Geometry for Architects

Mario Mainardis

May 3, 2023

Contents

Chapter 1

Introduction

This book is partly based on my lectures on geometry at the course of Architecture of the University of Udine. Teaching mathematics to students that are not directly interested in that subject is usually a very hard task for they, justly, do not see the point of loosing time on a hard and apparently sterile (on their opinion) theory. This has become expecially true in the last years, when the need of mathematical knowledge for practical uses seems to have drastically diminished, or at least changed: once, talking with me about mathematics and architecture, Giancarlo Carnevale, at that time dean of the Faculty of Architecture at the IUAV of Venice, clearly stated that students of architecture should be given course in mathematics more focused at the "cultural" aspect of this subject more than the practical one, since the latter could be easily accomplished by the growing facilities given by computer aided design, or left to civil engeneers.

There are however many ways to interpret what "cultural" aspects of mathematics are.

The more trivial ones tend to privilege those concepts of mathematics that sound impressive to the non mathematician: e.g. non Euclidean geometries, fractals, exotic topologies etc.. A great mathematician I was fortunate to have as a professor in Padua, Jacopo Barsotti, never disguised his perplexity for the use of the hyperbolic paraboloid (a surface rather typical for covers of small stadiums). These topics are typical ingredients of the meetings on Mathematics and Art, which are often amongst the most sterile and boring experiences for both true artists and true mathematicians and, in general that kind of "cultural" aspects of mathematics are of no use in the formation of a young architect. Maybe they might work for a purely academical carreer, provided the people who judge you are naive enough to be impressed by your words. On the practical side they usually have atrocious results (but remember, there are exceptions, as to every non mathematical assertion you'll read in the sequel). One of the purposes of this book is to prepare you so that you'll be able to detect and defend yourself from such a use of mathematics. I have to say that also mathematicians are often quite naive when they approach art: e.g. it is remarkable how a minor graphic designer such as Maurits Cornelis Escher has great success among

certain mathematicians.

At a deeper level we have science of perception. This will be discussed in the introduction of the chapter about symmetry. This is something an architect has to be aware of, though it is seldomly sufficent to produce good architecture.

There is an even deeper level, in which there is no difference between mathematics and architecture. Let me first give you an example: at the beginning of his carreer, my father worked for a engeneering company that designed highways, overpasses, junctions, bridges etc. There he had the idea of swapping the two directions of travel of the higways: There are many advantages in doing so:

1. The faster, and more dangerous, lanes are at the sides of the higways, while the slower lanes are closer to the center and they are separate by a double emergency lane, which makes, in case of accidents, more difficult for a vehicle to jump into a lane in the other direction of travel.

2. The emergency lane at the centre has double width, making it easier for emergency vehicles to reach the place of a possible accident.

3. There's no need to double resting stations (one, in the middle of the two lanes can work for both of them).

4. Junctions and overpasses are effectively simplified (I let you figure out how these will change).

Well, this is precisely the kind of intuition a mathematician needs for his job: on one side a perfect knowledge of your tools and, on the other, the ability of thinking differently. Indeed I am quite convinced that the mental processes used for designing are no way different from those used for mathematical thinking. In both cases you are given a problem with some initial conditions and you must solve it with the correct use of the tools you have at hand. The advantage of mathematics is that it is simpler (which does not necessarily mean easier) than architecture, in the sense that the initial conditions and the tools are more evident, time invariant, and precisely defined. Thus a good course in mathematics can be a good training for the future architect: I was grown up with an anecdote about Mies van der Rohe not accepting at his course those students who had not passed the mathematics exam, and the best compliment about my lectures I ever had was from Elena Z., a student of architecture in Udine, who, commenting my course (after she had passed my exam!), said that she felt as if having attended it made her become more intelligent.

Finally, architects work with spaces and shapes. But, although everyone believes to have clear in mind what space is, or what a shape is, any attempt to give a definition of space or shape will be vague and confusing (try this as an entertainement with your friends). In this book I also want to show how mathematicians tried to formalize these intuitive concepts, and how close to (or far from) they are to our intuition.

Chapter 2

The Fundaments

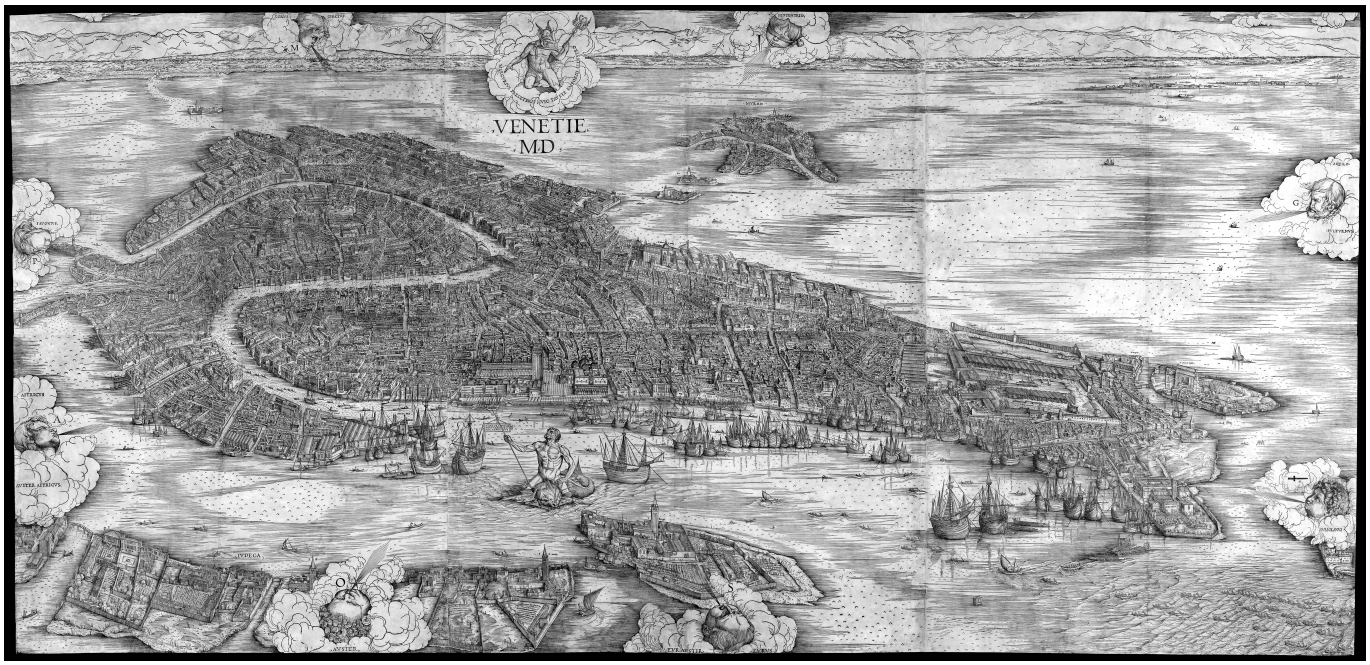


Figure 2.1: Jacopo de' Barbari ¨ *Veduta di Venezia a volo d'uccello*¨ (1500).

2.1 Introduction

Why do we start this section with a picture of Venice? Because similarly to contemporary mathematics Venice is a wonderful architecture whose fundaments

lie on mud. Similarly the fundaments of mathematics are based on certain *postulates* (the mud), that is assertions we accept as true as an act of faith

An important date for the fundaments of mathematics was the 8th August 1900, when, speaking at the International Congress of Mathematicians held in Paris, the German mathematician David Hilbert, proposed a list of problems that were open at that time (and some are still open now). The second of these problems concerned precisely the fundaments of mathematics. Taking as a model Euclidean Geometry, which is based on a formal system consisting on a finite set of primitive notions (such as point, line, circle etc.) and a finite set of axioms, i.e. rules that these primitive notions had to satisfy, (the five Postulates of Euclides), Hilbert was seeking for an analogous finite formal system as a fundament for mathematics in its whole. He required that this system should be as much intuitive and simple as possible and should satisfy the following two conditions:

> (Completeness:) On the basis of this formal system, one could decide the truth or the falseness of any statement within this formal system: for example the Goldbach Conjecture (which is still an open question) states that
>
> *every natural number greater than* 2 *is the sum of at most two prime numbers.*
>
> Is it true or not? Well, to fulfill completeness, on the basis of our axiomatic system we should be able, maybe after eons of efforts, either to prove the Goldbach Conjecture or prove that it is false.
>
> (Consistency:) The axioms in this formal system should not lead to a contradiction: for example one should not derive from that system statements like:
>
> *there exists objects such that* a *and* b a *is equal to* b *and* a *is different from* b
>
> which is manifestly contradictory.

Axiomatic set theory, developed by Abraham Fraenkel (1891-1965) and Ernst Zermelo (1871-1953) around 1908, seemed at a first glance to fulfill Hilbert's requirements, but, shortly after, in 1931, Kurt Gödel, showed that any axiomatic system satisfying those requirements would be too poor to be of any use. Namely, he showed that, given that any finite formal system, within which natural numbers could be defined, leads to the following consequences:

1. We cannot be prove that this system is consistent: there is no way to be sure that this system would not eventually lead to a contradiction[1]

[1]Here's an example of what could happen: in 1903 Gottlieb Frege (1848-1925), in an early attempt to provide an axiomatic system for set theory, was about to send to press the second

2. The system cannot be complete, i.e., there are undecidable statements (for example, the Goldbach Conjecture, which is still an open question, could be one of these statements[2])

Now, being not complete does not seem to me a major problem: if there are undecidable question, just add an answer (or its negation) as an axiom to that system: you will get two different theories (one accepting that answewr and one accepting its negation) and you can go on. On the other hand, the absence of the consistency condition is frightening: one works on the basis of an axiomatic system without being sure that that system is not contradictory. But that is precisely the act of faith of a mathematician[3]. Nevertheless, starting from it mathematics has developed into a wonderful (and useful) theory. That is what we shall do so in the sequel. Actually I shall require faith not just for the Zermelo-Fraenkel axiomatic system, but also on great part of the mathematics you have have been taught at school, without being given convincing proofs. Maybe that was one reason for many who have difficulties in learning mathematics at school. In the last section, I'll try to give a hint how natural, integer, and rational numbers can be constructed out of sets and how some of their basic properties, that we usually regard as obvious, can be derived from the Zermelo-Fraenkel axiomatic system.

2.2 Basic Logic

What do we mean by a *proof*? When we prove a theorem we perform a sequence of logical steps that start from the axioms, or from already proven results, and eventually lead to the assertion of the theorem. In this section I'll describe the basic tools that are used in the logical steps.

Take this example from arithmetics: We know from school that:

Fact 1 [Fundamental Theorem of Arithmetics] If p is a natural number greater than 1 then there exist a positive integer h and primes $p_1, p_2, \ldots, p_h$ such that

$$p = p_1 \leq p_2 \leq \cdots \leq p_h$$

volume of his *Grundgesetze der Arithmetik*, when Bertrand Russel (1872-1970) showed that a paradox could be derived from Frege's system: namely consider the set X whose elements are the sets that do not contain themselves as an element. There are obviously only two possibilities: either X contains itself as an element or X does not contain itself as an element. Now the first case leads to a contradiction, because if X contained itself as an element, then X would not contain itself as an element, for this is the property required for a set to be an element of X. But also the second case leads to a contradiction, for if X did not contain itself as an element, then X would satisfy the condition to be an element of X so it would contain itself as an element. Russel's paradox was eventually neutralized in the Zermelo-Fraenkel system, a description of which will be given in the sequel. Still we cannot be sure that in the future a new paradox will not pop out of that system.

[2](*Uncle Petros and Goldbach's Conjecture* written in 1992 by Greek author Apostolos Doxiadis is a worth reading novel about this conjecture.

[3]After all, this is also what we do in our every day's life, accepting what we are used to call the reality.

and

$$p = p_1 \cdot p_2 \cdot \cdots \cdot p_h$$

and if k is a positive integer and $q_1, q_2, \ldots, q_k$ are primes such that

$$p = q_1 \leq q_2 \leq \cdots \leq q_k$$

and

$$p = q_1 \cdot q_2 \cdot \cdots \cdot q_h,$$

then $h = k$ and $q_i = p_i$ for every $i \in \{1, \ldots, h\}$.

E.g.

$$120 = 2 \cdot 2 \cdot 2 \cdot 3 \cdot 5.$$

Fact 2 Every positive rational number z can be expressed as a fraction

$$z = \frac{p}{q}$$

where p and q are natural numbers greater than 0 and coprime to each other (i.e. no prime number divides both p and q).

Fact 3 Every rational number z different from 0 is either positive or negative and, if z is negative then $-z$ is positive.

Fact 4 If x and y are two rational numbers which are either both positive or negative then the product $x \cdot y$ is always positive and coincides with the product $(-x) \cdot (-y)$.

From these four facts (and elementary properties of the multiplication of rational numbers) we can derive that

{sqrt21}

Theorem 2.2.1 *There is no rational number whose square is* 2.

PROOF. Assume, by means of contradiction that there were a rational number, say z, whose square is 2. Since $0 \cdot 0 = 0$ and $0 \neq 2$, we have

$$z \neq 0.$$

So z is either positive or negative, therefore, by Fact 3, also $-z$ is either positive or negative. Whence, by Fact 4,

$$-z \cdot -z = z \cdot z = 2.$$

It follows that, possibly interchanging z with $-z$, we may assume that

z is a positive rational number.

By Fact 2 there exist two positive natural numbers p and q such that

$$z = \frac{p}{q}$$

By Fact 1 there exist prime numbers

$$p_1, p_2, \ldots, p_h \text{ and } q_1, q_2, \ldots, q_k,$$

such that

$$p = p_1 \cdot p_2 \cdot \cdots \cdot p_h \text{ and } q = q_1 \cdot q_2 \cdot \cdots \cdot q_k$$

with

$$p_1 \leq p_2 \leq \cdots \leq p_h \text{ and } q_1 \leq q_2 \leq \cdots \leq q_k.$$

It follows that

$$\begin{aligned} 2 &= \left(\frac{p}{q}\right)^2 \\ &= \frac{p^2}{q^2} \\ &= \frac{(p_1 \cdot p_2 \cdot \cdots \cdot p_h)^2}{(q_1 \cdot q_2 \cdot \cdots \cdot q_k)^2} \\ &= \frac{p_1^2 \cdot p_2^2 \cdot \cdots \cdot p_h^2}{q_1^2 \cdot q_2^2 \cdot \cdots \cdot q_k^2}. \\ &= \frac{p_1 \cdot p_1 \cdot p_2 \cdot p_2 \cdot \cdots \cdot p_h \cdot p_h}{q_1 \cdot q_1 \cdot q_2 \cdot q_2 \cdot \cdots \cdot q_k \cdot q_k}. \end{aligned}$$

Now, multiplying the first and the last members of the above equation by

$$q_1 \cdot q_1 \cdot q_2 \cdot q_2 \cdot \cdots \cdot q_k \cdot q_k$$

we get

$$2 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_2 \cdot \cdots \cdot q_k \cdot q_k = p_1 \cdot p_1 \cdot p_2 \cdot p_2 \cdot \cdots \cdot p_h \cdot p_h.$$

By the Fundamental Theorem of Arithmetics, $k + 1 = h$ and $2 = p_1$. Thus the last equation becomes

$$2 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_2 \cdot \cdots \cdot q_k \cdot q_k = 2 \cdot 2 \cdot p_2 \cdot p_2 \cdot \cdots \cdot p_h \cdot p_h$$

and, dividing by 2 both members, we get

$$q_1 \cdot q_1 \cdot q_2 \cdot q_2 \cdot \cdots \cdot q_k \cdot q_k = 2 \cdot p_2 \cdot p_2 \cdot \cdots \cdot p_h \cdot p_h.$$

Again, the Fundamental Theorem of Arithmetics implies that $q_1 = 2$, but we have seen that also $p_1 = 2$, so $q_1 = p_1$, which is a contradiction, because we assumed that p and q were coprime. ■ This was a proof *by contradiction* meaning

by that, that *i*f an assumption leads to a contradiction, *t*hen that assumption must be false and its negation true.

2.2.1 Implication

As in the above proof, in many occasions we will use in these lectures assertions of the kind

$$A \textbf{ implies } B,$$

or, equivalently,

$$\textbf{if } A \textbf{ then } B,$$

where A and B are two other assertions. In mathematical symbols, this will be denoted by

$$A \Rightarrow B.$$

For example assume I am standing in an open place without an umbrella and consider the sentence

If it rains (assertion A), **then** I will get wet (assertion B).

or, in mathematical symbols,

$$\text{It rains} \Rightarrow \text{I will get wet.}$$

This is an example of a logical implication, it means that whenever A happens (or is true), then also B should happen (should be true). Note that this does not mean that if A is not true then B is also not true: it might well happen that I get wet taking a bath on a sunny day. We say in this case that rain is a *sufficient* condition for me to get wet, but not a *necessary* condition (I can get wet in many other ways). Conversely, if A implies B, then B is a necessary condition for A: clearly it cannot be true that don't get wet when it rains (remember, I am standing in an open place without an umbrella). More generally let

$$C \text{ be the assertion } A \Rightarrow B$$

and

$$D \text{ be the assertion } \text{not}B \Rightarrow \text{ not}A$$

then C and D are *equivalent* in the sense that the truth of the assertion C is a necessary and sufficient condition for the truth of the assertion D, in other words C is true *if and only if* D is true. In symbols $C \Leftrightarrow D$. This is often useful, because it might happen that proving

$$\text{not}B \Rightarrow \text{ not}A$$

can be easier, or more intuitive, than proving

$$A \Rightarrow B.$$

Another example of necessary and sufficient conditions apt for architects comes from Ignazio Gardella (a famous Italian architect who was professor at the Istituto Universitario di Architettura of Venice during its golden years), showing the equivalence of beauty and function:

a terrace that is not beautiful is not functional,
a terrace that is not functional is not beautiful.

Note that this could be equivalently stated as

a terrace that is beautiful is also functional,
a terrace that is functional is also beautiful.

Or,

a necessary and sufficient condition for a terrace to be beautiful is to be functional.

Common mistakes by students trying to prove a sentence like

$$A \Rightarrow B$$

are proving

$$B \Rightarrow A$$

or

$$\text{not}A \Rightarrow \text{ not}B.$$

Similarly when proving the double implication

$$A \Leftrightarrow B$$

one has to prove first that

$$A \Rightarrow B$$

and then

$$B \Rightarrow A.$$

Often many students forget to prove one of the two above implications. The possible combinations of trueness of A, B and $A \Rightarrow$, can be summarized in the following truth table:

A	B	$A \Rightarrow B$
$True$	$True$	$True$
$True$	$False$	$False$
$False$	$True$	$True$
$False$	$False$	$True$

We have seen that the implication $A \Rightarrow B$ is true also when A is false and B is true (*if it rains I'll get wet* does not conflict with the fact that I can get wet on a sunny day). Similarly the implication $A \Rightarrow B$ is true also when both A and B are false: e.g.

If *I were a king (with absolute power)* (assertion A),
then *I could have you thrown in the dungeon* (assertion B),

is true for everyone, though for most people both assertions A (being a king) and B (having the power to throw someone in the dungeon) are false.

As an exercise, complete the truth table of $A \Leftrightarrow B$.

Note that there are certain assertions, called *paradoxes* for which it is not possible to tell if they are true or false. The most simple example is the following versione of the (selfdeclaring) *liar's paradox*:

This assertion is false

For if that assertion were true, it would be false and, conversely, if it were false, it would be true.

Finally notation is important: the symbols $\Rightarrow$ and $\Leftrightarrow$ have to be written with a double horizontal line, not a single one, since the symbol $\rightarrow$ (for example) has a different meaning.

