{u}+\mathbf{v}) \rangle = \langle f(\mathbf{u}), f(\mathbf{u}) \rangle + 2\langle f(\mathbf{u}), f(\mathbf{v}) \rangle + \langle f(\mathbf{v}), f(\mathbf{v}) \rangle.$$

Similarly,

$$\langle \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle + 2\langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle.$$

Therefore, it follows that

$$\langle f(\mathbf{u}), f(\mathbf{v}) \rangle = \langle \mathbf{u}, \mathbf{v} \rangle.$$

□

Theorem 3.48. *Let $f : \mathbf{V} \to \mathbf{V}$ be either orthogonal or unitary, and let λ be an eigenvalue of f. Then $|\lambda| = 1$.*

Proof. Take $\mathbf{v} \in \mathbf{V}$ to be an eigenvector corresponding with the eigenvalue λ, so that $f(\mathbf{v}) = \lambda \mathbf{v}$. We may assume that $\|\mathbf{v}\| = 1$; that is $\langle \mathbf{v}, \mathbf{v} \rangle = 1$. Then we have

$$1 = \langle \mathbf{v}, \mathbf{v} \rangle = \langle f(\mathbf{v}), f(\mathbf{v}) \rangle = \langle \lambda \mathbf{v}, \lambda \mathbf{v} \rangle = \overline{\lambda}\lambda \langle \mathbf{v}, \mathbf{v} \rangle = \overline{\lambda}\lambda = |\lambda|^2.$$

□

Finally, we have a further definition involving linear mappings and scalar products.

Definition. *Let the linear mapping $f : \mathbf{V} \to \mathbf{V}$ be such that*

$$\langle \mathbf{u}, f(\mathbf{v}) \rangle = \langle f(\mathbf{u}), \mathbf{v} \rangle,$$

for all $\mathbf{u}$, $\mathbf{v} \in \mathbf{V}$. Then the mapping f is a self-adjoint *mapping.*

3.18 Characterizing orthogonal, unitary, and Hermitian matrices

3.18.1 Orthogonal matrices

Let $\mathbf{V}$ be an n-dimensional real vector space (that is, over the real numbers $\mathbb{R}$), and let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be an orthonormal basis for $\mathbf{V}$. Let $f : \mathbf{V} \to \mathbf{V}$ be an orthogonal mapping, and let A be its matrix with respect to the basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. Then we say that A is an *orthogonal matrix*.

Theorem 3.49. *The $n \times n$ matrix A is orthogonal $\Leftrightarrow A^{-1} = A^t$. (Recall that if a_{ij} is the ij-th element of A, then the ij-the element of A^t is a_{ji}. That is, everything is "flipped over" the main diagonal in A.)*

Proof. For an orthogonal mapping f, we have $\langle \mathbf{u}, \mathbf{w} \rangle = \langle f(\mathbf{u}), f(\mathbf{w}) \rangle$, for all j and k. But in the matrix notation, the scalar product becomes the inner product. That is, if

$$\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \quad \text{and} \quad \mathbf{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix},$$

then

$$\langle \mathbf{u}, \mathbf{w} \rangle = \mathbf{u}^t \cdot \mathbf{w} = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \sum_{j=1}^{n} u_j w_j.$$

In particular, taking $\mathbf{u} = \mathbf{v}_j$ and $\mathbf{w} = \mathbf{v}_k$, we have

$$\langle \mathbf{v}_j, \mathbf{v}_k \rangle = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{otherwise.} \end{cases}$$

In other words, the matrix whose jk-th element is always $\langle \mathbf{v}_j, \mathbf{v}_k \rangle$ is the $n \times n$ identity matrix I_n. On the other hand,

$$f(\mathbf{v}_j) = A\mathbf{v}_j = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \cdot \mathbf{v}_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix}.$$

That is, we obtain the j-th column of the matrix A. Furthermore, since $\langle \mathbf{v}_j, \mathbf{v}_k \rangle = \langle f(\mathbf{v}_j), f(\mathbf{v}_k) \rangle$, we must have the matrix whose jk-th elements are $\langle f(\mathbf{v}_j), f(\mathbf{v}_k) \rangle$ being again the identity matrix. So

$$\begin{pmatrix} a_{1j} & \cdots & a_{nj} \end{pmatrix} \begin{pmatrix} a_{1k} \\ \vdots \\ a_{nk} \end{pmatrix} = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{otherwise.} \end{cases}$$