2.2.2 Quantifiers and negation

The first, and incorrect, formulation of the liar's paradox is Epimenides' paradox, after the Cretan philosopher Epimenides of Knossos. Epimenides stated

All Cretans are liars

[**?**] This is an apparent paradox, since Epimenides can be a liar, but there might exist some other Cretans that are not liars, so Epimenides would be a liar when he says *All Cretans are liars*, giving no paradox (had he simply said "*I am a liar*, it would have been a different matter, for then there would have been a paradox). Here we have introduced two important quantifiers *all* (or *for all*, *for every*) and *exists*. They are usually denoted by $\forall$ (for all) and $\exists$ (exists). Thus the sentence

All Cretans are liars

can be written in symbols

$\forall$ *Cretans are liars*

or, better,

$\forall$ *person* x*, such that* x *is a Cretan,* x *is a liar*,

that reads

for every person x*, such that* x *is a Cretan,* x *is a liar.*

As we mentioned above, this assertion is false when we find a Cretan that is not a liar. That is

there exists a person x *such that* x *is a Cretan and* x *is not a liar*,

in symbols

$\exists$ *a person* x *such that* x *is a Cretan and* x *is not a liar.*

This shows that whenever I need to prove that an assertion that *all* certain objects have a certain property (*all* Cretans are liars) is false, I just have to prove that there *exists* a *counterexample*, that is one of those objects that does not have that property (there *exists* a Cretan that is not a liar). Conversely, when I have to negate an assertion that says that there *exists* a certain object that has a certain property, I have to show that *all* objects do not have that property.

Usual mistakes by students is to confuse the quantifiers $\forall$ and $\exists$. There is a great difference between the sentences

Every Cretan is a liar

and

There is a Cretan who is a liar.

Also, when asked to negate a sentence of the kind

$\forall$ *fish* x, x *is blue*

(inaccurate) students often mistakenly write

$\forall$ *fish* x, x *is not blue*

instead of

$\exists$ a *fish* x, such that x *is not blue.*

Similarly when one has to negate a sentence like

A bird has feathers **and** *flies*

the correct negation is

A bird does not have feathers **or** *does not fly*

and not

A bird does not have feathers **and** *does not fly.*

In symbols

not(A **and** B) is equivalent to (not$A$ **or** not$B$)

and symmetrically

not(A **or** B) is equivalent to (not$A$ **and** not$B$)

2.3 Sets

The concepts of set and element of a set can be considered as the fundamental blocks of the whole mathematical buildingThey are (at least in naive set theory) *primitive notions* in the sense that they are not formally defined by means of other mathematical objects, but are considered as innate in our minds. We can only give descriptions or examples of these concepts: we can talk about the set of students in a class and its elements are the students belonging to that class, or the set of Lego blocks of a given shape, whose elements are the Lego blocks of that shape. Usually, after a certain number of examples, everyone seems to understand what these primitive notions are. In these notes, we presume that this has already been achieved at school and that the reader is familiar with the notions of set, element of a set, and equality between elements.

There is an *axiomatic* set theory which provides some rules *axioms* (or *postulates*) that sets have to satisfy. It would be too long to expose it here, but we wish to introduce in the sequel some of the less technical of these postulates as a taste.

2.3.1 Describing a set

If a is an element of a set A we write $a \in A$ and if two elements a and b of a set A are the equal we write $a = b$, if they are not we write $a \neq b$. Of course a set can be an elemet of another set (students are elements of a class and a class is an element of the set of all classes of a school), but, to avoid problems arising from the Russel Paradox, the following condition is required:

Axiom 2.1 (Regularity) *Given two sets A and B, if $A \in B$ then $B \notin A$.*

An easy way to define a set with few elements is to list the elements into curly braces: so, e.g.,

- $\{1, 3, 4, 5, 8\}$ is the set whose elements are precisely the numbers 1, 3, 4, 5, and 8.

When we use the curly braces notation, the order in which the elements are listed is not important: e.g.

$$\{1, 2, 3\}, \ \{3, 2, 1\}, \text{ and } \{2, 1, 3\}$$

all describe the same set: we do not bother about the order we list the elements in a set. Further no matter how many times the same element appears inside the braces, it just count for one: e.g.

$$\{1\} = \{1, 1\} = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$$

If we want to define larger, or possibly infinite, sets things can be more complicated: one way, often efficient, is to describe them giving only some elements of the set followed by three dots hoping that the reader's intuition will

make him understand what the set is: for example we can describe the set $\mathbb{N}$ of natural numbers as follows

$$\mathbb{N} := \{0, 1, 2, 3, ...\}$$

and most people will understand what we are meaning. Similarly the set $\mathbb{Z}$ of integer numbers can be described as

$$\mathbb{Z} := \{0, 1, -1, 2, -2, 3, -3, ...\}.$$

Note however that these are not s but just descriptions, since we are relying on the reader's intuition.

Another way to define a set is to give a property that only the elements of that set have. In this case we first declare after the opening brace which elements we consider and after the symbol | (which reads *such that*) we define the property these elements have to satisfy and close the braces. E.g. the set $\mathbb{Q}$ of rational numbers can be described as follows:

$$\mathbb{Q} := \{p/q | p \in \mathbb{Z}, q \in \mathbb{Z}, \text{ and } q \neq 0 \text{ and } p \text{ and } q \text{ e coprime}\} \qquad (2.1) \quad \texttt{\{Q\}}$$

which reads: $\mathbb{Q}$ *is the set of all fractions* p/q **such that** p *is an integer,* q *is an integer with* q *not equal to* 0 *and* p *and* q *are coprime.*

Other important sets of numbers, which we shall use in the sequel, are the set of real numbers (denoted by $\mathbb{R}$) and the set $\mathbb{C}$ of complex numbers[4].

2.3.2 Subsets and the empty set

Given two sets A and B we say that A *is contained* in B (or that A is a *subset* of B), if every element of A is also an element of B, that is if the implication

$$x \in A \Rightarrow x \in B$$

is true. In this case we write $A \subseteq B$. The two sets A and B are said to be *equal* if both assertions $A \subseteq B$ and $B \subseteq A$ are true.

Examples:

- $\{1, 2, 3\} \subseteq \{1, 2, 3, 4, 5\}$
- $\{1, 2, 3\}$ is not a subset of $\{1, 3, 4, 5\}$ for 2 is an element of $\{1, 2, 3\}$, but not an element of $\{1, 3, 4, 5\}$.

The *empty set* is the set that contains no elements and it denoted by $\emptyset$. Formally, the empty set id defined as

$$\emptyset = \{A | A \text{ is a set and } A \neq A\}.$$

Since no set can be different from itself, the emptyset contains no element.

[4]As for sets we also presume the reader is already acquainted with basic notions of arithmetics in $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$, such as the basic properties of addition, multiplication, and the usual ordering (associativity, symmetry, distributivity, and when, given two numbers, one number is larger than the other). In the last section we shall give a clue how natural, integer and rational numbers can be formally defined out of sets and how complex numbers can be defined out of real numbers.

Axiom 2.2 (EMPTY SET) *The empty set exists as a set and is a subset of every set.*

Note that

$$\{\emptyset\} \neq \emptyset,$$

for $\{\emptyset\}$ contains one element (the emptyset itself!). Iterating this process we have a chain

{natural}
$$\emptyset \in \{\emptyset\} \in \{\{\emptyset\}\} \in \{\{\{\emptyset\}\}\} \in \cdots . \tag{2.2}$$

of sets each two different form each other. The series (2.2) suggests a way we can construct the natural numbers out of set theory: just call

$$0 := \emptyset,\ \ 1 := \{\emptyset\},\ \ 2 := \{\{\emptyset\}\},\ \ 3 := \{\{\{\emptyset\}\}\}, \ldots \text{ and so on}$$

It is however more convenient to construct the natural numbers out of sets in a slightly different way, as we shall do in Section 2.6 of this chapter.

2.3.3 Union, intersection, and difference of sets

The *union* $A \cup B$ of A and B is the set of elements that are either elements of A or elements of B. E.g.,

- $\{1,2,3,6\} \cup \{1,2,4,5\} = \{1,2,3,4,5,6\}$,
- $\{1,2,3,6\} \cup \{1,2,3,6\} = \{1,2,3,6\}$,
- for every set A, $A \cup \emptyset = A$.

The *intersection* $A \cap B$ of A and B is the set of elements that are contained both in A and B. E.g.,

- $\{1,2,3,6\} \cap \{1,2,4,5\} = \{1,2\}$,
- $\{1,2,3,6\} \cap \{4,5\} = \emptyset$,
- $\{1,2,3,6\} \cap \{1,2,3,4,5,6\} = \{1,2,3,6\}$.

The *difference* $A \setminus B$ of A and B is the set of the elements of A that are not elements of B. E.g.

- $\{1,2,3,6\} \setminus \{1,2,4,5\} = \{3,6\}$,
- $\{1,2,3,6\} \setminus \{1,2,3,4,5,6,\} = \emptyset$,
- $\{1,2,3,6\} \setminus \emptyset = \{1,2,3,6\}$

{unint}
Lemma 2.3.1 *Let A and B be sets with $A \subseteq B$, then*

1. $A \cup B = B$
2. $A \cap B = A$

PROOF. In order to prove the first assertion, we have to show that every element of $B \subseteq A \cup B$ and, viceversa, every element of $A \cup B \subseteq B$. The first inclusion is obvious, viceversa, assume $x \in A \cup B$. Then either $x \in A$ or $x \in B$. If $x \in B$, we are done, if $x \in A$, then, since $A \subseteq B$, x is also an element of B. Since in both cases x is an element of B, we are done. Now consider the second assertion: since $A \subseteq A$ and $A \subseteq B$, we have $A \subseteq A \cap B$. , assume $x \in A \cap B$. Then $x \in A$, so $A \cap B \subseteq A$. ∎

{diff}

Lemma 2.3.2 $A \setminus B = A \setminus (A \cap B)$

PROOF. We have to prove that

$$A \setminus B \subseteq A \setminus (A \cap B) \text{ and } A \setminus (A \cap B) \subseteq A \setminus B. \tag{2.3}$$ {doubleinc}

Let $a \in A \setminus B$ then $a \in A$ and $a \notin B$. In particular, since $A \cap B \subseteq B$, $a \notin A \cap B$, so $a \in A \setminus (A \cap B)$, which shows the first inclusion of (2.3). Conversely, assume $a \in A \setminus (A \cap B)$. Then $a \in A$ and $a \notin A \cap B$. But this implies that $a \notin B$, so $a \in A \setminus B$. ∎

2.3.4 Pairs and n-tuples

Given an element a_1 of a set A_1 and an element a_2 of a set A_2, we can formally define the *ordered pair* (a_1, a_2) as the set $\{a_1, \{a_1, a_2\}\}$. Here we have an asymmetrical rôle of a_1 and a_2, for the fundamental property of ordered pairs is

{pairs}

Lemma 2.3.3 *If (a_1, a_2) and (b_1, b_2) are two ordered pairs such that $(a_1, a_2) = (b_1, b_2)$, then $a_1 = b_1$ and $a_2 = b_2$.*

PROOF. By the definition of ordered pair,

$$(a_1, a_2) = (b_1, b_2) \Leftrightarrow \{a_1, \{a_1, a_2\}\} = \{b_1, \{b_1, b_2\}\}.$$

By definition of equality between sets, a_1 has to be an element of the set $\{b_1, \{b_1, b_2\}\}$, so either

$$a_1 = b_1 \text{ and } \{a_1, a_2\} = \{b_1, b_2\},$$

or

$$a_1 = \{b_1, b_2\} \text{ and } \{a_1, a_2\} = b_1.$$

But this second possibility cannot occour, for otherwise

$$a_1 \in \{a_1, a_2\} = b_1 \in \{b_1, b_2\} = a_1,$$

that is $a_1 \in b_1 \in a_1$ contradicting the Regularity Axiom. We now distinguish two cases: $a_1 = a_2$ and $a_1 \neq a_2$. Assume first $a_1 = a_2$, then $\{a_1\} = \{a_1, a_2\} =$

$\{b_1, b_2\} = \{a_1, b_2\}$. By definition of equality betwen sets, we must have $b_2 \in \{a_1\}$, but, since a_1 is the unique element of $\{a_1\}$, we have also $b_2 = a_2$ and we are done. Assume finally $a_1 \neq a_2$. Since $\{a_1, a_2\} = \{b_1, b_2\}$ and $a_1 = b_1$, again by definition of equality between sets, we have $b_2 \in \{a_1, a_2\}$, so

$$\text{either } b_2 = a_1 \text{ or } b_2 = a_2.$$

But the first possibility cannot occour, since, if $b_2 = a_1$, then

$$a_2 \in \{a_1, a_2\} = \{b_1, b_2\} = \{a_1, a_1\} = \{a_1\}$$

whence $a_1 = a_2$, which cannot happen, since we assumed $a_1 \neq a_2$. So $a_1 = b_1$ and $a_2 = b_2$. ∎

Given an ordered pair (a_1, a_2), a_1 is called the first coordinate of the pair and a_2 its second coordinate.

The construction of an ordered pair can be easily generalised to construct define triples, quadruples, or, more generally, n-tuples for every natural number n: just define

- the ordered triple (a_1, a_2, a_3) as the pair $((a_1, a_2), a_3)$,
- the ordered quadruple (a_1, a_2, a_3, a_4) as the pair $((a_1, a_2, a_3), a_4)$

  or, more generally, for every natural number n different from 0,

- the ordered n-tuple $(a_1, a_2, \ldots, a_{n-1}, a_n)$ as the pair $((a_1, a_2, \ldots, a_{n-1}), a_n)$

The three dots ... in the last formula mean that we should imagine that their place is filled by all the a_i's missing from a_2 to a_n: since n is variable, we don't know how large is n, so we don't know how many a_i's are needed to fill the formula but, if, e.g., $n = 5$, when we read

$$a_1, a_2, \ldots, a_5,$$

we should understand

$$a_1, a_2, a_3, a_4, a_5.$$

Remember to use exactly three dots, not one more not one less. For completeness, we also define the 1-*tuple* (or *singleton*) (a) as the set $\{a\}$.

{ntuples}

Corollary 2.3.4 *If $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$ are two n-tuples such that $(a_1, a_2, \ldots, a_n) = (b_1, b_2, \ldots, b_n)$, then $a_i = b_i$ for every $i \in \{1, \ldots, n\}$*

PROOF. Assume the assertion were not true, and choose the minimal natural number n such that there are two n-tuples $(a_1, a_2, \ldots, a_{n-1}, a_n)$ and $(b_1, b_2, \ldots, b_{n-1}, b_n)$ are two n-tuples such that $(a_1, a_2, \ldots, a_n) = (b_1, b_2, \ldots, b_n)$ but $a_i \neq b_i$ for some $i \in \{1, \ldots, n\}$. By the definition of equality between sets,

n cannot be 1 and, by Lemma 2.3.3, n cannot be 2. So we can assume that n is an integer greater than 2. Since, by definition,

$$\begin{aligned}((a_1, a_2, \ldots, a_{n-1})a_n) &= (a_1, a_2, \ldots, a_{n-1}, a_n) \\ &= (b_1, b_2, \ldots, b_{n-1}, b_n) \\ &= ((b_1, b_2, \ldots, b_{n-1}), b_n)\end{aligned}$$

By Lemma 2.3.3, we have

$$(a_1, \ldots, a_{n-1}) = (b_1, \ldots, b_{n-1}) \text{ and } a_n = b_n.$$

But now $n-1$ is smaller than n, so by the minimal choice of n, $a_1 = b_1$, $a_2 = b_2$ etc. ■

The above proof is a proof by *induction*. It uses the fact that, by Lemma **??**, every nonempty subset of the natural numbers has a minimun element. Assume we have a set

$$\mathcal{A} := \{A_1, A_2, A_3 \ldots\}$$

of assertions A_i, indexed by the natural numbers, and we want to prove that all of them are true. We can proceed by means of contradiction, that is showing that if we assume that some of them are false we get a contradiction. In this case take the subset $\mathcal{F}$ of all assertions that are false, then there should be a minimum n such that A_n is in $\mathcal{F}$, that is A_n is false. This means that for every m smaller than n A_m has to be true. If this information leads us to a contradiction, we are done. For example, in the above proof,

A_1 was the assertion $[(a_1) = (b_1)] \Rightarrow [a_1 = b_1]$,

A_2 was the assertion $[(a_1, a_2) = (b_1, b_2)] \Rightarrow [a_1 = b_1 \text{ and } a_2 = b_2]$,

A_3 was the assertion $[(a_1, a_2, a_3) = (b_1, b_2, b_3)] \Rightarrow [a_1 = b_1, a_2 = b_2, a_3 = b_3]$,

and so on.

Another way to see this kind of argument is the following: Suppose A_1 is true and that, for every positive integer n, the implication

$$(A_n \text{ is true } \Rightarrow A_{n+1} \text{ is true})$$

is true, then all A_i's for every positive integer i are true.

2.3.5 Cartesian products of sets

Given two sets A and B the *Cartesian product* of A and B (in symbols $A \times B$) is the set of all pairs whose first coordinate is an element of A and whose second coordinate is an element of B:

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

{cartprod}

Example 2.3.1 *Suppose* $A = \{a_1, a_2, a_3, a_4\}$ *and* $B = \{b_1, b_2, b_3\}$*, then*

$$\begin{aligned} A \times B &= \{(a_1, b_1), (a_2, b_1), (a_3, b_1), (a_4, b_1), \\ &\quad (a_1, b_2), (a_2, b_2), (a_3, b_2), (a_4, b_2), \\ &\quad (a_1, b_3), (a_2, b_3), (a_3, b_3), (a_4, b_3)\}. \end{aligned}$$

The name "Cartesian" is not casual, since, the Cartesian plane is actually the direct product of two lines perpendicular to each other. Analogously, we can visualize the Cartesian product of two (finite) sets A *and* B *in the following way: we list the elements of the set* A *on a horizontal line and the elements of* B *is a vertical line, and each pair* (a_i, b_j) *of elements of* $A \times B$ *in the box of the rectangle corresponding to the same column of* a_i *and the same row of* b_j*, as in the picture below:*

b_m	(a_1, b_m)	(a_2, b_m)	$\dots$	(a_{n-1}, b_{m-1})	(a_n, b_{m-1})
b_{m-1}	(a_1, b_{m-1})	(a_2, b_{m-1})	$\dots$	(a_{n-1}, b_{m-1})	(a_n, b_{m-1})
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
b_2	(a_1, b_2)	$(a_2, b_2 s)$	$\dots$	(a_{n-1}, b_2)	(a_n, b_2)
b_1	(a_1, b_1)	(a_2, b_1)	$\dots$	(a_{n-1}, b_1)	(a_n, b_1)
	a_1	a_2	$\dots$	a_{n-1}	a_n

More generally, given a positive integer n and n sets

$$A_1, A_2, \dots, A_n$$

we define the *cartesian product*

$$A_1 \times A_2 \times \cdots \times A_n$$

of $A_1, A_2, \dots, A_n$, as the set (of n-tuples)

$$\{(a_1, a_2, \dots, a_n) | a_1 \in A_1, a_2 \in A_2, \dots, a_n \in A_n\}.$$

If all A_i's are equal to a set A, we'll also write A^n for $A_1 \times A_2 \times \cdots \times A_n$, as in the case where $n = 2$.

Cartesian products will play a fundamental rôle in the sequel.

2.3.6 The power set

Let X be a set, we define the *power set* 2^X of X as the set of all subsets of X:

$$2^X := \{Y | Y \subseteq X\}$$

For example, if

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

its subsets are

$$\emptyset,\ \{1\},\ \{2\},\ \{3\},\ \{2, 3\},\ \{1, 3\},\ \{1, 2\},\ \text{and } \{1, 2, 3\},$$

thus

$$2^X = \{\emptyset, \{1\}, \{2\}, \{3\}, \{2, 3\}, \{1, 3\}, \{1, 2\}, \{1, 2, 3\}\}.$$

2.4 Correspondences and maps

2.4.1 Correspondences

Intuitively a correspondence between two sets A and B is a "rule" ρ that "associates" some elements of A to some elements of B. As we will see in the sequel, this general concept includes many important other ones which are of fundamental importance for mathematics, namely, the concepts of function, relation, equivalence and order.

{elem}

Example 2.4.1 *Let $A = \{1, 2, 3, 4\}$ and $B = \{h, k, l, m, n\}$, we can consider the correspondence ϵ that associates 1 to h 2 to k, 3 to n and 4 to h, k, l, m, and n.*

Example 2.4.2 (Lego bricks and their colours) *If A is a set of Lego bricks and B is the set of their colours, we can consider the correspondence χ that associates to each Lego brick its colour.*

Example 2.4.3 (Mother and child) *If A is the set of women in a community and B the set of children, we can consider the correspondence μ that associates to each mother their children.*

Example 2.4.4 (Military hierarchy) *If A is the set of all personnel in an army and $B = A$, we can consider the correspondence κ that associates two members a and b if and only if a commands b.*

Example 2.4.5 (Train timetables) *A (simplified) train timetable is a correspondence τ associates to every train its the departure time so it is a correspondence between the set of trains and the set of their departure times.*

Example 2.4.6 (Cartesian coordinates in three dimensions) *This is a correspondence that associates to each point in the three dimensional space a triple of real numbers (the coordinates of that point).*

Example 2.4.7 (Charts) *A cartographic map of a territory can be regarded as a correspondence π between the set of points of that territory and the set of points of a piece of paper (the map itself).*

So, apparently, we have a clear idea of what a correspondence is. The problem is, as usual, that what we have given above is not a formal definition, we are just postponing the issue of defining what a "correspondence" is to the problems of defining what "rule" and "link" are. On the other hand, observe we have a complete control of a correspondence ρ betwen the sets A and B when we can assign to every element a of A the elements of B that correspond to a and to every element b of B the elements of a that correspond to B. That is if we know the subset Γ_ρ of all pairs (a, b) in the cartesian product $A \times B$ such that a corresponds to b. The subset Γ_ρ is usually called the *graphic* of the correspondence

ρ. So, e.g., the correspondence ϵ of the elementary example above is perfectly determined once we know the subset

$$\Gamma_\epsilon := \{(1,h),(1,k),(3,n),(4,h),(4,k),(4,l),(4,m),(4,n)\}$$

of the Cartesian product $\{1,2,3,4\} \times \{h,k,l,m,n\}$.

Similarly, the correspondence between Lego bricks and their colour is completely known when we know the set of all pairs (a,b) where a is a Lego brick and b is the colour of a.

But if the graphic Γ_ρ, which is something we can formally define as a subset of the Cartesian product $A \times B$, encodes all information about what we have tried describe as a correspondence, why don't forget about the latter and identify a correspondence with its graphic? That is precisely what we shall do: a *correspondence* between two sets A and B is a subset of the Cartesian product $A \times B$ and we say that two elements $a \in A$ and $b \in B$ are *associated via* ρ if and only if the pair (a,b) is an element of the set ρ (in Example 2.4.1 this means that we identify ϵ with the set Γ_ϵ). So the military hierarchy κ in an army is just the set set of all pairs (a,b) where a and b are members of that army and a commands b.

If ρ is a correspondence between two sets A and B we call A the *domain* of ρ and B the *codomain* of ρ. The correspondence

$$\rho^{-1} := \{(b,a) | (a,b) \in \rho\}$$

between B and A is called the *inverse correspondence* of ρ. A correspondence between a set A and itself is called a *relation*.

There are many classes of correspondences between sets. For the moment I'll focus on one which is probably the most important: the class of functions.

2.4.2 Functions

Intuitively a function from a set A to a set B is something that projects elements of A to elements of B. Formally, a a *function* (or a *map*) *from* a set A *to* a set B is a correspondence f between A and B such that

for **every** element a of its domain A there is a **unique** element $b \in B$ such that $(a,b) \in f$.

In this case this unique element b is called the *image* of a via f and it is denoted by $f(a)$. To denote a function f from a set A to a set B we'll write

$$f\colon A \longrightarrow B$$

or

$$A \xrightarrow{f} B$$

and to denote that the element $b \in B$ is the image of $a \in A$ we'll also write

$$a \mapsto b$$

and say that f *maps* a to b.

So, e.g., if f is the function that maps every real number to its square, we will denote it by

$$\begin{array}{rccc} f\colon & \mathbb{R} & \longrightarrow & \mathbb{R} \\ & x & \mapsto & x^2 \end{array}$$

We shall use the words "function" or "map" indifferently.

Example 2.4.8 *The correspondence ϵ in Example 2.4.1 is not a function for two reasons:*

1. *there is no pair having* 2 *as a first coordinate (so it is not true that* **every** *element of the domain has an image in the codomain, because* 2 *is an element of the domain that has no image);*
2. *there are two pairs having* 1 *as the first coordinate* $(1, h)$ *and* $(1, k)$ *(so the "image" is not* **unique***).*

Example 2.4.9 (Lego bricks and colours) *In this case χ is a function for* **every** *Lego brick has a* **unique** *colour (transparent Lego bricks are not considered).*

Example 2.4.10 (Mother and child) *μ is a function if and only if every woman in that community has a unique child. If there are women without childs or if there are women that have more than one child μ is not a function. On the other hand, if there are no orphans in that community, the inverse correspondence μ^{-1} is a function (there is only one mother!).*

Example 2.4.11 (Military hierarchy) *much as people in the army would disagree, κ is not a function: a general commands many soldiers (***uniqueness** *fails) and a simple soldier commands nobody (***existence** *fails). The inverse correspondence κ^{-1} is not a function either (a soldier is commanded by all his superiors, which can be many) and the chief of the army is commanded by no one else in the army.*

Example 2.4.12 (Train timetables) *τ is a function but τ^{-1} does not need to be a function (there might be two different trains leaving at the same time).*

Example 2.4.13 (Cartesian coordinates) *This is a function between the points in the space and the set $\mathbb{R}^3$ of triples of real numbers. Its inverse correspondence is also a function.*

Example 2.4.14 (Charts) *These are functions between the set of the points of the territory mapped and the set of the points in the chart. Incidentally the term "map" as a synonimous of "function" comes from this example.*

Example 2.4.15 (The identity map) *Given a set A the correspondence between A and itself defined by*

$$id_A := \{(a, a) | a \in A\}$$

is a function and it is called the identity function on A. Thus id_A is the function

$$\begin{array}{rccc} id_A\colon & A & \longrightarrow & A \\ & a & \mapsto & a \end{array}$$

that maps every element a of A to itself. Its inverse correspondence is also a map and coincides with the identity ma itself.

Example 2.4.16 (Operations) *Operations are* **very important examples of maps**. *Given a set A and operation on A is a map whose domain is the cartesian product $A \times A$ and the codomain is the set A. Usually operations are denoted with symbols like $+, -, \circ, \times, \cdot, *$, etc. and, given an operation*

$$*\colon A \times A \longrightarrow A,$$

for a pair (a, b) in $A \times A$ one denotes its image by $a * b$ instead of $*(a, b)$. So, for example, addition in $\mathbb{N}$ or $\mathbb{Z}$ is an operation.

We will often have to check if two maps f and g are equal. In practice we will have to check if they have the same domain and if, for every element a of their domain, $f(a) = g(a)$.

A map $f\colon A \to B$ is called *injective* if it maps distinct elements of A to distinct elements of B. In practice, to see if the function f is injective we have to prove that if a_1 and a_2 are two elements of A such that $f(a_1) = f(a_2)$, then $a_1 = a_2$.

Example 2.4.17 (Lego bricks and colours) *The map χ is not necessarily injective (different bricks can have the same colour).*

Example 2.4.18 (Child and Mother) *The map μ^{-1} is not injective for two different children have the same mother. On the other hand, if every mother has just one child μ^{-1} is injective.*

Example 2.4.19 (Train timetables) *Again the map τ is not injective, for different train can depart at the same time.*

Example 2.4.20 (Cartesian coordinates and maps) *These are injective.*

Example 2.4.21 (Operations) *In general operations are not injective: e.g. consider the addition in $\mathbb{Z}$: $1 + 3 = 4 = 2 + 2$, so the different pairs $(1, 3)$ and $(2, 2)$ have the same image.*

Dually a map $f\colon A \to B$ is called *surjective* if, for every element b of the codomain B, there is an element $a \in A$ such that $f(a) = b$.

Examples and non examples:

Example 2.4.22 (Lego bricks and colours) *The map χ is not surjective (there are colours which are not colours of any Lego brick).*

Example 2.4.23 (Child and Mother) *The map μ^{-1} is surjective only if every woman in that community has a child.*

Example 2.4.24 (Train timetables) *The map τ is surjective if the codomain is the set of the departure times of the trains depart at the same time.*

Example 2.4.25 (Cartesian coordinates and maps) *These are surjective (the paper map is assumed to have no border).*

Example 2.4.26 (Operations) *The addition in $\mathbb{Z}$ is surjective: for every $z \in \mathbb{Z}$, $z = z + 0$, so z is the image of the pair $(z, 0)$.*

A map $f\colon A \longrightarrow B$ is that is both injective and surjective is called *bijective* (or *one-to-one*). This means that for every element of a there is a unique element of b that corresponds to a and for every element of b there is a unique element of A that corresponds to b.

{inversemap}

Lemma 2.4.1 *A map $f\colon A \longrightarrow B$ is bijective if and only if the inverse correspondence f^{-1} is also a bijective map.*

PROOF. Assume first f is bijective. We have to prove that, for every $b \in B$, there is a unique $a \in A$ such that $(b, a) \in f^{-1}$. Since f is surjective, for every element b of B there is an element a of A such that $(a, b) \in f$, whence $(b, a) \in f^{-1}$. Since f is injective this element a is unique, so f^{-1} is a function. Note that, since f is a function, f^{-1} is bijective. Indeed, let b_1 and b_2 be elements of B. Set $a_1 := f^{-1}(b_1)$ and $a_2 := f^{-1}b_2$. Assume $f^{-1}(b_1) = f^{-1}(b_2)$, that is $a_1 = a_2$. Then

$$b_1 = f(f^{-1}(b_1))) = f(a_1) = f(a_2) = f^{-1}(f(b_2)) = b_2$$

so f^{-1} is injective. Finally if $a \in A$ then $a = f^{-1}(f(a))$ so f^{-1} is also surjective, hence bijective. ∎

Examples and non examples:

- **Lego bricks and colours** The map χ is not surjective (there are colours which are not colours of any Lego brick).
- **Child and Mother** The map μ^{-1} is surjective only if every woman in that community has a child.
- **Train timetables** The map τ is surjective if the codomain is the set of the departure times of the trains depart at the same time.
- **Cartesian coordinates and maps** These are surjective (the paper map is assumed to have no border).

Example 2.4.27 (Operations) *The addition in $\mathbb{Z}$is not bijective, since it is not injective.*

Restriction to a subset

Assume

$$f\colon A \to B$$

is a map, and U is a subset of A, Using the formal definition of a function as a correspondence as a subset of the cartesian product we have

$$f = \{(a, f(a)) | a \in A\}$$

Denote by

$$f_{|_U} := \{(u, f(u)) | u \in U\}$$

Clearly $f_{|_U}$ is a correspondence between U and B and it is immediate to see that $f_{|_U}$ is actually a map from U to B. The map $f_{|_U}$ is called the *restriction* of f to U. In the usual notation for maps, $f_{|_U}$ is the map

$$\begin{array}{rccc} f_{|_U}\colon & U & \to & B \\ & u & \mapsto & f(u) \end{array}$$

In other words, $f_{|_U}$ is the map having U as domain instead of A, B as codomain, and, for every element u in U the image $f_{|_U}(u)$ of u via $f_{|_U}$ is the same as the image $f(u)$ of u via f. So, for example, if f is the map

$$\begin{array}{rccc} f\colon & \{1,2,3\} & \to & \{a,b\} \\ & 1 & \mapsto & a \\ & 2 & \mapsto & b \\ & 3 & \mapsto & a \end{array}$$

then $f_{|_{\{1,2\}}}$ is the map

$$\begin{array}{rccc} f_{|_{\{1,2\}}}\colon & \{1,2\} & \to & \{a,b\} \\ & 1 & \mapsto & a \\ & 2 & \mapsto & b \end{array}$$

Composition of maps

Assume now we have three sets A, B and C and two maps

$$f\colon A \longrightarrow B \text{ and } g\colon B \longrightarrow C$$

Then we can define a map $g \circ f$ as follows:

$$\begin{array}{rccc} g \circ f\colon & A & \longrightarrow & B \\ & a & \mapsto & g(f(a)) \end{array}$$

So, to compute the image of an element a of A, we first take the image $f(a)$ of a via f. Then, since $f(a)$ is an element of the set B, we take the image $g(f(a))$ of $f(a)$ via g.

The map $g \circ f$ is called the *composition* of the maps g and f. Note that we can make the composition $g \circ f$ if and only if, for every element of the domain of f, $f(a)$ is an element of the domain of g.

Examples

- Let

$$f\colon \begin{array}{ccc} \mathbb{Z} & \longrightarrow & \mathbb{Z} \\ z & \mapsto & z+3 \end{array} \quad \text{and} \quad g\colon \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \mapsto & x^2 \end{array}$$

  then

$$g \circ f\colon \begin{array}{ccc} \mathbb{Z} & \longrightarrow & \mathbb{R} \\ z & \mapsto & (z+3)^2 \end{array}$$

- if $f\colon A \longrightarrow B$ is bijective, then $f \circ f^{-1} = id_B$ and $f^{-1} \circ f = id_A$.

{compinj}

Lemma 2.4.2 *Let A, B, and C be sets and let $f\colon A \to B$ and $g\colon B \to C$ be maps. Then*

1. *if f and g are injective, then so is $g \circ f$;*
2. *if f and g are surjective, then so is $g \circ f$;*
3. *if f and g are bijective, then so is $g \circ f$.*

{compass}

Lemma 2.4.3 *Let A, B, C, and D be sets and let $f\colon A \to B$, $g\colon B \to C$, and $h\colon C \to D$ be maps. Then*

$$h \circ (g \circ f) = (h \circ g) \circ f$$

{compidentity0}

Lemma 2.4.4 *Let A be a set and σ a function from A to A. Then*

$$\sigma \circ id_A = \sigma = id_A \circ \sigma.$$

PROOF. We prove the first equality, the proof of the second one is similar and is left as an exercise. Since $\sigma \circ id_A$ and σ are both functions from A to A, we just need to prove that, for every $a \in A$

$$\sigma \circ id_A(a) = \sigma(a).$$

But this follows immediately from the definition of composition of functions and the fact that $id_A(a) = a$ for every $a \in A$, indeed

$$\sigma \circ id_A(a) = \sigma(id_A(a)) = \sigma(a).$$

■

{compidentity1}

Lemma 2.4.5 *Let A and B be sets and $f\colon A \to B$ be a bijective map. Then*

1. *$f \circ f^{-1} = id_B$,*
2. *$f^{-1} \circ f = id_A$.*

PROOF. We prove the first assertion, the proof of the second one is similar and is left as an exercise. Since $f \circ f^{-1}$ and id_B have the same domain and the same codomain (namely the set B), we only need to prove that, for every $b \in B$,

$$f \circ f^{-1}(b) = id_B(b).$$

Indeed, let $b \in B$ and let $a := f^{-1}(b)$. Then, by the definition of the inverse map, $f(a) = b$, so

$$f \circ f^{-1}(b) = f(f^{-1}(b)) = f(a) = b = id_B(b).$$

■

{compidentity2}

Lemma 2.4.6 *Let A and B be sets and $f\colon A \to B$ a function. Then f is bijective if and only if there is a map $g\colon B \to A$ such that*

{compoo}
$$f \circ g = id_B \text{ and } g \circ f = id_A. \tag{2.4}$$

PROOF. If f is bijective, then Equation 2.4 is satisfied taking f^{-1} as g. Conversely, assume g is a map from B to A that satisfies Equation 2.4. We prove first that f is injective. Indeed let a and a' be elements of A such that $f(a) = f(a')$. Then

$$a = id_A(a) = g \circ f(a) = g(f(a)) = g(f(a')) = g \circ f(a') = id_A(a') = a',$$

which proves that f is injective. For the surjectivity, let $b \in B$ then

$$f(g(b)) = f \circ g(b) = id_B(b) = b,$$

so b is the image via f of $g(b)$ for every $b \in B$, thus f is also surjective, whence bijective. ■

If $f\colon A \longrightarrow B$ is a function and H is a subset of A, the *image* of H is the subset, denoted by $f(H)$, of B defined as follows:

$$f(H) := \{b \in B| \text{ there exists } a \in A \text{ such that } b = f(a)\}$$

Conversely, if K is a subset of G the *inverse image* of K is the subset $f^{-1}(K)$ of A defined as follows:

$$f^{-1}(K) := \{a \in A | f(a) \in K\}$$

Warning: by an unfortunate case due to tradition, the symbol f^{-1} is used in two different ways: one for defining the inverse correspondence of f, the other for defining the inverse image, which are two different things:

1. if $f\colon A \to B$ is a bijective map, for every element $b \in B$, the image $f^{-1}(b)$ of b via the inverse map f^{-1} is the *element* a of A such that $f(a) = b$,

   while

2. if $f\colon A \to B$ is any function, then, for every subset V of B, the preimage $f^{-1}(V)$ of V via f is the subset U of A such that $f(u) \in V$ for every $u \in U$.

2.4.3 Counting

{counting}

From a mathematical point of view, the only feature of a set which is important is its number of elements, in the sense that e.g. for a mathematician, as a difference to, say, a big game hunter, there's essentially no difference between a set of five bullets and a set of five lyons. The only property that counts is that both sets contain five elements. Now it should be clear that two sets have the same number of elements if and only if there's a bijection between them[5]: (the big game hunter can save himself if and only if he kills each of the five lyons with exactly one bullet). I believe that all big game hunters, and many other people with them, would agree that counting is useful. But, what does actually counting mean? And how do we make use of the knowledge of the number of elements in a set?

To answer the first question, just think what we do when we count the number of elements in a set: usually we point our index finger to the first object and say "one", then we shift our index finger to the next object and say "two" and so on until we have exhausted all the elements in that set. But this is just associating to each element of that set a natural number in the usual order, starting from 1. So we say that the set $\{a, b, c, d, e\}$ has five elements because there is bijective function from the set $\{a, b, c, d, e\}$ to the set $\{1, 2, 3, 4, 5\}$. We'll take this as a definition:

Let A be a non empty set and assume there is a natural number n and a bijection between A and the set $\{1, \ldots, n\}$ then we'll say that A is *finite* and its *cardinality* (its *order* or its *size*) is n. If there is no natural number n such that there is a bijection between A and $\{1, \ldots, n\}$ we say that A is an *infinite* set and denote its cardinality by ∞. Finally we say that the cardinality of the empty set is 0. For every set A we denote its cardinality by $|A|$. Thus, e.g.,

$$|\emptyset| = 0, |\{\emptyset\}| = 1, \{1, 2, 3, 5\} = 4, |\{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}| = 1$$

and they are all finite sets, while

$$\mathbb{N},\ \mathbb{Z},\ \mathbb{Q},\ \text{and } \mathbb{R}$$

are examples of infinite sets.

The meticulous reader would notice is however a problem: we haven't yet proved the apparently obvious fact that,

(*) *if A is finite, there is a unique $n \in \mathbb{N}$ such that there is a bijection between A and $\{1, \ldots, n\}$.*

If that weren't so, we could have different cardinalities for the same finite set. Fortunately (*) is true but a formal proof of this, and of other facts that we shall use in this subsection, is by no means trivial and requires a construction of the natural numbers out of set theory. In order to avoid burdening the readers with such details, we shall postpone them to the last section of this chapter, reserved to the daring readers, and ask the meticulous ones a further act of faith.

[5] Actually, this is the formal definition for two sets to have the same number of elements.

Let's now turn to the second question: how do we make use of the knowledge of the number of elements in a set?