But now, if you think about it, you see that this is just one part of the matrix multiplication $A^t A$. All together, we have

$$A^t A = \begin{pmatrix} a_{11} & \cdots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{nn} \end{pmatrix} \cdot \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = I_n.$$

Thus we conclude that $A^{-1} = A^t$. (Note: this was only the proof that f orthogonal $\Rightarrow A^{-1} = A^t$. The proof in the other direction, going backwards through our argument, is easy, and is left as an exercise for you.) $\square$

3.18.2 Unitary matrices

Theorem 3.50. *The $n \times n$ matrix A is unitary $\Leftrightarrow A^{-1} = \overline{A}^t$. (The matrix $\overline{A}$ is obtained by taking the complex conjugates of all its elements.)*

Proof. Entirely analogous with the case of orthogonal matrices. One must note however, that the inner product in the complex case is

$$\langle \mathbf{u}, \mathbf{w} \rangle = \overline{\mathbf{u}}^t \cdot \mathbf{w} = (\overline{u}_1 \quad \cdots \quad \overline{u}_n) \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \sum_{j=1}^{n} \overline{u}_j w_j.$$

□

3.18.3 Hermitian and symmetric matrices

Finally, we say that a matrix is *Hermitian* if it represents a self-adjoint mapping $f : \mathbf{V} \to \mathbf{V}$ with respect to an orthonormal basis of $\mathbf{V}$.

Theorem 3.51. *The $n \times n$ matrix A is Hermitian $\Leftrightarrow A = \overline{A}^t$.*

Proof. This is again a matter of translating the condition $\langle \mathbf{v}_j, f(\mathbf{v}_k) \rangle = \langle f(\mathbf{v}_j), \mathbf{v}_k \rangle$ into matrix notation, where f is the linear mapping which is represented by the matrix A, with respect to the orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. We have

$$\langle \mathbf{v}_j, f(\mathbf{v}_k) \rangle = \overline{\mathbf{v}}_j^t \cdot A\mathbf{v}_k = \overline{\mathbf{v}}_j^t \begin{pmatrix} a_{1k} \\ \vdots \\ a_{n_k} \end{pmatrix} = a_{jk}.$$

On the other hand

$$\langle f(\mathbf{v}_j), \mathbf{v}_k \rangle = \overline{A\mathbf{v}_j}^t \cdot \mathbf{v}_k = (\overline{a}_{1j} \quad \cdots \quad \overline{a}_{nj}) \cdot \mathbf{v}_k = \overline{a}_{kj}.$$

□

In particular, we see that in the real case, self-adjoint matrices are symmetric.

3.19 Which matrices can be diagonalized?

The complete answer to this question is a bit too complicated for me to explain to you in the short time we have in this semester. It all has to do with a thing called the "minimal polynomial".

Now we have seen that not all *orthogonal* matrices can be diagonalized. (Think about the rotations of $\mathbb{R}^2$.) On the other hand, we can prove that all *unitary*, and also all *Hermitian* matrices can be diagonalized.

Of course, a matrix M is only a representation of a linear mapping $f : \mathbf{V} \to \mathbf{V}$ with respect to a given basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of the vector space $\mathbf{V}$. So the idea that the matrix can be diagonalized is that it is similar to a diagonal matrix. That is, there exists another matrix S, such that $S^{-1}MS$ is diagonal.

$$S^{-1}MS = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

But this means that there must be a basis for $\mathbf{V}$, consisting entirely of eigenvectors.

In this section we will consider complex vector spaces — that is, $\mathbf{V}$ is a vector space over the complex numbers $\mathbb{C}$. The vector space $\mathbf{V}$ will be assumed to have a scalar product associated with it, and the bases we consider will be orthonormal.

We begin with a definition.

Definition. *Let $\mathbf{W} \subset \mathbf{V}$ be a subspace of $\mathbf{V}$. Let*

$$\mathbf{W}^\perp = \{\mathbf{v} \in \mathbf{V} : \langle \mathbf{v}, \mathbf{w} \rangle = 0, \forall \mathbf{w} \in \mathbf{W}\}.$$

Then $\mathbf{W}^\perp$ is called the perpendicular space *to $\mathbf{W}$.*

It is a rather trivial matter to verify that $\mathbf{W}^\perp$ is itself a subspace of $\mathbf{V}$, and furthermore $\mathbf{W} \cap \mathbf{W}^\perp = \{\mathbf{0}\}$. In fact, we have:

Theorem 3.52. $\mathbf{V} = \mathbf{W} \oplus \mathbf{W}^\perp$.