Well, counting the elements of a set gives a measure of that set: if counting the number of elements of a set A we end up with a number that is not the one we expected, then we know that there is something wrong. For example, imagine a morning roll call in an army barrack. Let X be the set of soldiers in that barrack and Y the set of soldiers answering the call. If the number of elements of Y is the same as the number of elements in the set X, then $Y = X$ and everything is ok, otherwise there are soldiers missing. We have applied the following fact

{pig}

Lemma 2.4.7 *Let Y be a subset of a finite set X. Assume Y has the same number of elements as X, then $Y = X$.*

{pigoo}

Lemma 2.4.8 *Let X and Y be finite sets. Then there is a bijection $f\colon X \to Y$ if and only if $|X| = |Y|$.*

PROOF. Let $n = |X|$. Assume first that $|X| = |Y|$. Then there are bijections

$$f_X\colon X \to \{1, \ldots, n\} \text{ and } f_Y\colon Y \to \{1, \ldots, n\}.$$

By Lemma 2.4.1

$$f_Y^{-1}\colon \{1, \ldots, n\} \to Y$$

is also bijective, whence, by Lemma 2.4.2 the map

$$f_Y^{-1} \circ f_X\colon X \to Y$$

is bijective. Conversely, assume $g\colon X \to Y$ is a bijection. Again, by Lemma 2.4.1

$$g^{-1}\colon Y \to X$$

is also bijective. let f_X be as above, then, by Lemma 2.4.2, the map

$$f_X \circ g^{-1}\colon Y \to \{1, \ldots, n\}$$

is bijective, hence $|Y| = n = |X|$. ∎

As above, a formal proof of this result will be postponed in the last section of this chapter.

More generally, the following result holds:

{pigeon}

Lemma 2.4.9 (THE PIGEONHOLE PRINCIPLE FOR SETS) *Let X and Y be two finite sets with the same cardinality and $\phi\colon X \to Y$ a map. Then the following assertions are equivalent:*

1. *ϕ is injective;*
2. *ϕ is surjective;*

3. ϕ is bijective;

PROOF. Clearly 3 implies 1 and 2. Conversely, if we prove the equivalence of 1 and 2, then also 3 will follow. So assume ϕ is injective. Then no two distinct elements of X will have the same image. Therefore $|\phi(X)| = |X| = |Y|$, that is $\phi(X)$ is a subset of Y with the same cardinality as Y. By Corollary 2.6.5 this is possible if and only if $\phi(X) = Y$. Vice versa, for every $x \in X$, pick one and only one element y_x in $\{\phi^{-1}(\{x\})$ Then the map

$$\begin{array}{rccc} \psi\colon & X & \to & Y \\ & x & \mapsto & y_x \end{array}$$

is injective, so, by the first part of this proof, also bijective. Now observe that, for every $x \in X$, since $y_x \in \phi^{-1}(x)$,

$$\phi \circ \psi(x) = \phi(y_x) = x,$$

that is

$$\phi \circ \psi(x) = id_X,$$

which means that ϕ is precisely the inverse map ψ^{-1} of ψ, hence the result follows by Lemma 2.4.1. ∎

The name *Pigeonhole* is due to the following example, which also illustrates the proof: assume we have, say, 9 pigeons in 9 holes, and consider the map ϕ that associates to each pigeon the hole the pigeon is in. Of course we allow more pigeons to stay in the same hole, but, if we want to fill all the holes (surjectivity of ϕ) then each hole can contain only one pigeon (injectivity of ϕ). Vice versa, if we want each hole to contain only one pigeon (injectivity of ϕ), then we need to fill all the holes (surjectivity of ϕ).

We remark that **the Pigeonhole principle holds only for finite sets**[6]. Indeed an infinite set X admits maps from X to X that are injective, but not surjective and maps that are surjective but not injective. E.g. the map

$$\begin{array}{rccc} \mu\colon & \mathbb{N} & \to & \mathbb{N} \\ & x & \mapsto & x+1 \end{array}$$

is not surjective, for $0 \in \mathbb{N}$ but $\mu^{-1}(\{0\}) = \emptyset$, but it is injective: if $x, y \in \mathbb{N}$ and $\mu(x) = \mu(y)$, then

$$x = (x+1) - 1 = \mu(x) - 1 = \mu(y) - 1 = (y+1) - 1 = y.$$

Similarly the map

$$\nu\colon \quad \mathbb{N} \quad \to \quad \mathbb{N}$$

[6]Actually there is an equivalent definition of finite set that says: a set X is finite if and only if every injective map from X to X is also surjective (or equivalently, if and only if every surjective map from X to X is also injective).

defined by

$$\begin{cases} \nu(0) = 0 \\ \text{ and} \\ \nu(x) = x - 1, \quad \text{if } x > 0 \end{cases}$$

is surjective, but not injective since $\nu(0) = 0 = \nu(1)$.

We state now some obvious, though important, counting property of finite sets that we'll use and generalize in the sequel.

{diff0}

Lemma 2.4.10 *If A and B are finite sets with $A \cap B = \emptyset$, then $|A \cup B| = |A| + |B|$*

PROOF. Just count the elements of $A \cup B$ beginning first with the elements of A and then the elements of B. ∎

{diff1}

Lemma 2.4.11 *If A is finite set, $|A \setminus B| = |A| - |A \cap B|$*

PROOF. Clearly

$$A = (A \cap B) \cup (A \setminus (A \cap B)) \text{ and } (A \cap B) \cap (A \setminus (A \cap B)) = \emptyset,$$

so the result follows from Lemma 2.4.10. ∎

{Wittset}

Lemma 2.4.12 (THE WITT PROPERTY FOR FINITE SETS) *Let B and C be two subsets of a finite set A. Assume*

$$f\colon B \to C$$

is a bijection. Then there are bijections

$$\bar{f}\colon A \to A$$

that extend f, i.e. $\bar{f}(b) = f(b)$ for every $b \in B$

PROOF. Since f is a bijection, by Lemma 2.4.8, $|B| = |C|$, whence, by Lemma 2.4.11, also $|A \setminus B| = |A \setminus C|$. So, again by Lemma 2.4.8, there is a bijection

$$g\colon A \setminus B \to A \setminus C.$$

Now define

$$\bar{f}\colon A \to A$$

as follows:

$$\begin{cases} \bar{f}(a) = f(a) & \text{if } a \in B \\ \textit{and} \\ \bar{f}(a) = g(a) & \text{if } a \in A \setminus B \end{cases}$$

By definition $\bar{f}$ extends f. We prove that $\bar{f}$ is injective: let $x, y \in A$ with $x \neq y$. We distinguish three cases:

Case 1 If $x, y \in B$ then, since f is injective, $f(x) \neq f(y)$, whence $\bar{f}(x) = f(x) \neq f(y) = \bar{f}(y)$.

Case 2 If $x, y \in A \setminus B$ then, since g is injective, $g(x) \neq g(y)$, whence $\bar{f}(x) = g(x) \neq g(y) = \bar{f}(y)$.

Case 3 Assume finally that $x \in B$ and $y \in A \setminus B$ then $\bar{f}(x) = f(x) \in C$ and $\bar{f}(y) = g(y) \in A \setminus C$, whence, again $\bar{f}(x) \neq \bar{f}(y)$ since $A \cap (A \setminus C) = \emptyset$.

We finally prove that $\bar{f}$ is also surjective, hence bijective. Let $z \in A$, again we distinguish two cases:

Case 1 If $z \in C$ then, since f is surjective, there exists $x \in B$ such that $f(x) = z$, whence $\bar{f}(x) = f(x) = z$.

Case 2 If $z \in A \setminus C$ then, since g is surjective, there exists $y \in A \setminus B$ such that $g(y) = z$, whence $\bar{f}(y) = g(y) = z$.

■

Note that the map $\bar{f}$ extending f is general in not unique, since there may be more than one bijection g from $A \setminus B$ to $A \setminus C$.

{Grassets}

Theorem 2.4.13 (The Grassmann's Theorem for finite sets) *If A and B are finite sets, then $A \cup B$ and $A \cap B$ are finite sets and*

$$|A \cup B| = |A| + (|B| - |A \cap B|).$$

Proof. Count the elements of $A \cup B$ first count ingthe elements of B and then the elements of A which do not belong to B, that is the elements of $A \setminus B$. By Lemma 2.4.11 $|A \setminus B| = |A| - |(A \cap B)|$, So

$$|A \cup B| = |B| + (|A| - |A \cap B|) = |A| + (|B| - |A \cap B|).$$

■

Note also that, if A and B are finite sets, then $|A \times B|$ is precisely $|A| \times |B|$. Example 2.3.1, and the way the pairs are displaced, gives us a hint how to prove that: namely, for every element a of the set A, there are $|B|$ distinct pairs (one for each element of B) and any two of the $|A|$ distinct elements of A give two distinct pairs which makes in total $|A| \times |B|$ distinct pairs.

We finish this subsection summarizing some important facts.

1. The mathematical property of a finite set is its number of elements.
2. Two finite sets have the same number of elements if and only if there is a bijection between them.
3. A subset Y of a finite set X is equal to X if and only if Y and X have the same number of elements.

4. For a map between two finite sets that have the same number of elements injectivity is equivalent to surjectivity (hence to bijectivity).

5. the sets $\{1, \ldots, n\}$ for $n \in \mathbb{N} \setminus \{0\}$ constitute a set of representatives of the finite non empty sets, in the sense that given any finite non empty set X there is a unique positive integer n and a bijection ϕ between X and $\{1, \ldots, n\}$, and viceversa, for every positive integer n, there are sets of cardinality n (e.g. $\{1, \ldots, n\}$).

6. The Grassman Theorem for finite sets holds: if A and B are two finite sets, then $|A \cup B| = |A| + |B| - |A \cap B|$.

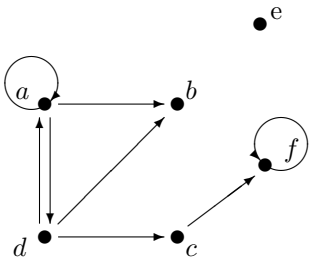
We shall see in Chapter 4 that analogous results hold for vector spaces over the real numbers. However these are infinite sets, so we cannot speak of cardinality. The corresponding concept will turn to be the *dimension*.

2.5 Relations and graphs

A correspondence between a set A and itself is called a *relation on* A. Given a relation ρ on a set A, there is a nice way to visualize it: draw the elements of A as points in a plane and, given two points a and b of A join the point a to the point b with an arrow from a to b. So, for example, if $A = \{a, b, c, d, e, f\}$ and

$$\rho_1 = \{(a, a), (a, b), (a, d), (d, a), (d, b), (d, c), (c, f), (f, f)\}$$

then we can represent ρ as follows:

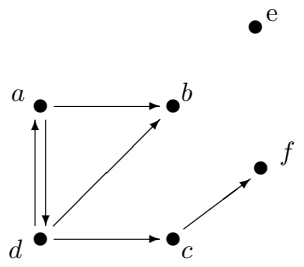


The pair (A, ρ) is called the *digraph*, short for *directed graph*, associated to the relation ρ, the elements of A (corresponding to the points) are called *vertices* of the digraph (A, ρ) and the elements of ρ (corresponding to the arrows) are called the *directed edges* of (A, ρ). The edges of the type (x, x) for $x \in A$ are called *loops*. For example (a, a) is a loop in ρ_1.

A relation ρ on a set A is called *reflexive* if for every $a \in A$ $a\rho a$. Again ρ_1 is not reflexive for, e.g., $(b, b) \notin \rho$. If we know that a relation is reflexive, it is customary to omit loops, so, for example, if

$$\rho_2 := \{(a, a), (b, b), (c, c), (d, d), (e, e), (f, f), (a, b), (a, d), (d, a), (d, b), (d, c), (c, f)\}$$

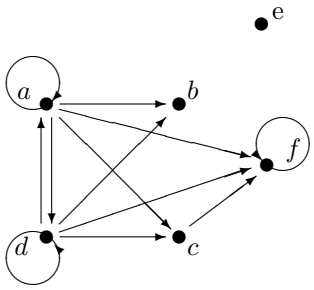
we shall represent it as follows:



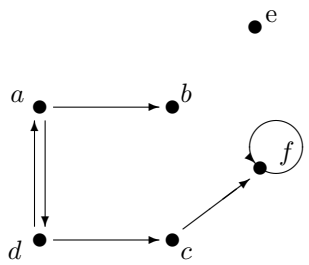
A relation ρ on a set A is called *transitive* if for every $a, b, c \in A$, whenever (a, b) and (b, c) are in ρ, then also $(a, c) \in \rho$. Note that neither ρ_1 nor ρ_2 are transitive, since, for example, $a\rho_1 d$ and $d\rho_1 c$, but $(a, c) \notin \rho_1$. We shall prove that, for every relation ρ there is a unique transitive relation $\bar{\rho}$ containing ρ and such that no relation σ containing ρ and properly contained in $\bar{\rho}$ is transitive. The relation $\bar{\rho}$ is called the *transitive closure* of ρ. Intuitively, $\bar{\rho}$ can be obtained adjoining to ρ all relations necessary to make it transitive. For example

$$\bar{\rho}_1 = \{(a, a), (a, b), (a, c), (a, d), (a, f), (d, a), (d, d), (d, b), (d, c), (d, f), (c, f)\}.$$

The relation $\bar{\rho}$, should be represented as follows:



As we can see, transitive relations tend to have many directed edges, most of which can be deduced from the existence of other edges and the transitivity. Thus if we know that the relation is transitive, often the directed edges whose existence can be deduced by other edges and transitivity can be omitted, so the a transitive relation $\bar{\rho}_1$ can also be represented as follows:

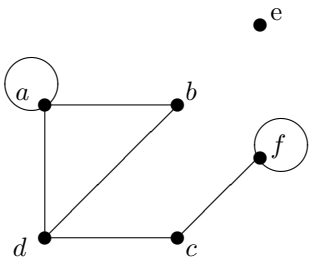


A transitive relation ρ on a set A is called *antisymmetric* if, for every $a, b \in A$, whenever $a\rho b$ and $b\rho a$, then $b = a$. Intuitively, this means that, in the associated digraph, once you go along an arrow (a directed edge), there is no way come back. The relation ρ_1 is not antisymmetric because $a\rho_1 d$ and $d\rho_1 a$, but $a \neq d$.

At the opposite side, a relation (not necessarily transitive) ρ on a set A is called *symmetric* if $\rho = \rho^{-1}$, or, equivalently, if, for every $a, b \in A$, $a\rho b$ implies $b\rho a$. In this case we shall represent the elements of ρ with lines instead of arrows. Note that ρ_1 is not symmetric, since, for example, $(c, f) \in \rho_1$, but $(f, c) \notin \rho_1$. On the other hand, the relation

$$\rho_4 := \{(a,a),(a,b),(b,a),(a,d),(d,a),(d,c),(c,d),(c,f),(f,c),(f,f)\}$$

on the set $\{a, b, c, d, e, f\}$ is symmetric and is represented as follows



2.5.1 Orderings

A *partial ordering* (or an *partial order relation*) on a set A is a relation on A that satisfies the following relations:

reflexivity: for every element a of A, $a\rho a$;

antisymmetry: for every a, b in A, $a\rho b$ and $b\rho a$ imply $a = b$;

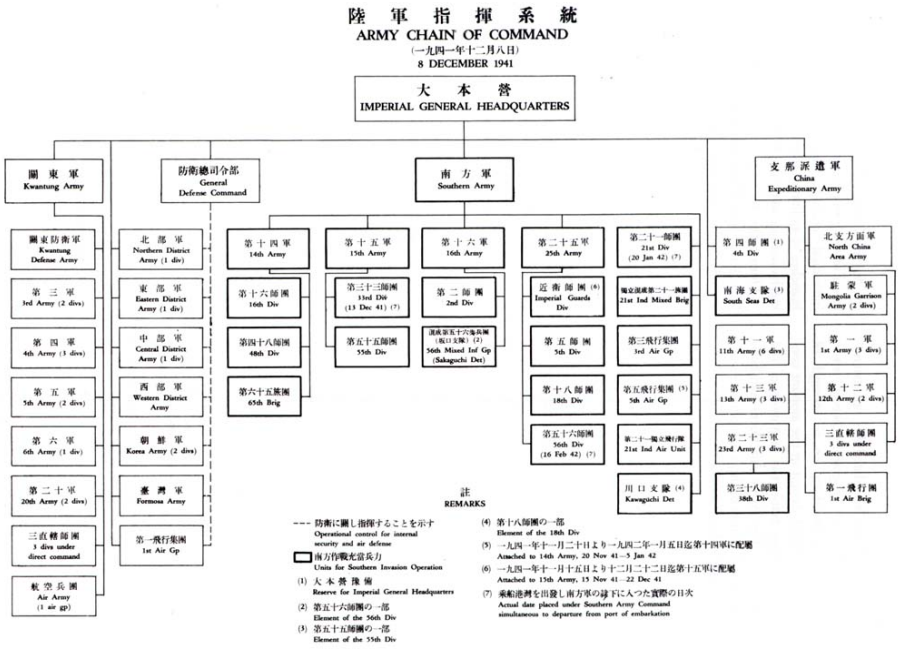
transitivity: for every a, b, c in A, if $a\rho b$ and $b\rho c$, then also $a\rho c$

The (visual representation of the) digraph associated to an ordering is called the *Hasse diagram*

Example 2.5.1 (Chain of commands) *In the set of people in an army A define a relation κ as follows: for any two members a and b of A*

$$a\kappa b \text{ if and only if } a \text{ commands } b$$

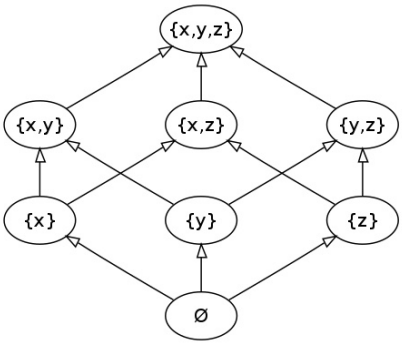
so if a is a general and b is a captain $a\kappa b$. Then κ is an ordering on the army A (we admit that anyone commands himself or herself, to guarantee symmetry, and that two different members of that army of the same grade cannot command each other, to guarantee antisymmetry). Figure 2.5.1 is the Hasse diagram of the relation κ in the Japanese Army in 1941 as a digraph. The arrows (which are missing) should be pointed downwards.



Example 2.5.2 (Inclusion) *Let A be a set and 2^A be the set of all subsets of A. Define a relation σ on 2^A as follows: for every X, Y subsets of A*

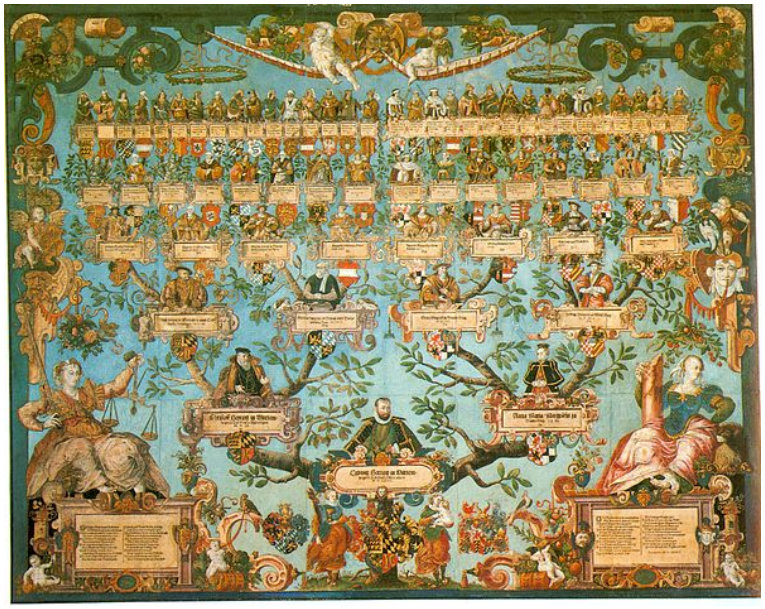
$$X\sigma Y \text{ if and only if } X \subseteq Y.$$

This relation is clearly reflexive. It is antisymmetric by the definition of equality between subsets and it is also obviously transitive. In Figure 2.5.2 we see the Hasse diagram associated to this ordering.



Example 2.5.3 (Family trees) *In Figure 2.5.3 we see the family tree of of Ludwig III, Duke of Württenberg, this also can be interpreted as a Hasse diagram associated to an ordering relation δ on the set fo the descendants of Ludwig III (arrows going upwards). Clearly the relation is*

$$a\delta b \text{ if and only if } a = b \text{ or } a \text{ is a descendant of } b$$



Example 2.5.4 (Multiples) *Let $\mathbb{N}$ be the set of natural numbers, and A be a subset of $\mathbb{N}$ define a relation μ on A as follows:*

$$\text{for every } a, b \in A,\ \ a\mu b \text{ if and only if there exists } m \in \mathbb{N} \text{ such that } a = mb$$

(that is a is a multiple of b). We prove that this is an order relation on A. Let a, b, and c be elements of A.