Proof. Let $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ be some orthonormal basis for the vector space $\mathbf{W}$. This can be extended to a basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{w}_{m+1}, \ldots, \mathbf{w}_n\}$ of $\mathbf{V}$. Assuming the Gram-Schmidt process has been used, we may assume that this is an orthonormal basis. The claim is then that $\{\mathbf{w}_{m+1}, \ldots, \mathbf{w}_n\}$ is a basis for $\mathbf{W}^\perp$.

Now clearly, since $\langle \mathbf{w}_j, \mathbf{w}_k \rangle = 0$, for $j \neq k$, we have that $\{\mathbf{w}_{m+1}, \ldots, \mathbf{w}_n\} \subset \mathbf{W}^\perp$. If $\mathbf{u} \in \mathbf{W}^\perp$ is some arbitrary vector in $\mathbf{W}^\perp$, then we have

$$\mathbf{u} = \sum_{j=1}^{n} \langle \mathbf{w}_j, \mathbf{u} \rangle \mathbf{w}_j = \sum_{j=m+1}^{n} \langle \mathbf{w}_j, \mathbf{u} \rangle \mathbf{w}_j,$$

since $\langle \mathbf{w}_j, \mathbf{u} \rangle = 0$ if $j \leq m$. (Remember, $\mathbf{u} \in \mathbf{W}^\perp$.) Therefore, $\{\mathbf{w}_{m+1}, \ldots, \mathbf{w}_n\}$ is a linearly independent, orthonormal set which generates $\mathbf{W}^\perp$, so it is a basis. And so we have $\mathbf{V} = \mathbf{W} \oplus \mathbf{W}^\perp$. □

Theorem 3.53. *Let $f : \mathbf{V} \to \mathbf{V}$ be a unitary mapping ($\mathbf{V}$ is a vector space over the complex numbers $\mathbb{C}$). Then there exists an orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ for $\mathbf{V}$ consisting of eigenvectors under f. That is to say, the matrix of f with respect to this basis is a diagonal matrix.*

Proof. If the dimension of $\mathbf{V}$ is zero or one, then obviously there is nothing to prove. So let us assume that the dimension n is at least two, and we prove things by induction on the number n. That is, we assume that the theorem is true for spaces of dimension less than n.

Now, according to the fundamental theorem of algebra[11], the characteristic polynomial of f has a zero, λ say, which is then an eigenvalue for f. So there must be some non-zero vector $\mathbf{v}_n \in \mathbf{V}$, with $f(\mathbf{v}_n) = \lambda \mathbf{v}_n$. By dividing by the norm of $\mathbf{v}_n$ if necessary, we may assume that $\|\mathbf{v}_n\| = 1$.

Let $\mathbf{W} \subset \mathbf{V}$ be the 1-dimensional subspace generated by the vector $\mathbf{v}_n$. Then $\mathbf{W}^\perp$ is an $n-1$ dimensional subspace. We have that $\mathbf{W}^\perp$ is invariant under f. That is, if $\mathbf{u} \in \mathbf{W}^\perp$ is some arbitrary vector, then $f(\mathbf{u}) \in \mathbf{W}^\perp$ as well. This follows since

$$\lambda \langle f(\mathbf{u}), \mathbf{v}_n \rangle = \langle f(\mathbf{u}), \lambda \mathbf{v}_n \rangle = \langle f(\mathbf{u}), f(\mathbf{v}_n) \rangle = \langle \mathbf{u}, \mathbf{v}_n \rangle = 0.$$

But we have already seen that for an eigenvalue λ of a unitary mapping, we must have $|\lambda| = 1$. Therefore we must have $\langle f(\mathbf{u}), \mathbf{v}_n \rangle = 0$.

[11] This says that every polynomial $P(z) \in \mathbb{C}[z]$ of degree at least one has a zero. That is, there exists some $\lambda \in \mathbb{C}$ with $P(\lambda) = 0$. The proof is not difficult, but it takes too much time to explain in this lecture.

So we can consider f, restricted to $\mathbf{W}^\perp$, and using the inductive hypothesis, we obtain an orthonormal basis of eigenvectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}\}$ for $\mathbf{W}^\perp$. Therefore, adding in the last vector $\mathbf{v}_n$, we have an orthonormal basis of eigenvectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ for $\mathbf{V}$. □

Theorem 3.54. *All Hermitian matrices can be diagonalized.*

Proof. This is similar to the last one. Again, we use induction on n, the dimension of the vector space $\mathbf{V}$. We have a self-adjoint mapping $f : \mathbf{V} \to \mathbf{V}$. If n is zero or one, then we are finished. Therefore we assume that $n \geq 2$.

Again, we observe that the characteristic polynomial of f must have a zero, hence there exists some eigenvalue λ, and an eigenvector $\mathbf{v}_n$ of f, which has norm equal to one, where $f(\mathbf{v}_n) = \lambda \mathbf{v}_n$. Again take $\mathbf{W}$ to be the one dimensional subspace of $\mathbf{V}$ generated by $\mathbf{v}_n$. Let $\mathbf{W}^\perp$ be the perpendicular subspace. It is only necessary to show that, again, $\mathbf{W}^\perp$ is invariant under f. But this is easy. Let $\mathbf{u} \in \mathbf{W}^\perp$ be given. Then we have

$$\langle f(\mathbf{u}), \mathbf{v}_n \rangle = \langle \mathbf{u}, f(\mathbf{v}_n) \rangle = \langle \mathbf{u}, \lambda \mathbf{v}_n \rangle = \lambda \langle \mathbf{u}, \mathbf{v}_n \rangle = \lambda \cdot 0 = 0.$$

The rest of the proof follows as before. □

In the particular case where we have only real numbers (which of course are a subset of the complex numbers), then we have a symmetric matrix.

Corollary. *All real symmetric matrices can be diagonalized.*

Note furthermore, that even in the case of a unitary matrix, the symmetry condition, namely $a_{jk} = \overline{a}_{kj}$, implies that on the diagonal, we have $a_{jj} = \overline{a}_{jj}$ for all j. That is, the diagonal elements are all real numbers. But these are the eigenvalues. Therefore we have:

Corollary. *The eigenvalues of a self-adjoint matrix — that is, a symmetric or a Hermitian matrix — are all real numbers.*

Orthogonal matrices revisited

Let A be an $n \times n$ orthogonal matrix. That is, it consists of real numbers, and we have $A^t = A^{-1}$. In general, it cannot be diagonalized. But on the other hand, it can be brought into the following form by means of similarity transformations.

$$A'' = \begin{pmatrix} \pm 1 & & & & & \\ & \ddots & & & 0 & \\ & & \pm 1 & & & \\ & & & R_1 & & \\ & 0 & & & \ddots & \\ & & & & & R_p \end{pmatrix},$$

where each R_j is a 2×2 block of the form

$$\begin{pmatrix} \cos\theta & \pm \sin\theta \\ \sin\theta & \mp \cos\theta \end{pmatrix}.$$

To see this, start by imagining that A represents the orthogonal mapping $f : \mathbb{R}^n \to \mathbb{R}^n$ with respect to the canonical basis of $\mathbb{R}^n$. Now consider the *symmetric* matrix

$$B = A + A^t = A + A^{-1}.$$

This matrix represents another linear mapping, call it $g : \mathbb{R}^n \to \mathbb{R}^n$, again with respect to the canonical basis of $\mathbb{R}^n$.

But, as we have just seen, B *can* be diagonalized. In particular, there exists some vector $\mathbf{v} \in \mathbb{R}^n$ with $g(\mathbf{v}) = \lambda g(\mathbf{v})$, for some $\lambda \in \mathbb{R}$. We now proceed by induction on the number n. There are two cases to consider:

- $\mathbf{v}$ is also an eigenvector for f, or
- it isn't.