Clearly $a = 1a$ so, since $1 \in \mathbb{N}$, $a\mu a$, giving reflexivity.

Now assume $a\mu b$ and $b\mu a$. The first relation implies that there exists $m \in \mathbb{N}$ such that

$$a = mb, \tag{2.5}$$ {anb}

the second relation implies that there exists $n \in \mathbb{N}$ such that

$$b = na \tag{2.6}$$ {bma}

Substituting in Equation (2.5) the expression for b given in Equation (2.6), we get

$$a = mb = mna$$

which implies that $mn = 1$ and, since both m and n are elements of $\mathbb{N}$, the only possibility is that $m = n = 1$, whence $a = b$, giving antisymmetry.

Finally assume $a\mu b$ and $b\mu c$. As above he first relation implies that there exists $m \in \mathbb{N}$ such that

$$a = mb, \tag{2.7}$$ {amb}

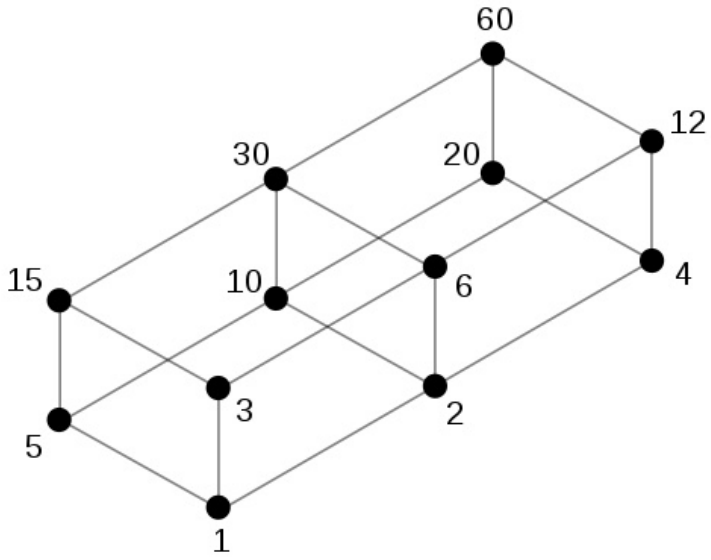
the second relation implies that there exists $n \in \mathbb{N}$ such that

$$b = na \tag{2.8}$$ {bnc}

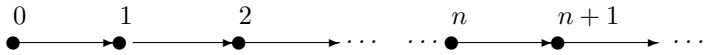
Substituting in Equation (2.8) the expression for b given in Equation (2.7), we get

$$c = mb = mna$$

since both m and n are elements of $\mathbb{N}$, their product mn is also an element of $\mathbb{N}$, $a\mu c$, giving transitivity. Here is the Hasse diagram of μ on the set of divisors of 60 (arrows should go downwards)



Example 2.5.5 (The usual ordering of natural numbers) *We shall give a formal definition of this ordering in Section 2.6. For the moment, this is the ordering* $\leq$ *which we learnt at school, for which* $0 \leq 1$, $2 \leq 10001$, ... *etc. The Hasse diagram is*



The last example is an extreme situation of ordering:

Example 2.5.6 (The equality) *Let* A *be a set and* ϵ *be the equality relation on* A*, that is, for every* a *and* b *in* A*,*

$$a \epsilon b \text{ if and only if } a = b.$$

Let a*,* b*, and* c *be elements of a set* A *clearly* $a = a$*, so* $a \epsilon a$*, and, obviously, if* $a \epsilon b$ *and* $b \epsilon a$*, then* $a = b$*. Finally if* $a \epsilon b$ *and* $b \epsilon c$ *then* $a = b$ *and* $b = c$*, whence* $a = c$*. So equality is also a partial ordering. The Hasse diagram on* A *has no arrows (except for loops, that are omitted, as usual, in case of reflexive relations).*

A *poset* (short for *partially ordered set*) is a pair A, ρ where A is a set and ρ is an order relation on A. Let (A, ρ) be a partially ordered set and B a subset of A (it might help to think ρ as $\leq$). If a and b are two elements of A we say that a and b are *comparable* if either $a \rho b$ or $b \rho a$ (including the case $b = a$). An element m of B is said

maximal if for every $b \in B$, $m \rho b$ implies $m = b$;

maximum or *largest element* if if for every $b \in B$, $b \rho m$;

minimal if for every $b \in B$, $b \rho m$ implies $b = m$;

minimum or *least element* if if for every $b \in B$, $m \rho b$;

Example 2.5.7 (Chain of commands) *The chief of the Army is the maximum and the unique maximal element. The minimal elements are the simple soldiers. If there are more than two simple soldiers (which is a reasonable assumption for a decent army), there is no least element (two different soldiers are not comparable).*

Example 2.5.8 (Inclusion) *A itself is the maximum and the emptyset is the least element*

Example 2.5.9 (Family trees) *This is opposite to the chain of commands: There is a least element (the founder of the dynasty) which is also the unique minimum, and (possibly) many maximal elements (the descendents of the family that have no childs). If the minimal elements are more than one, there is no maximum.*

Example 2.5.10 (Multiples) *1. In $\mathbb{N} \setminus \{1\}$ is the minimum and (strange as it might sound) 0 is the maximum.*

In $\mathbb{N} \setminus \{1\}$ the minimal elements are the prime numbers and 0 is the maximum.

In $\mathbb{N} \setminus \{1, 0\}$ the minimal elements are the prime numbers and there are no maximal elements nor a maximum.

Example 2.5.11 (The usual ordering of natural numbers) *Here 0 is the minimum and there are no maximal elements nor a maximum.*

Example 2.5.12 (The equality) *Here there is no maximum and no minimum (except in case the set contains only one element) and all the elements are both maximal and minimal.*

Note that in all the above examples, when it exists, the maximum (resp. the minimum) is unique. This is true in general:

Lemma 2.5.1 (Uniqueness of maximum and minimum) *Let (A, ρ) be a poset and assume a and b are maximums of A (resp. minimums of A). Then $a = b$*

PROOF. Assume a and b are maximums of A. Since a is a maximum of A and $b \in A$, we have

$$b \rho a,$$

conversely, since b is also a maximum of A and $a \in A$, we have

$$a \rho b,$$

whence $a = b$ by antisymmetry. The proof of the uniqueness of the minimum is the same. ∎

A *totally ordered set* is poset (A, ρ) in which every two elements are comparable. In the above examples, only the poset $\mathbb{N}$ with the usual ordering is totally ordered (the others can be totally ordered only in some trivial cases).

2.5.2 Equivalences and partitions

It is often useful to weaken the concept of equality. For example when we play with Lego bricks and want to put one Lego brick in some position we don't need to put that particular Lego brick, but we are (usually) satisfied with any Lego brick of the same shape and colour of the one we had chosen. We say that two Lego bricks of the same shape and colour are *equivalent* for us. In the sequel of this subsection we'll formalize this concept. Recall that a relation on a set A is a correspondence with domain A and codomain A. If ρ is a relation on the set A and a and b are elements of A such that $(a, b) \in \rho$ then one usually writes $a \rho b$. A relation ρ on a set A is called an *equivalence relation* (or simply an *equivalence*) if the following conditions are satisfied:

reflexivity: for every element a of A, $a \rho a$;

symmetry: for every a, b in A, $a \rho b$ if and only if $b \rho a$;

transitivity: for every a, b, c in A, if $a \rho b$ and $b \rho c$, then also $a \rho c$

In order to stress that equivalences are a sort of weakening of equality, the usual symbols for equivalences are $\sim$, $\approx$, and $\equiv$. There are two extreme equivalences on a set A:

Example 2.5.13 (The equality) *let a, b, and c be elements of a set A clearly $a = a$, and if $a = b$, then $b = a$, finally if $a = b$ and $b = c$ then $a = c$. As a subset of $A \times A$ the equality is the diagonal subset $\{(a, a) | a \in A\}$.*

Example 2.5.14 *At the other extreme we have* **the full cartesian product** *$A \times A$. In this case any two elements of A are equivalent.*

But there are other more interesting examples:

Example 2.5.15 (The equivalence associated to a map) *let $f \colon A \to B$ be a map between two sets A and B, define the equivalence associated to f as the relation $\sim_f$ on A defined, for every a_1 and a_2 in A, by*

$$a_1 \sim_f a_2 \text{ if and only if } f(a_1) = f(a_2)$$

Clearly for every a_1, a_2, and a_3 in A we have $f(a_1) = f(a_1)$, whence $a_1 \sim_f a_1$. Also, if $a_1 \sim_f a_2$, then $f(a_1) = f(a_2)$, whence $f(a_2) = f(a_1)$, that is $a_2 \sim_f a_1$. Finally if $a_1 \sim_f a_2$ and $a_2 \sim_f a_3$, then $f(a_1) = f(a_2)$ and $f(a_2) = f(a_3)$, whence $f(a_1) = f(a_3)$, that is $a_1 \sim_f a_3$.

Example 2.5.16 (The congruence modulo n) *Let n be an integer and define, for every $a, b \in \mathbb{Z}$, the relation $\equiv_n$ on Z as follows:*

$$a \equiv_n b \text{ if and only if } a - b \text{ is a multiple of } n.$$

So, for example, if $n = 2$ any two integers are equivalent if they are either both even or both odd, for the difference of two even or two odd numbers is always a multiple of 2 (we include 0 as a multiple of 2, since $0 = 2 \cdot 0$). Let us prove that $\equiv_n$ is an equivalence. Let a,b, and c be integers. Since $a - a = 0$ and 0 is a multiple of n, we have $a \equiv_n a$. Now assume $a \equiv_n b$ then $a - b$ is a multiple of n, say $a - b = hn$ with $h \in \mathbb{Z}$. then $b - a = (-h)n$ which is also a multiple of n, thus $b \equiv_n a$. Finally suppose $a \equiv_n b$ and $b \equiv_n c$, then $a - b$ and $b - c$ are multiples of n, say $a - b = hn$ and $b - c = kn$ with h and k in $\mathbb{Z}$. Then

$$a - c = (a - b) + (b - c) = hn + kn = (h + k)n,$$

whence also $a - c$ is a multiple of n and so $a \equiv_n c$.

Let $\sim$ be an equivalence on a set A. For every $a \in A$ define the *equivalence class* $\{a\}_\sim$ of a as follows:

$$[a]_\sim := \{b \in A | a \sim b\}.$$

Example 2.5.17 (Equivalence of Lego bricks by shape and colours) *Let $\sim$ be the equivalence on the set L of Lego bricks by shape and colour, that is any two Lego bricks are equivalent if and only if they have the same shape and the same colour. If a is a Lego brick then $[a]_\sim$ is the subset of L of all blocks that have the same shape and colour as a.*

Example 2.5.18 (Train connection) *Let A be a set of cities, define $\sim_t$ to be the relation on A defined , for every $a, b \in A$, as follows:*

$$a \sim_t b$$

if and only if either $a = b$ or one can from a reach b by train and vice versa.

By the way it has been defined, this is clearly a reflexive and symmetric relation on the set A. Moreover it is also transitive, for, if from the city a I can reach by train the city b and from b I can reach by train the city c, then I can also from a reach the city c by train (possibly changing train at b). Given $a \in A$, the equivalence class $[a]_{\sim_t}$ of a is the set of all cities of A that are linked to a by an operating railway line

Example 2.5.19 (Equality) *In this case each equivalence class* $[a]_=$ *contains only the element* a*.*

Example 2.5.20 (The full cartesian product $A \times A$**)** *Opposite to the equality, since every* a *and* b *in* A *are equivalent,* A *is the unique equivalence class and* $[a] = A$ *for every* $a \in A$*.*

Example 2.5.21 (The equivalence $\sim_f$ **associated to a map)** *let* $f\colon A \to B$ *be a map, then, for every* $a \in A$*,*

$$[a]_{\sim_f} = \{a' \in A | f(a) = f(a')\},$$

that is

$$[a]_{\sim_f} = f^{-1}(\{f(a)\}).$$

Example 2.5.22 (Conguence modulo 2**)** *Here there are two equivalence classes:*

the class $[0]_{\equiv_2}$ *consisting of all even integer numbers and*

the class $[1]_{\equiv_2}$ *consisting of all odd integer numbers.*

Let $\sim$ be an equivalence relation on a set A. Denote by $A/\sim$ the set of all equivalence classes $[a]_\sim$ of A. $A/\sim$ is called the *factor set* of A, modulo the equivalence $\sim$. Note that $A/\sim$ is a set of subsets of A.

Example 2.5.23 (Equivalence of Lego bricks by shape and colours) *In this case we can figure the bricks parted into boxes each containing Lego bricks that have the same shape and colour and the factor set is the set of the contents of the boxes (i.e. the elements of the factor set are the sets of Lego bricks that have the same shape and colour).*

Example 2.5.24 (Equality) *In this case* $A/= $ *is the set* $\{\{a\} | a \in A\}$*.*

Example 2.5.25 (The full cartesian product $A \times A$**)** *in this case the factoor set is* $\{A\}$*.*

Example 2.5.26 (The equivalence $\sim_f$ **associated to a map** f**)** *In this case the factor set* $A/\sim_f$ *is the set*

$$\{f^{-1}(\{b\}) | b \in B\}.$$

Example 2.5.27 (Conguence modulo 2**)** *in this case* $\mathbb{Z}/\equiv_2$ *is the set that contains two elements, one element of* $\mathbb{Z}/\equiv_2$ *is the set of even integers, the other element is the set of odd integers.*

The concept of equivalence on a set A is strictly linked with the concept of a partition of the set A. Let A be a set, a *partition* $\mathcal{P}$ of A is a set of subsets of A such that

(i) for every $X \in \mathcal{P}$, $X \neq \emptyset$;

(ii) for every $a \in A$ there is an element X of $\mathcal{P}$ such that $a \in X$;

(iii) for every X and Y in $\mathcal{P}$, either $X = Y$, or $X \cap Y = \emptyset$.

The most important property of the factor set is the following:

{equivpart}

Lemma 2.5.2 *Let A be a set and $\sim$ an equivalence on A. Then, for every $a \in A$,*

1. *for every $a \in A$ $a \in [a]_\sim$;*
2. *for every $a \in A$ and $b \in [a]_\sim$, we have $[a]_\sim = [b]_\sim$;*
3. *in particular $A/\sim$ is a partition of A.*

PROOF. By reflexivity $a \sim a$, hence $a \in [a]_\sim$, proving *1.*. Assume $b \in [a] \sim$ then $a \sim b$, whence $b \sim a$ by simmetry. Now let $c \in [a]_\sim$ then $a \sim c$ and, since $b \sim a$ we get, by transitivity, $b \sim c$, whence $c \in [b]_\sim$. This proves that $[a]_\sim \subseteq [b]_\sim$. Using symmetry and interchanging a and b, we also get the opposite inclusion, proving *2.*. Now *1.* shows that every element of the factor set $A/\sim$ is not empty and for every $a \in A$ there is an element (namely $[a]_\sim$) of $A/\sim$ that contains a as an element. Finally, assume $[a]_\sim \cap [b]_\sim \neq \emptyset$, then, for every $c \in [a]_\sim \cap [b]_\sim$ we have, by *2.*

$$[a]_\sim = [c]_\sim = [b]_\sim,$$

so $A/\sim$ is a partition of A. ∎

Conversely, assume $\mathcal{P}$ is a partition of a set A and define a relation $\sim_\mathcal{P}$ on A as follows: for every a and b elements of A,

$a \sim_\mathcal{P} b$ if and only if there exists an element $X \in \mathcal{P}$ such that $\{a, b\} \subseteq X$

{parteequiv}

Lemma 2.5.3 *Let $\mathcal{P}$ be a partition of a set A, then*

1. *$\sim_\mathcal{P}$ is an equivalence on A;*
2. *$\mathcal{P} = A/\sim_\mathcal{P}$.*

PROOF. Since, for every $a \in A$ there is an element X of $\mathcal{P}$ that contains a, we have $a \sim_\mathcal{P} a$, proving reflexivity. Assume a and b are elements of A such that $a \sim_\mathcal{P} b$, then there exists an element Y of $\mathcal{P}$ such that

$$\{b, a\} = \{a, b\} \subseteq Y,$$

so also $b \sim_\mathcal{P} b$, proving symmetry. Finally assume a, b, and c are elements of A such that

$$a \sim_\mathcal{P} b \text{ and } b \sim_\mathcal{P} c.$$

Then there are elements X and Y in $\mathcal{P}$ such that

$$\{a, b\} \subseteq X \text{ and } \{b, c\} \subseteq Y.$$

In particular $c \in X \cap Y$ and so, since $\mathcal{P}$ is a partition, it must be $X = Y$, whence also $\{a, c\} \subseteq X$, that is $a \sim_{\mathcal{P}} c$, proving transitivity and *1.*. By definition, for every $a \in A$, $[a]_{\sim_{\mathcal{P}}} = X$ where X is the unique element of $\mathcal{P}$ that contains a, giving *2.*. ∎

In other words, the map $\sim \mapsto A/\sim$ is a bijection between the set of equivalence classes of A and the set of partitions of A. The equivalence $\sim_{\mathcal{P}}$ is called the *equivalence on A associated to the partition $\mathcal{P}$ of A*.

2.5.3 The First Homomorphism Theorem for sets

{subsec:firstiososet}

As an application of the concepts of equivalence and partition, we prove now a fundamental, though elementary result.

Given an equivalence $\sim$ on a set A, the map

$$\begin{array}{rccc} \pi_\sim\colon & A & \to & A/\sim \\ & a & \mapsto & [a]_\sim \end{array}$$

that associates each element a of A its equivalence class $[a]_\sim$ is called the *canonical projection* of A onto $A/\sim$ (note that $\pi_\sim$ has to be surjective by the definition of $A/\sim$).

{firstiso}

Theorem 2.5.4 (The First Homomorphism Theorem for sets) *Let $f\colon A \to B$ be a map between two sets A and B and let $\pi = \pi_{\sim_f}$ be the canonical projection of A onto $A/\sim_f$. Then there is a unique map*

$$\bar{f}\colon A/\sim_f$$

such that $\bar{f} \circ \pi = f$. Moreover $\bar{f}$ is injective and $im(\bar{f}) = im(f)$, in particular $\bar{f}$ is a bijection between $A/\sim_f$ and $im(f)$.

Proof. Let $\bar{f}$ be the correspondence between $A/\sim_f$ and B whose elements are the pairs $([a]_{\sim_f}, f(c))$ with $c \in [a]_{\sim_f}$, for every $a \in A$. We prove that $\bar{f}$ is a function, that is, for every $[a]_{\sim_f} \in A/\sim_f$ there is a unique $b \in B$ such that $([a]_{\sim_f}, b)$ is an element of $\bar{f}$. The existence of b is immediate taking $b = f(a)$. To prove the uniqueness of b follows from the fact that if $c \in [a]_{\sim_f}$, then $f(c) = f(a)$ by definition of $\sim_f$. So $\bar{f}$ is a map such that

$$\bar{f}[a]_{\sim_f} = f(a)$$

for every $a \in A$. In particular f and $\bar{f}$ have the same image. Now f and $\bar{f} \circ \pi$ have the same domain (the set A) and the same codomain (the set B) and, for every $a \in A$,

$$\bar{f} \circ \pi(a) = \bar{f}(\pi(a)) = \bar{f}[a]_{\sim_f} = f(a),$$

so $\bar{f} \circ \pi = f$. Finally if assume $[a]_{\sim_f}$ and $[a']_{\sim_f}$ are elements of $A/\sim_f$ such that

$$\bar{f}([a]_{\sim_f}) = \bar{f}([a']_{\sim_f})$$

then, by definition of $\bar{f}$, $f(a) = f(a')$, whence, by definition of $\sim_f$,

$$[a]_{\sim_f} = [a']_{\sim_f},$$

proving that $\bar{f}$ is injective. ∎

The above theorem is extremely useful. To understand its meaning, consider the following situation: there is a courier that has a set of parcels in his storage area and he has to send these parcels to different cities. He has two ways to organize this: either send each parcel individually to its city of destination, or partition all the parcels into blocks containing those parcels that have the same destination and, for each block, send all parcels in a block together to the city of destination. Clearly the number of blocks is equal to the number of the cities of destination.