The first case is easy. Let $\mathbf{W} \subset \mathbf{V}$ be simply $\mathbf{W} = [\mathbf{v}]$. i.e. this is just the set of all scalar multiples of $\mathbf{v}$. Let $\mathbf{W}^\perp$ be the perpendicular space to $\mathbf{W}$. (That is, $\mathbf{w} \in \mathbf{W}^\perp$ means that $\langle \mathbf{w}, \mathbf{v} \rangle = 0$.) But it is easy to see that $\mathbf{W}^\perp$ is also invariant under f. This follows by observing first of all that $f(\mathbf{v}) = \alpha \mathbf{v}$, with $\alpha = \pm 1$. (Remember that the eigenvalues of orthogonal mappings have absolute value 1.) Now take $\mathbf{w} \in \mathbf{W}^\perp$. Then $\langle f(\mathbf{w}), \mathbf{v} \rangle = \alpha^{-1} \langle f(\mathbf{w}), \alpha \mathbf{v} \rangle = \alpha^{-1} \langle f(\mathbf{w}), f(\mathbf{v}) \rangle = \alpha^{-1} \langle \mathbf{w}, \mathbf{v} \rangle = \alpha^{-1} \cdot 0 = 0$. Thus, by changing the basis of $\mathbb{R}^n$ to being an orthonormal basis, starting with $\mathbf{v}$ (which we can assume has been normalized), we obtain that the original matrix is similar to the matrix

$$\begin{pmatrix} \alpha & 0 \\ 0 & A^* \end{pmatrix},$$

where A^* is an $(n-1) \times (n-1)$ orthogonal matrix, which, according to the inductive hypothesis, can be transformed into the required form.

If $\mathbf{v}$ is *not* an eigenvector of f, then, still, we know it is an eigenvector of g, and furthermore $g = f + f^{-1}$. In particular, $g(\mathbf{v}) = \lambda \mathbf{v} = f(\mathbf{v}) + f^{-1}(\mathbf{v})$. That is,

$$f(f(\mathbf{v})) = \lambda f(\mathbf{v}) - \mathbf{v}.$$

So this time, let $\mathbf{W} = [\mathbf{v}, f(\mathbf{v})]$. This is a 2-dimensional subspace of $\mathbf{V}$. Again, consider $\mathbf{W}^\perp$. We have $\mathbf{V} = \mathbf{W} \oplus \mathbf{W}^\perp$. So we must show that $\mathbf{W}^\perp$ is invariant under f. Now we have another two cases to consider:

- $\lambda = 0$, and
- $\lambda \neq 0$.

So if $\lambda = 0$ then we have $f(f(\mathbf{v})) = -\mathbf{v}$. Therefore, again taking $\mathbf{w} \in \mathbf{W}^\perp$, we have $\langle f(\mathbf{w}), \mathbf{v} \rangle = \langle f(\mathbf{w}), -f(f(\mathbf{v})) \rangle = -\langle \mathbf{w}, f(\mathbf{v}) \rangle = 0$. (Remember that $\mathbf{w} \in \mathbf{W}^\perp$, so that $\langle \mathbf{w}, f(\mathbf{v}) \rangle = 0$.) Of course we also have $\langle f(\mathbf{w}), f(\mathbf{v}) \rangle = \langle \mathbf{w}, \mathbf{v} \rangle = 0$.

On the other hand, if $\lambda \neq 0$ then we have $\mathbf{v} = \lambda f(\mathbf{v}) - f(f(\mathbf{v}))$ so that $\langle f(\mathbf{w}), \mathbf{v} \rangle = \langle f(\mathbf{w}), \lambda f(\mathbf{v}) - f(f(\mathbf{v})) \rangle = \lambda \langle f(\mathbf{w}), f(\mathbf{v}) \rangle - \langle f(\mathbf{w}), f(f(\mathbf{v})) \rangle$, and we have seen that both of these scalar products are zero. Finally, we again have $\langle f(\mathbf{w}), f(\mathbf{v}) \rangle = \langle \mathbf{w}, \mathbf{v} \rangle = 0$.

Therefore we have shown that $\mathbf{V} = \mathbf{W} \oplus \mathbf{W}^\perp$, where both of these subspaces are invariant under the orthogonal mapping f. By our inductive hypothesis, there is an orthonormal basis for f restricted to the $n-2$ dimensional subspace $\mathbf{W}^\perp$ such that the matrix has the required form. As far as $\mathbf{W}$ is concerned, we are back in the simple situation of an orthogonal mapping $\mathbb{R}^2 \to \mathbb{R}^2$, and the matrix for this has the form of one of our 2×2 blocks.