Now let A be the set of parcels in the storage area and let f be the map that assigns to each parcel its city of destination (the map f sends individually each parcel to its destination). On the other hand define an equivalence $\sim$ on the set A of parcels by $a \sim a'$ if and only if a and a' have to be sent to the same city (a moment's thought shows that $\sim$ is precisely the equivalence $\sim_f$ associated to the map f). The map π is the map that sends each parcel into its equivalence class. The blocks are precisely the equivalence classes $[a]_\sim$ for $a \in A$ and $\bar{f}$ is the map that sends each equivalence class (block) to its destination.

2.5.4 Graphs

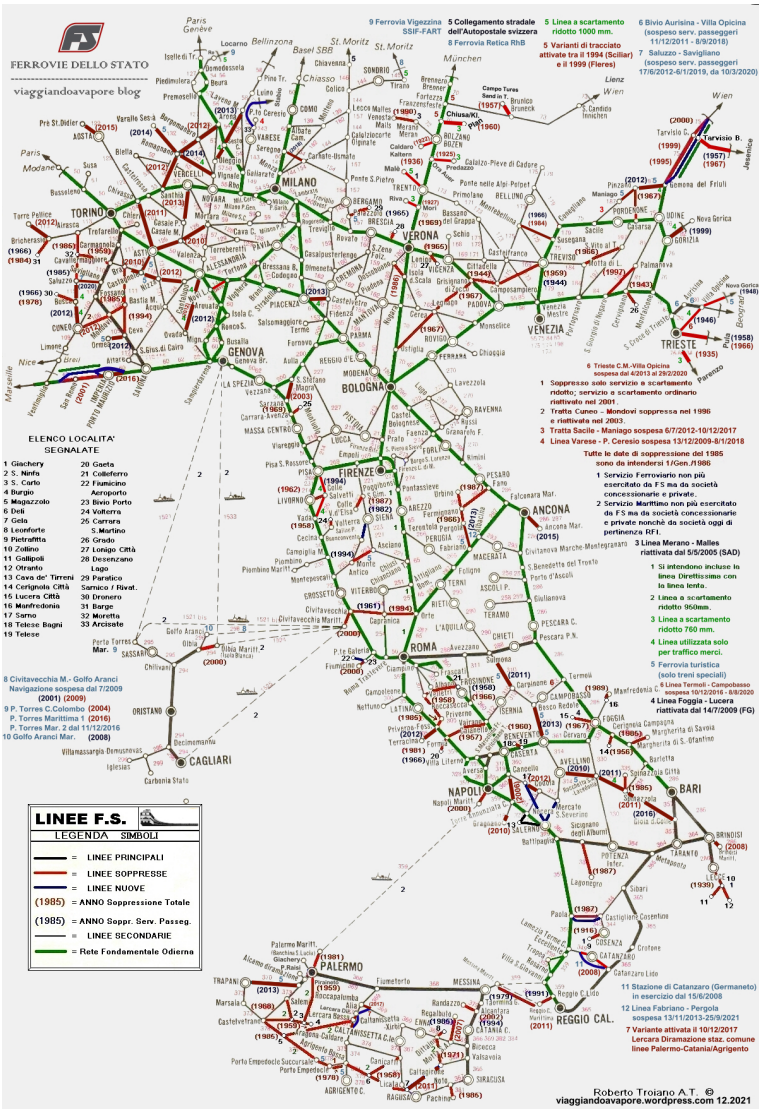
{sec:graphs}

As mentioned in the introduction of this section, given an equivalence relation $\sim$ on a set A is symmetric, the digraph associated to it contains redundant information: namely, since the relation is symmetric, we do not need to specify the direction of the edges. Further we do not need to indicate loops, for there is a loop at every edge. It is therefore convenient to associate to $\sim$ a slightly different object, i.e. an (undirected) graph. A *graph* Γ is a pair (V, E) where V is a set and E is a set of nonempty subsets containing 2 elements of E. The elements of V are called *vertices* of Γ and the elements of E are called *edges* of Γ. Since (as a difference to a pair) a set does not distinguish the order of its elements ($\{a, b\} = \{b, a\}$) in this way we forget the direction of the edges and, since every edge contains two elements, there are no loops in a graph. Also, in the case of an equivalence $\sim$ on a set A, when we want to visualize the associated graph $\Gamma_\sim = (A, E_\sim)$[7] graphically, we can avoid drawing loops and superfluous edges: if $\{a, b\}$ and $\{b, c\}$ are in $E_\sim$ then also $\{a, c\} \in E_\sim$ so we can avois drawing one of these three edges. A *subgaph* of a graph $\Gamma := (V, E)$ is a graph $\Gamma' := (V', E')$ such that $V' \subseteq V$ and $E' \subseteq E$. Finally we say that a graph $\Gamma := (V, E)$ is *finite* if the set of vertices V is finite.

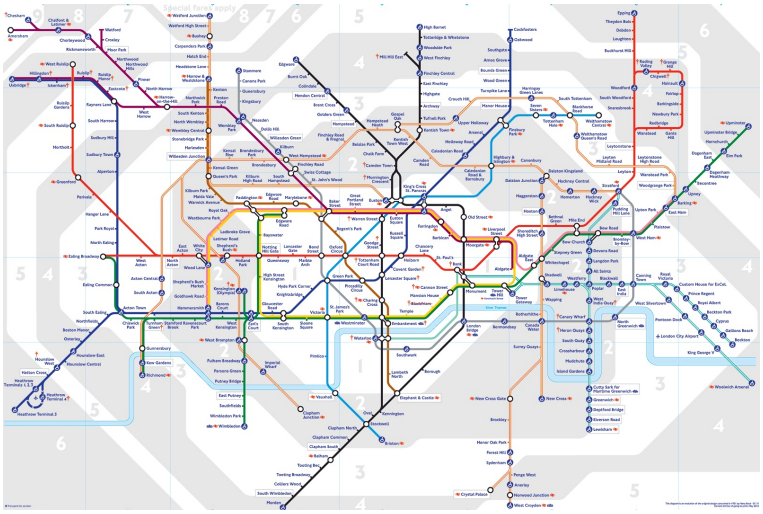
[7] Clearly $E_\sim$ is the set $\{\{a, b\} | a, b \in A, \text{ and } a \sim b\}$

Warning: *In the sequel when we speak of graphs we shall always assume that these graphs are finite.*

Example 2.5.28 (Train connection) *Let A be the set of cities in Italy that have a train station, and let $\sim_t$ the equivalence relation defined in the prevoius subsection. Then the graph associated to $\sim_t$ is the following (consider only black, red and green lines and not dotted lines, nor arrows):*



Example 2.5.29 (The Tube) *Another famous example is the map of the London Underground (forget colours):*



Given a graph $\Gamma := (V, E)$, two vertices a and b and a positive integer n, a *path of length* n in Γ from a to b is an n-tuple

$$p := (\{a_0, a_1\}, \{a_1, a_2\}, \ldots, \{a_{n-1}, a_n\})$$

of edges in E such that

1. $a = a_0$ and $b = a_n$ and
2. for every $i \in \{1, \ldots, n-1\}$, $\{a_{i-1}, a_i\} \neq \{a_i a_{i+1}$.

If the path p is as above, we shall say that p *involves* the edges

$$\{a_0, a_1\}, \{a_1, a_2\}, \ldots, \{a_{n-1}, a_n\}$$

and that the points

$$a_0, a_1, \ldots, a_n$$

are *along* p. Finally the *reverse path* of p is the path

$$p^{-1} := (\{a_n, a_{n-1}\}, \{a_{n-1}, a_{n-2}\}, \ldots, \{a_1, a_0\})$$

In Example 2.5.29, the triple

$$(\{\text{Marble Arch, Bond Street}\}, \{\text{Bond Street, Green Park}\}, \{\text{Green Park, Oxford Circus}\})$$

is a path from Marble Arch to Oxford Circus. We say that a graph Γ is *connected* if, given any two vertices a and b in Γ there is a path in Γ from a to b. Note that the graph in Example 2.5.28 is not connected, since there is no path from Rome to Cagliari. Given a graph $\Gamma = (V, E)$, the relation $\sim_\Gamma$, defined, for every a, b in V, by

$$a \sim_\Gamma b \text{ if and only if there is a path from } a \text{ to } b$$

is an equivalence relation on V and, for every $a \in A$ the equivalence class $[a]_{\sim_\Gamma}$ of a is called the *connected component* containing a. In Example 2.5.28 there are two connected components: one containing Rome (and all other cities in the Italian peninsula and Sicily[8]) and one containing Cagliari (and all other cities in Sardinia). Let n be a positive integer, we say that two different vertices a and b of a graph Γ have *distance* n (and write $d(a, b) = n$), if

1. there is a path of length n from a to b,
2. no other path from a to b has length less than n.

For a vertex a of Γ we define $d(a, a) := 0$ and say that a has distance 0 from itself. If there is no path between two vertices a and b of a graph Γ (or, equivalently, if a belong to two different connected components of Γ) we say that the distance between a and b is *infinite*. For example

$$d(\text{Marble Arch, Bond Street}) = 1,\ d(\text{Marble Arch, Oxford Circus}) = 2$$

(note that the pair

$$(\{\text{Marble Arch, Bond Street}\}, \{\text{Bond Street, Oxford Circus}\})$$

is also a path from Marble Arch to Oxford Street, and it is the shortest one).

A *cycle* in a graph is a path from a vertex a to a itself: in Example 2.5.29, if we exclude the edges from Paddington to Hammersmith in the Circle Line (in yellow), the remaining edges are a cycle from e.g. Victoria to Victoria (or from and to any intermediate stop in that circle). The *girth* of a graph Γ is the length of the shortest cycle in Γ. Check that the graph of the London Underground has girth 3.

{cyclege}

Lemma 2.5.5 *Let $\Gamma := (V, E)$ be a connected graph,*

$$p := (\{a_0, a_1\}, \{a_1, a_2\}, \ldots, \{a_n, a_0\})$$

a cycle in Γ. Then the graph $\Gamma' := (V, E \setminus \{\{a_n, a_0\}\})$ is still connected.

PROOF. Note that Γ and Γ' have the same set V of vertices. Let $a, b \in V$ we show that there is a path in Γ' from a to b. Since Γ is connected, there is a path

$$p_1 :=:= (\{b_0, b_1\}, \{b_1, b_2\}, \ldots, \{b_{t-1}, b_t\})$$

[8] actually the edge {Reggio Calabria Marittima, Messina Marittima} is not by rail but by ferry, however, since the trains are transported in the ferry, we consider it as an edge too.

from $a(= b_0)$ to $b(= b_t)$. If p_1 does not involve $\{a_0, a_n\}$, then p_1 is also a path in Γ' and we are done. Assume p_1 involves $\{a_n, a_0\}$, say $\{a_n, a_0\} = \{b_i, b_{i+1}\}$.

Now, the idea of the proof is to walk along p_1 until one gets to the vertex b_i, and has to cross the edge $\{a_n, a_0\}\{b_i, b_{i+1}\}$. At this point one can walk along the cycle p, or its reversed cycle, to reach b avoiding the edge $\{a_n, a_0\}$.

Let us formalize this: since $\{a_n, a_0\} = \{b_i, b_{i+1}\}$.

$$\text{either } a_0 = b_i \text{ and } a_n = b_{i+1}, \text{ or } a_n = b_i \text{ and } a_0 = b_{i+1}.$$

In the first case, the path

$$(\{b_0, b_1\}, \ldots, \{b_{i-1}, b_i\}, \{a_0, a_1\}, \ldots, \{a_{n-1}, a_n\}, \{b_{i+1}i, b_i + 2\}, \ldots, \{b_{t-1}, b_t\})$$

is a path from a to b and does not involve $\{a_n, a_0\}$, so it is a path in Γ'. The same argument, reverting the cycle p gives the second case. ∎

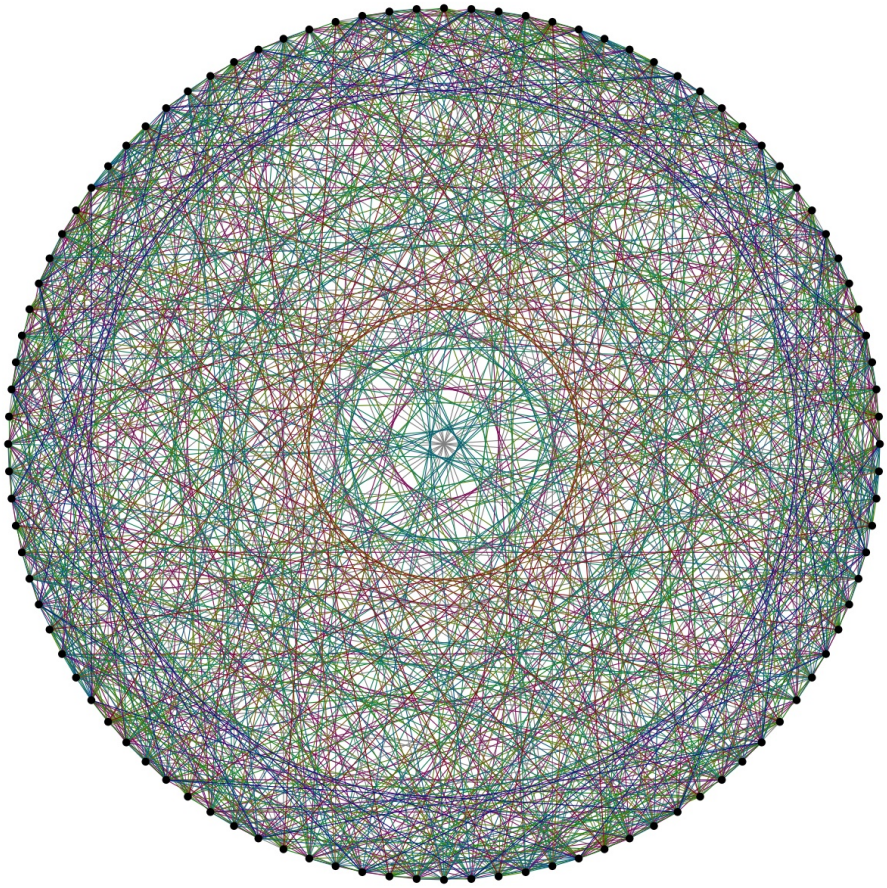
Given a vertex a of a graph Γ, the *valency* of a is the number of edges of Γ containing a. For example the valency of Oxford Circus in Example 2.5.29 of the London Underground is 6, for Oxford Circus is contained in the six edges

1. {Oxford Circus, Regent's Park},
2. {Oxford Circus, Warren Street},
3. {Oxford Circus, Tottenham Court Road},
4. {Oxford Circus, Piccadilly Circus},
5. {Oxford Circus, Green Park},
6. {Oxford Circus, Bond Street},

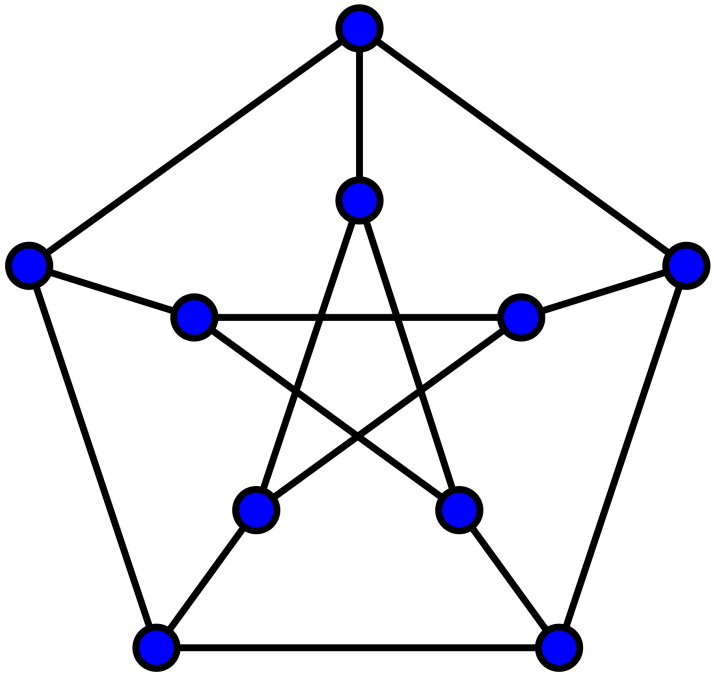
Similarly Regent's Park has valency 2, since it is contained in

1. {Oxford Circus, Regent's Park},
2. {Baker Street, Regent's Park},

A graph is regular if the valency is every vertex has the same valency (that is the valency is constant on vertices). Here's a picture of a regular graph with 100 vertices of valency 22 and 1100 edges (the *Higman-Sims graph*, arising from Group Theory):



And here's another important regular graph of valency 3 and girth 5 that we shall discuss on the next chapter: the Petersen Graph:



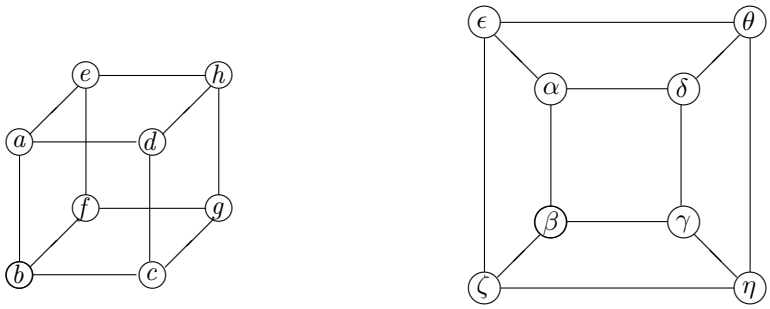
Given two graphs $\Gamma_1 := (V_1, E_1)$ and $\Gamma_2 := (V_2, E_2)$ a map $\phi\colon V_1 \to V_2$ is called a *morphism* of graphs if for every edge $\{a, b\}$ in E_1, $\{\phi(a), \phi(b)\}$ is an edge in E_2. In other words a morphism of graphs is a map between two graphs that *preserves* adjacency.

Example 2.5.30 *Given the two graphs*

the map ϕ that sends a to α, b to β, c to γ, d to δ, and e to ϵ is a (bijective) morphism of graphs. On the other hand, though ϕ^{-1} is still a bijective map, it is not a morphism of graphs, since the edge $\{\gamma, \delta\}$ is mapped to the set $\{c, d\}$, which is not an edge in the first graph.

An *isomorphism* between two graphs is a bijective morphism whose inverse map is still a morphism.

Example 2.5.31 *Consider the following two graphs:*



Check that the map ϕ that sends a to α, b to β, c to γ, d to δ, e to ϵ, f to ζ, g to η, and h to θ is an isomorphism of graphs.

If γ_1 and Γ_2 are graphs and there exists an isomorphism between Γ_1 and Γ_2, we say that the two graphs are *isomorphic*. I want to spend a few words about the concept of isomorphism, which is central in the whole algebra: we shall soon see the analogue for groups and vector spaces. The reason is that when we consider an object from the point of view of (in this case) graphs, we are only interested in the graph theoretical properties of this object, that is

1. which are the vertices,
2. which are the edges.

All other properties (e.g. the way the vertices are named, the way we represent it etc.) do not interest us. This is because every graph theoretical property which is satisfied by a particular graph Γ is also satisfied (possibly after renaming) by any graph isomorphic to Γ.

2.5.5 The Seven Bridges of Königsberg

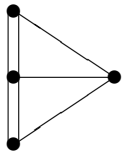
As the reader might expect so far, Graph theory is an immensely vaste research area in Mathematics, with applications ranging from Linguistic to Social Sciences, Physics, Chemistry and, of course, Mathematics. It would be impossible to give even the basics here. Nevertheless I would like to give, as a taste, two simple applications of this theory: Euler's solution to the Königsberg Bridges Problem and the Classification of the Platonic Solids.

I shall give a slightly modified version of the original problem here, but I shall make evident how the solution given here implies the solution of the general problem. The reason is that, in order to solve the general problem, I would need

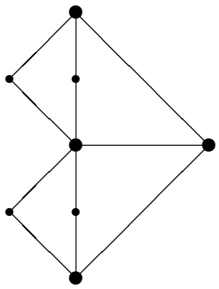
to introduce multigraphs, i.e. graphs that admit more than one edge between two vertices.

Here's the original problem: the city of Königsberg (now Kaliningrad) is crossed by the river Pregel which forms two islands (Kneiphof and Lomse) that are connected to each other and to the rest of the city by seven bridges. In the map below, the island in the centre is Kneiphof, and right of it, connected by one bridge is the island of Lomse, there is also another bridge connecting Lomse to the lower part of the city which is, unfortunaltely, not represented in this map. Reportedly the people of Königsberg used to spend their Sundays looking for a path through the city that crossed once and only once each of the seven bridges. This was very good for their health, but also an immense waste of time and energies (and, living in a touristic city as Venice is, I can easily figure out how overcrowded that bridges should have been on Sundays). Probably annoyed by that, the mathematician Leonard Euler decided to put a stop to all that mess, giving in 1736 a mathematical proof of the non existence of such a path [4]. His proof had enormous consequences for mathematics, setting the fundaments of Graph Theory and foreshadowing modern Algebraic Topology.

We can schematize this situation with the following multigraph, the points being the parts of the city of Kaliningrad and the edges being the bridges:



The problem is now to find an Eulerian path in the above multigraph, i.e. a path that involves once, and only once, all the edges of the above multigraph. In order to avoid using multigraphs, we add, in the above multigraph, four more edges in the middle of the double edges, thus obtaining only single edges. In the picture below, these four additional edges are drawn as smaller dots.



It should be clear that the existence of an Eulerian path in the multigraph in the first picture is equivalent to the existence of an Eulerian path in the graph in the second picture. Of course, one can try all possible paths (they are finitely many) and find out that none of them is Eulerian.

But Euler was more clever and gave a general solution: Euler's observation was that, if I start from one vertex (i.e. one of the two islands or one of the two mainland's bank of Königsberg), say a, and arrive to the vertex b via an Eulerian path, then any other vertex, say c, different from a and b should be connected with an even number of bridges to the others, (since, for each bridge I cross to get into c I also need a bridge to get out of it). In other words, the valency of all vertices should be even, except possibly for a and b. This is clearly not the case for the Königsberg since all vertices (except for the additional ones) have odd valency and these are more than two. The nice thing is that also the converse is true. Euler just stated the converse, but he did not gave any proof. The first proof was given by Carl Hierholzer in 1873 (see [5].

Given a connected graph $\Gamma := (V, E)$, and a path

$$p := (\{a_0, a_1\}, \{a_1, a_2\}, \ldots, \{a_{n-1}, a_n\})$$

in Γ, we say that p is *Eulerian* if

1. $E = \{\{a_0, a_1\}, \{a_1, a_2\}, \ldots, \{a_{n-1}, a_n\}\}$

2. for all $i, j \in \{0, n-1\}$, if $i \neq j$, then $\{a_i, a_{i+1}\} \neq \{a_j, a_{j+1}\}$

We first sketch a proof of the converse when the path is a cycle.

{Euler1}

Theorem 2.5.6 *Let Γ be a connected graph. Then there exists an Eulerian cycle in Γ if and only if all vertices of Γ have even valency.*

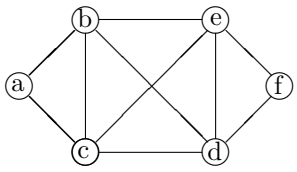
PROOF. First choose any vertex a_1 of Γ and start a random path p_1 from it in such a way that no edge appears more than one in that path. Keep on along this path until you get stuck, say at point a'. Observe that $a' = a_1$ since, by the even valency, every time you get into a vertex x different from a, you'll always find a new edge that lets you out of x (cfr. Euler's remark).

If p_1 involves all the edges of Γ you are done.

Otherwise, since Γ is connected, there has to be a vertex a_2 along p_1 contained in an edge not involved in p_1. Start another cycle p_2 from a_2 in such a way that no edge involved in p_1 is involved in p_2. Then "glue" together at the point a_2 the two cycles p_1 and p_2, obtaining a new cycle, say p_3, involving once and only once all the edges involved in p_1 and p_2.

Connectedness of Γ ensures that repeating this procedure a finite number of times gives eventually an Eulerian cycle in Γ ∎

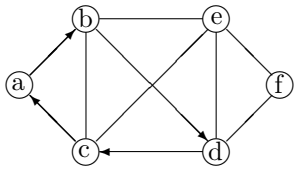
For example, if Γ is the graph below, choose any random path starting at a and never crossing the same edge more than once.



Suppose you've choosen $a_1 = a$ and

$$p_1 := (\{a, b\}, \{b, d\}, \{d, c\}, \{c, a\}).$$

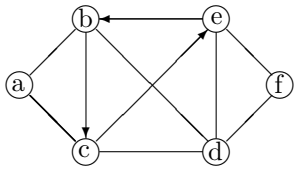
The path p_1 is represented by the arrows in the picture below:



Since p_1 does not involve all the edges of Γ, there must be a vertex x along p_1 such that x is contained in an edge not involved in p_1. For example $\{b, c\}$ is not involved in p_1, so we can choose $x = b$. Now choose a random cycle p_2 that starts and finishes at b and has no edge in common with p_1. For example

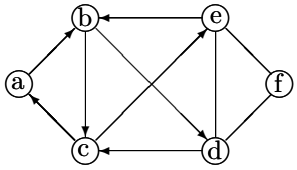
$$p_2 := (\{bc,\}, \{c, e\}, \{e, b\})$$

as in the picture below:



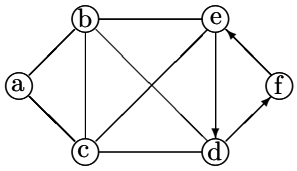
Now, gluing together at the vertex b the two cycles p_1 and p_2 gives the cycle

$$p_3 := (\{a, b\}, \{bc,\}, \{c, e\}, \{e, b\}, \{b, d\}, \{d, c\}, \{c, a\}),$$



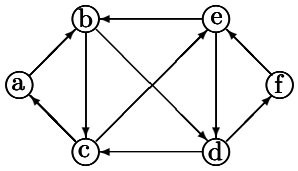
Again the cycle p_3 does not involve all the edges of Γ, and indeed there are vertices (e and d) along p_3 which are contained in edges not involved in p_3. No problem, do the same procedure as above with one of the two vertices e or d: e.g. choose the following cycle starting at e:

$$p_4 := (\{e, d, \}, \{d, f\}, \{f, e\})$$



Finally, gluing together p_3 and p_4 at the vertex e, we obtain the desired Eulerian cycle

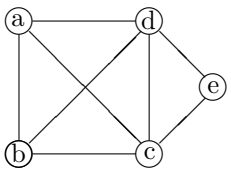
$$(\{a, b\}, \{bc, \}, \{c, e\}, \{e, d, \}, \{d, f\}, \{f, e\}, \{e, b\}, \{b, d\}, \{d, c\}, \{c, a\}),$$



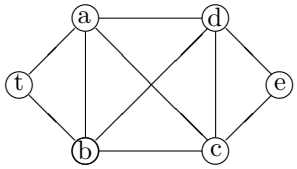
Observe that, if a graph Γ has an Eulerian cycle p, any vertex a along p can be the starting (and finishing) point of an Eulerian path (just shift the cycle p so that it starts at the point a).

Now it is easy to turn to the existence of Eulerian paths. Indeed, if all the vertices of Γ have even valency, then Γ has an Eulerian cycle, which is also an Eulerian path. Suppose the graph Γ has all vertices of even valency except for two, say a and b (a and b distinct). Adjoin to Γ an auxiliary vertex t and the two edges $\{t, a\}$ and $\{t, b\}$. The vertices of new graph Γ' have all even valencies. Indeed all vertices of Γ' except for a, b, and t, have the same valencies as before adjoining t. On the other hand a and b are contained in one extra edge in Γ', hence their valencies become even as vertices in Γ' and clearly t has valency 2. By Theorem 2.5.6 Γ' has an Eulerian cycle p. By the above remark, we can assume that this cycle starts at the vertex t and, possibly reverting p, we may assume that its first edge is $\{t, a\}$. Now removing the edges $\{a, t\}$ and $\{b, t\}$ from p gives an Eulerian path in Γ from a to b.

We show this with an example: Assume Γ is the graph below



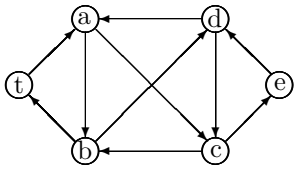
In the above graph all vertices have even valency except for a and b, which have valency 3. Now adjoin to this graph an auxiliary vertex t and the edges $\{t, a\}$ and $\{t, b\}$, as in the picture below:



The new graph has all vertices of even valency, hence it admits an Eulerian cycle, as in the previous example:

.

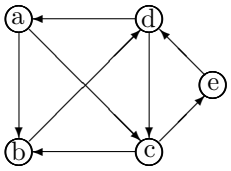
$$(\{t,a\},\{a,b,\},\{b,d\},\{d,c,\},\{c,e\},\{e,d\},\{d,a\},\{a,c\},\{c,b\},\{b,t\}),$$



Now remove the from this cycle the vertex t and the edges $\{t,a\}$ and $\{t,b\}$, and obtain the path

$$(\{a,b,\},\{b,d\},\{d,c,\},\{c,e\},\{e,d\},\{d,a\},\{a,c\},\{c,b\}),$$

as in the picture below:



We have thus proven

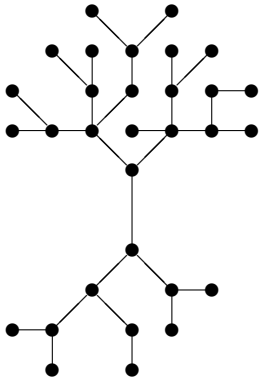
{euler}

Theorem 2.5.7 *Let* Γ *be a connected graph. Then there exists an Eulerian path in* Γ *if and only if all vertices of* Γ *except possibly two, have even valency.*

Epilogue If you look, for example in Google Maps, at a contemporary map of Königsberg, now Kaliningrad, you'll realize that two of the bridges connecting Kneiphof to the upper and lower river banks have been demolished for some unknown reasons: maybe, convinced by Euler about the impossibility of the existence of an Eulerian path, they decided to change the bridges configuration in order to admit such a path. Will the residents find this path now? And, if so, where should it start?

2.5.6 Trees and Planar Graphs

A *tree* is a connected graph that has no circuits. The name has origin from the resemblance of these graphs with real trees, where the edges are the branches, the roots and the trunk and the vertices are the bifurcations of the branches, the roots and the trunk, as in the picture below.



A vertex of valency 1 in a tree is called a *leaf*. Since we assume that trees are finite (see the warning at the beginning of Section 2.5.4), every tree has leaves (we leave the formal proof of this fact as an exercise). Note that if $\Gamma := (V, E)$ is a tree with $|V| > 1$, a is a leaf, and $\{a, b\}$ is the unique edge containing a, then the subgraph $\Gamma' := (V \setminus \{a\}, E \setminus \{\{a, b\}\})$ is still connected and it is a tree, since any cycle in Γ' would be a cycle in Γ too, which is impossible. Trees have an important though elementary combinatorial property:

{Eulertree}

Theorem 2.5.8 (Euler's Formula for Trees) *Let $\Gamma := (V, E)$ be a tree, then*

{Eulertree}
$$|V| - |E| = 1. \tag{2.9}$$

PROOF. We prove the result by induction on $|V|$. If $|V| = 1$ then Γ has one vertex and no edges, and the result follows trivially. Assume $|A| > 1$. Let a be a leaf in Γ and let $\{a, b\}$ the edge containing a. Consider the subgraph $\Gamma' := (V \setminus \{a\}, E \setminus \{\{a, b\}\})$. By the above remark, Γ' is still a tree, so, by the inductive hypothesis, we have

$$1 = |V \setminus \{a\}| - |E \setminus \{\{a, b\}\}| = (|V| - 1) - (|E| - 1) = |V| - |E|.$$

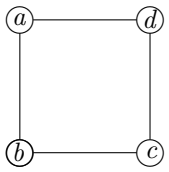
■

A *planar* graph is a graph that can be represented in the plane in such a way that for no pair of edges the line segments representing them intersect only at their endpoints[9].

[9]This is actually not a definition, at least for the way we have defined graphs (besides we haven't defined yet what a plane is). There is a formal definition of planar graph, but

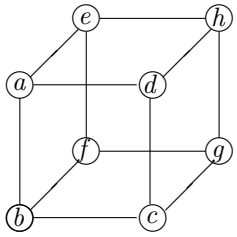
Example 2.5.32 (Trees) *Trees are planar graphs.*

Example 2.5.33 (Square) *The square is a planar graph:*

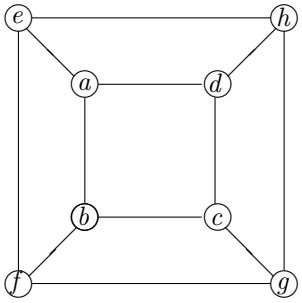


{cube}

Example 2.5.34 (The cube) *Take the graph whose vertices are the vertices of a cube and whose edges are the pairs of endpoints of the edges of the cube:*



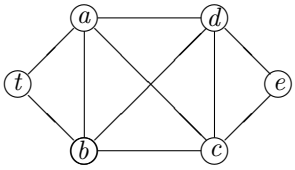
This is also a planar graph since it can also be represented via its orthographic projection as follows:



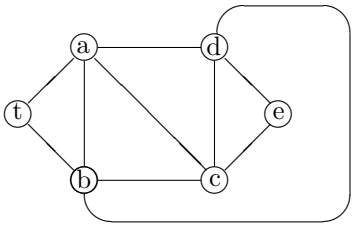
{planar}

Example 2.5.35 *The following graph is also planar*

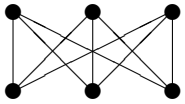
this would bring us much to far away, so we content with this one, trying to neutralize its vagueness with some examples.



Inded, though the edges $\{a, c\}$ and $\{b, d\}$ seem to intersect at their center points, we can draw the same graph making the edge $\{b, d\}$ going around the edges c and e:



Example 2.5.36 *And here's an example of a graph that is not planar (check it by yourself):*



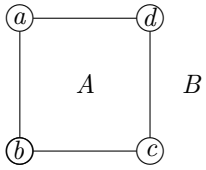
Let $\Gamma := (V, E)$ be a planar graph and represent it on the plane so that the line segments representing two different edges intersect only in their endpoints and consider the set Π obtained taking away the points of all the line segments representing the edges of Γ from the plane. The set Π is partitioned into several *connected components* (or it faces of Γ) that is subsets U such that for every pair (a, b) of points in U there is a (possibly curved) line segment with a and b as endpoints which is fully contained in U. A way to figure out this is to imagine the line segments representing the edges of Γ as canals and the faces as the islands separated by these canals. Connectivity of the faces means that I can walk from any point of an island to any other point of that island without crossing any canal[10].

{treees}

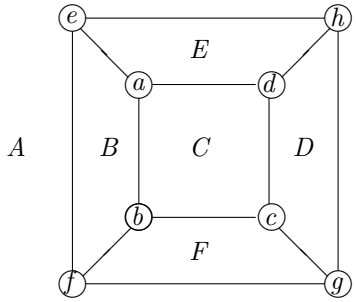
Example 2.5.37 (Trees) *A tree does not disconnect the plane, so a tree has only one face. Conversely, since any cycle in a graph would disconnect the plane (i.e. produce at least two faces), any connected graph with one face is a tree.*

[10]These concepts will be formalized in Section **??**.

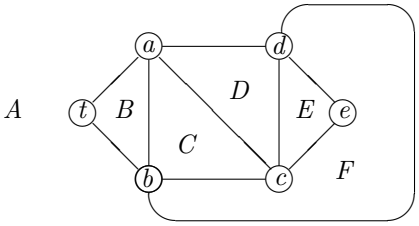
Example 2.5.38 (Square) *The square has two faces: the inside A of the square and its outside B.*



Example 2.5.39 (The cube) *The planar graph constructed in Example 2.5.34 has 6 faces, labelled A, B, C, D, E, and F below and corresponding to the six faces of the cube (the face A is the face with vertices e, f, g, h):*



Example 2.5.40 *The planar graph in Example 2.5.35 has 6 faces labelled A, B, C, D, E, and F below:*



{Eulerform}

Theorem 2.5.9 (Euler's Formula for Planar Graphs) *Let $\Gamma := (V, E)$ be a planar graph and let F be the set of faces of Γ. Then*

$$(|V| - |E|) + |F| = 2 \tag{2.10}$$ {Eulerfo}

Proof. We give a sketch of the proof by induction on the number $|F|$ of faces of Γ. If $|F| = 1$, then Γ is a tree (see Example 2.5.37), thus, by Theorem 2.9, $|V| - |E| = 1$, whence

$$(|V| - |E|) + |F| = 1 + 1 = 2.$$

Assume $|F| = n > 1$ and the result true for every connected graph with less than n faces. Since $n > 1$, Γ has at least one cycle p. Let $\{a, b\}$ be an edge in p and consider the subgraph $\Gamma' := (V, E \setminus \{\{a, b\}\})$. By Lemma 2.5.5 Γ' is still a connected graph and, by removing the edge $\{a, b\}$, we join two faces of Γ getting one face of Γ'. Thus Γ' is a connected graph with $|V|$ vertices, $|E| - 1$ edges, and $|F| - 1$ faces. By the inductive hypothesis applied to Γ', we have

$$(|V| - |E|) + |F| = (|V| - (|E| - 1)) + (|F| - 1) = 2.$$

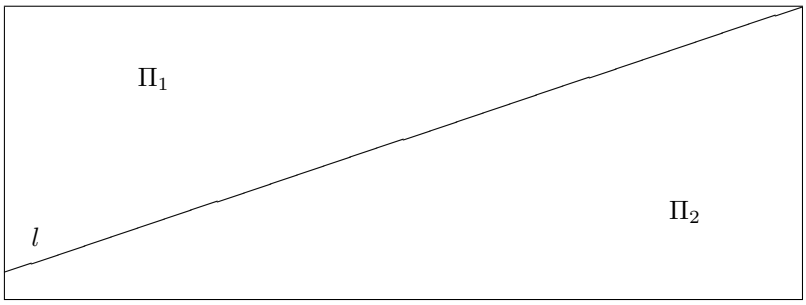
■

{subsec:platonic}

2.5.7 Platonic Solids

Warning: in this subsection I will need some concepts of Euclidean geometry (such as *line, plane, space, circle, etc.*) that we shall formalize only in chapters [**?**, **?**, **?**]. However, assuming that the reader is already confident with these concepts from high school, I decided to put this subsection here not to break continuity in the topic of graphs. The reader who is not confident with the aforementioned concepts might wish to skip this subsection and read it after chapters [**?**, **?**, **?**].

We first define what a convex polygon is. This is a subset of the plane that is bounded by straight lines. More precisely, note that very straight line l in a plane Π divides the plane into two half planes Π_1 and Π_2 such that $\Pi_1 \cap \Pi_2 = l$:



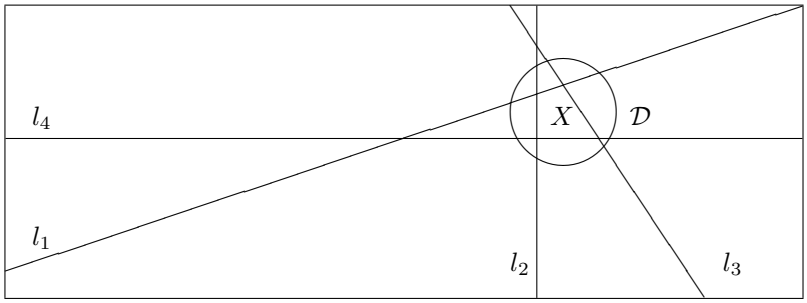
In this case we say that Π_1 and Π_2 are the half planes *associated* to the line l. A subset X of the plane Π is called a *convex polygon* if

1. Convexity: there are lines $l_1, \ldots, l_n$ such that X is the intersection of half planes associated to the lines $l_1, \ldots, l_n$,

2. Finiteness: there is a disk[11] $\mathcal{D}$ in Π such that X is contained in $\mathcal{D}$.

[11] Given a positive real number d and a point c in the Euclidean plane Π, a *disk* of *radius* d and center c is the set of points of Π whose distance from c is less or equal d. This second condition is imposed to exclude infinitely large polygons

Example 2.5.41 *The area marked by X in the picture below is a polygon*



If a builder's image might help you, figure out n bulldozers with infinitely large blades, pushing sand along different, but convergent directions. When they stop the sand will be concentrated in a convex polygon, whose edges are the traces of the blades.

Since any two distinct lines intersect at most in one point, if l_i and l_j are two distinct lines in $\{l_1, \ldots, l_n\}$, then $|l_i \cap l_j \cap X| \leq 1$. If $|l_i \cap l_j \cap X| = 1$ and $a \in l_i \cap l_j \cap X$, we say that v is a *vertex* of X. Moreover, if $l_i \cap X$ is not empty, we shall call $l_i \cap X$ an *edge* of X[12]. A *flag* in X is a pair that (v, e) where v is a vertex of X and e is an edge of X containing v.

In a similar way, any plane Π in the three dimensional Euclidean space E divides the space into two half-spaces E_1 and E_2 intersecting in Π. As above we shall call E_1 and E_2 the half-spaces *associated* to the plane Π. A subset X of E is called a *convex polyhedron* if

1. CONVEXITY: there are planes $\Pi_1, \ldots, \Pi_n$ such that X is the intersection of half-planes associated to $\Pi_1, \ldots, \Pi_n$,

2. FINITENESS: there is a sphere $\mathcal{S}$ such that X is contained in $\mathcal{S}$.

The intersection of each half plane with X is called a *face* of X. Let f_1 and f_2 be two distinct faces of X and let Π_1 and Π_2 the planes containing f_1 and f_2, respectively. Since X is convex, one can show that $f_1 \cap f_2$ is either empty, or is a line segment of the line $\Pi_1 \cap \Pi_2$. In this latter case, we'll call $f_1 \cap f_2$ an *edge* of X. Note that, again by the convexity, an edge of X is contained in exactly two faces. If the intersection of two distinct edges of X is not empty, then it must contain only one point, which is called a *vertex* of X. As for edges and faces, every vertex is contained in exactly two edges. A *flag* in X is a triple (v, e, f)

[12]Note that we use the words *vertex*, *edge*, and, below, *face* in two different contexts: graphs (and planar graphs) and polygons (and polyhedra). Obviously, changing the context, their meanings change too: e.g. an edge in a convex polyhedron is a line segment and contains infinitely many points, while an edge in a graph contains only two points. Nevertheless these concepts are related to each other: in fact, as in Example 2.5.34, the vertices (resp. edges, resp. faces) of the cube correspond to the vertices (resp. edges, resp. faces) of graph (the orthographic projection) associated to it.

where v is a vertex in X, e is an edge in X containing v, and f is a face in X containing e.

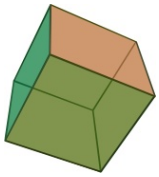
A permutation ϕ of the points in the three dimensional space E is called an *isometry* if it preserves the distances between points: that is if a and b are points in E, then the distance between a and b is the same as the distance between $\phi(a)$ and $\phi(b)$. A polygon X is called *regular* if given any two flags (v, e) and (v', e') in X, there is an isometry ϕ of E such that $v' = \phi(v)$ (note that this definition is equivalent to the definition of regular polygon given at school).

A polyhedron X is called *regular* if given any two flags (v, e, f) and (v', e', f') in X, there is an isometry ϕ of E such that $v' = \phi(v)$, $e' = \phi(e)$, and $f' = \phi(f)$ (intuitively a convex polyhedron is regular if it is highly symmetric since its faces are *congruent* regular polygons which are *assembled* in the same way around each vertex).

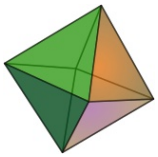
A convex regular polyhedra in a three dimensional Euclidean space E is called a *Platonic solid*. Here are the Platonic Solids:

Example 2.5.42 (The Tetrahedron) *The tetrahedron is a convex regular polyhedron with four vertices, six edges and four faces. Each vertex is contained in exactly three edges, each edge is contained in exactly two faces, and each face contains exactly three edges and three vertices.*

Example 2.5.43 (The Cube) *The cube is a convex regular polyhedron with eight vertices, twelve edges and six faces. Each vertex is contained in exactly three edges, each edge is contained in exactly two faces, and each face contains exactly four edges and four vertices.*



Example 2.5.44 (The Octahedron) *The octahedron is a convex regular polyhedron with six vertices, twelve edges and eight faces. Each vertex is contained in exactly four edges, each edge is contained in exactly two faces, and each face contains exactly three edges and three vertices.*



Example 2.5.45 (The Dodecahedron) *The dodecahedron is a convex regular polyhedron with twenty vertices, thirty edges and twelve faces. Each vertex is contained in exactly three edges, each edge is contained in exactly two faces, and each face contains exactly five edges and five vertices.*

Example 2.5.46 (The Icosahedron) *The icosahedron is a regular convex polyhedron with twelve vertices, thirty edges, and twenty faces. Each vertex is contained in exactly five edges, each edge is contained in exactly two faces, and each face contains exactly three edges and three vertices.*

Are there any others? No! These are the only platonic solids. This fact is quite fascinating (at least for someone, including me). They were known already in ancient times and have been thoroughly investigated by Greek philosophers and mathematicians. Theaetetus was probably the first to show the existence and uniqueness of the five Platonic solids. Plato discussed them in one of his dialogues (*Timaeus*) where he associates to each of the four classical elements (earth, air, water, and fire), a platonic solid, (the cube, the octahedron, the icosahedron and the tetrahedron respectively). About the dodecahedron, Plato says it was used by "god to arrange the constellations in the whole even". This probably led Aristoteles to add a new element, the ether, of which heavens are constituted. Probably on the basis of Theaetetus work, Euclid gave a complete mathematical description of these solids in the last book (XII) of his *Elements*[13].

[13]Hermann Weyl wrote that the Swiss mathematician "...Andreas Speiser has advocated the view that the construction of the five regular solids is the chief goal of the deductive system of geometry as erected by the Greeks canonized in Euclid's *Elements*" [12, p.74]

We shall prove the *uniqueness* of the platonic solid first associating to every platonic solid X a planar graph $\Gamma(X)$ (its *orthographic projection*) and then using Euler's Formula for Planar Graphs 2.5.9 we'll show that there are (up to isomorphism) only five distinct possibilities for $\Gamma(X)$, whence it will follow that there are only five shapes of platonic solids.

Given a convex polihedron X let $\Gamma(X)$ be the graph defined in the following way:

> the vertices of $\Gamma(X)$ are the vertices of X
>
> given two distinct vertices a and b of $\Gamma(X)$, the set $\{a, b\}$ is an edge of $\Gamma(X)$ if and only if $[a, b]$ is an edge of X

It is not immediate to prove that $\Gamma(X)$ is planar, but we can figure that out imagining the edges of X made of an elastic band. Then choose a face and stretch the edges of X that face by pulling towards diverging directions the vertices contained in that chosen face until you'll have reduced it to (the representation of) a planar graph (the orthographic projection of X)[14]. Note that if f is a face of X, the set of its vertices is a face of the planar graph $\Gamma(X)$.

Now let X be a Platonic solid with l vertices, m edges, and n faces. As remarked above, the planar graph $\Gamma(X)$ has also l vertices, m edges and n faces. Moreover, since X is regular, $\Gamma(X)$ is also regular and every face of $\Gamma(X)$ contains the same number of vertices. Let p be the valency of the vertices of $\Gamma(X)$ and let q be the number of vertices (and of edges) contained in each face of $\Gamma(X)$. We now apply a fundamental combinatorial trick:

COUNT THE NUMBER OF EDGES IN TWO WAYS AND COMPARE THE RESULTS

First way: we know that every vertex is contained in exactly p edges, and every edge contains exactly two vertices, so, multiplying the number of edges by the valency we get twice the number of edges: $2m = pl$, whence

{edge1}
$$l = \frac{2m}{p} \tag{2.11}$$

Second way: we also know that each face contains exactly q faces and each edge is contained in two different faces. So, if we multiply q by the number n of faces, we get that $2m = nq$, whence

{edge2}
$$n = \frac{2m}{q} \tag{2.12}$$

But now substituting Equations (2.11) and (2.12) in Euler's Formula for Planar Graphs 2.5.9 we get

{edge3}
$$m + 2 = l + n - 2 = \frac{2m}{p} + \frac{2m}{q} = 2m\left(\frac{1}{p} + \frac{1}{q}\right). \tag{2.13}$$

[14] *Stretching* is what mathematicians call a *continuous transformation* i.e. a transformation of a shape that does not pierce, nor tear this shape. Topology is the branch of mathematics that studies the shapes up to continuous transormations. I shall give a taste of it in Chapter **??**.

Since $m > 1$ we can simplify the above equation and get

{edge4}
$$\frac{1}{p} + \frac{1}{q} = \frac{1}{2} + \frac{1}{m} \tag{2.14}$$

which leads to the famous[15] Diophantine inequality:

{edge5}
$$\frac{1}{p} + \frac{1}{q} > \frac{1}{2} \tag{2.15}$$

Diophantine means that the solutions p and q (and m for Equation (2.14)) should be found within the integer numbers, and actually p has to be at least 2 because every vertex is the intersection of (at least) two edges, and q is at least 3, since a face is a polygon which has at least three edges. This makes the possibilites quite restricted. Let's check:

1. Assume by contradiction that $p = 2$: then, by Equation (2.14), $m = q$. Let f be a face of $\Gamma(X)$. Then all the q edges of $\Gamma(X)$ are contained in f, but then all the vertices of $\Gamma(X)$ have to be contained in f. So $\Gamma(X)$ has q vertices and q edges and only one face so it cannot be associated to a Platonic solid. Thus

   p and q have to be at least 3.

2. Assume, by contradiction, that $p \geq 3$ and $q \geq 6$: then $\frac{1}{p} + \frac{1}{q} \leq \frac{1}{3} + \frac{1}{6} = \frac{1}{2}$, which is excluded by the Inequality (2.15).

3. Similarly it cannot be $p \geq 6$ and $q \geq 3$, so

   p and q can be at most 5.

4. Assume $p = 3$ and $q = 3$: then, by Equations (2.11), (2.12), and (2.14), $l = 4$, $m = 6$, and $n = 4$. So the unique possibility for X is to be the tetrahedron.

5. Assume $p = 3$ and $q = 4$: then, by Equations (2.11), (2.12), and (2.14), $l = 8$, $m = 12$, and $n = 6$. So the unique possibility for X is to be the square.

6. Assume $p = 3$ and $q = 5$: then, by Equations (2.11), (2.12), and (2.14), $l = 20$, $m = 30$, and $n = 12$. So the unique possibility for X is to be the dodecahedron.

7. Assume $p = 4$ and $q = 3$: then, by Equations (2.11), (2.12), and (2.14), $l = 6$, $m = 12$, and $n = 8$. So the unique possibility for X is to be the octahedron.

[15] Famous, because it arises also in other branches of mathematics, e.g. the classification of the semisimple Lie algebras over the complex field

8. Assume, by contradiction, that $p \geq 4$ and $q \geq 4$: then $\frac{1}{p} + \frac{1}{q} \leq \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$, which is excluded by the Inequality (2.15).

9. Finally assume $p = 5$ and $q = 3$: then, by Equations (2.11), (2.12), and (2.14), $l = 12$, $m = 30$, and $n = 20$. So the unique possibility for X is to be the icosahedron.

2.5.8 Why orthographic projection?

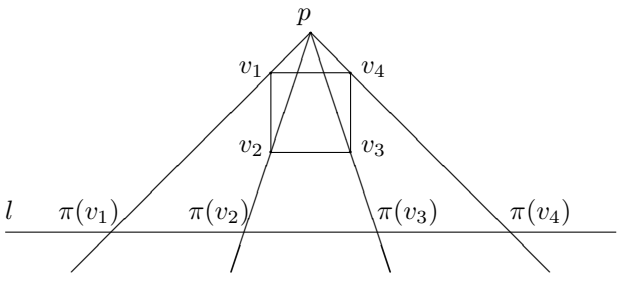
Let X be convex polygon with n vertices and n edges. Let, $v_1, \ldots, v_n$ be its vertices. Set $v_{n+1} := v_1$ and denote its edges $e_1, \ldots, e_n$ in such a way that, for every $i \in \{1, \ldots, n\}$, v_i and v_{i+1} are the endpoints of e_i. We can *project* X on a straight line as follows: Take a point p not contained in X and take a straight line l not containing p. Since p is not in X, for every vertex v of X there is a unique straight line $l_{p,v}$ through p and v and, since l does not contain p, $l_{p,v}$ and l intersect exactly in one point. Call $\phi(v)$ this point and let $\phi\colon Y \to l$ be the map that associates to every vertex v of X the point $\phi(v)$ of l. Let $\Gamma(X)$ be the graph whose vertices are $\pi(v_1), \ldots, \pi(v_n)$ and whose edges are the sets $\{\phi(v_i), \phi(v_i + 1)\}$, for $i \in \{1, \ldots, n\}$. Since X is convex, we can choose p in such a way that ϕ is injective and, for every pair of different edges e_i and e_j, the line segments of l with endpoints $\pi(v_i)$, $\pi(v_{i+1})$ and $\pi(v_j)$, $\pi(v_{j+1})$ intersect at most in one of their endpoints. In this case, we shall call $\Gamma(X)$ the *orthographic projection* of X.

Example 2.5.47 (The orthographic projection of a square) *we illustrate the above procedure with a picture. X is the square with vertices v_1, v_2, v_3, v_4. The graph $\Gamma(X)$ is the graph whose vertices are*

$$\phi(v_1),\ \phi(v_2),\ \phi(v_3),\ \text{and}\ \phi(v_4)$$

and whose edges are

$$\{\phi(v_1), \phi(v_2)\},\ \{\phi(v_2), \phi(v_3\},\ \{\phi(v_3), \phi(v_4)\},\ \text{and}\ \{\phi(v_1), \phi(v_4)\}.$$



We can proceed analogously in three dimensions. Let X be a convex polyhedron, V the set of vertices of X, p a point not in X, and Π a plane not containing p. For each vertex x of X let $\pi(x)$ be the intersection. Since p is not in X, for every vertex x of X there is a unique straight line $l_{p,x}$ through p

and x and, since Π does not contain p and is parallel only to the edge e, $l_{p,x}$ and l intersect exactly in one point. Call $\phi(x)$ this point and let $\phi\colon V \to l$ be the map that associates to every vertex x of X the point $\phi(x)$ of l. Let $\Gamma(X)$ be the graph whose vertices are $\pi(v_1), \ldots, \pi(v_n)$ and whose edges are the sets $\{\phi(v_i), \phi(v_i+1)\}$, for $i \in \{1, \ldots, n\}$. As above, since X is convex, we can choose p in such a way that ϕ is injective and Γ is a planar graph. In this case, we shall call $\Gamma(X)$ the *orthographic projection* of X. Note that also the edges (resp. the faces) of X are mapped bijectively to edges (resp. faces) of $\Gamma(X)$.

Example 2.5.48 *Here is the orthographic projection of a tetrahedron with*

vertices a, b, c, and d,

edges $[a,b]$, $[a,c]$, $[a,d]$, $[b,c]$, $[b,d]$, and $[c,d]$,

faces $[a,b,c]$, $[a,b,d]$, $[a,c,d]$, and $[b,c,d]$.

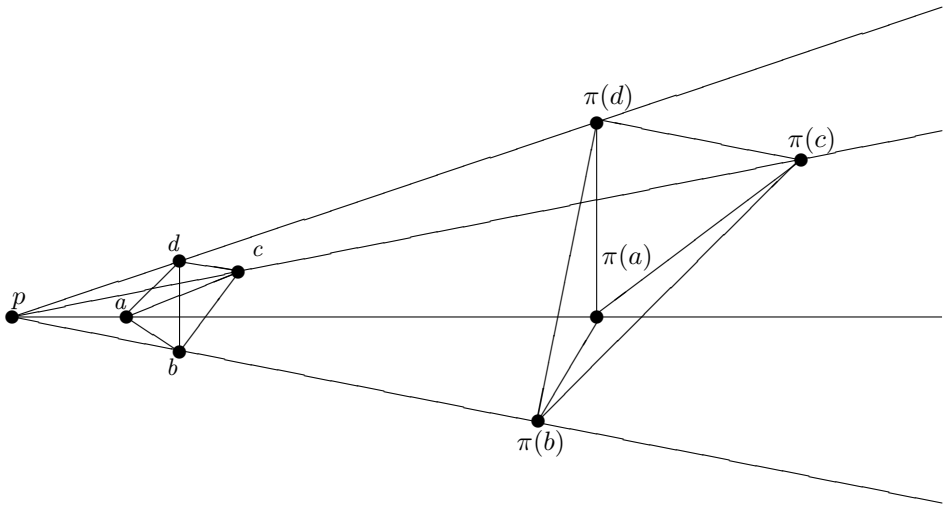
Its orthographic projection is the planar graph with

vertices $\pi(a)$, $\pi(b)$, $\pi(c)$, and $\pi(d)$,

edges $\{\pi(a),\pi(b)\}$, $\{\pi(a),\pi(c)\}$, $\{\pi(a),\pi(d)\}$, $\{\pi(b),\pi(c)\}$, $\{\pi(b),\pi(d)\}$, and $\{\pi(c),\pi(d)\}$,

faces $\{\pi(a),\pi(b),\pi(c)\}$, $\{\pi(a),\pi(b),\pi(d)\}$, $\{\pi(a),\pi(c),\pi(d)\}$, and $\{\pi(b),\pi(c),\pi(d)\}$

(In the picture below, imagine the points $\pi(a)$, $\pi(b)$, $\pi(c)$, and $\pi(d)$ as lying on the same plane).



2.6 Construction of $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$

{nnumbers}

In this secion we give a sketch of how the sets of natural, integer, rational, real, and complex numbers can be constructed out of set theory. The interested reader can consult any of the many treatments on axiomatic set theory, e.g. [8, Appendix]).

2.6.1 The natural numbers

The formal definition of the set $\mathbb{N}$ of natural numbers is the set whose elements are

$$\begin{aligned} 0 &:= \emptyset, \\ 1 &:= \{\emptyset\}, \\ 2 &:= \{\emptyset, \{\emptyset\}\}, \\ 3 &:= \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}, \\ 4 &:= \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}\}, \end{aligned}$$

and so on by the general rule

{sum1}

$$n + 1 := \{n, \{n\}\}. \tag{2.16}$$

Note that

1. (Total Ordering given any two natural numbers m and n, either $m \in n$, or $n \in m$ or $m = n$;
2. $0 = \emptyset \in n$ for every natural number n,
3. (Minimum Principle) given any nonempty subset X of the set $\mathbb{N}$ of the natural numbers, the intersection m of all the elements of X is still a natural number which is called the *minimum element* of X.

For m and n in $\mathbb{N}$, we'll write $m \leq_{\mathbb{N}} n$ (or simply $m \leq n$) if $m \in n$ or $m = n$. By the Total Ordering, $(\mathbb{N}, \leq_{\mathbb{N}})$ is a totally ordered set. A totally ordered set satisfying the Minimum Principle is called a *well ordered set*. So $(\mathbb{N}, \leq_{\mathbb{N}})$ is also a well ordered set.

We are now ready to prove some of the results we used in Subsection 2.4.3:

{uniquen0}

Lemma 2.6.1 *Let m and n be positive integers with $m \leq n$ and let $f\colon Y \to \{1, \ldots, n\}$ be a surjective map. Then $m = n$ and f is also injective, hence bijective.*

Proof. Note first that we may assume, without loss of generality, that $m \in f^{-1}\{n\}$, for, otherwise, we may consider the map

$$\sigma\colon \{1, \ldots n\} \to \{1, \ldots, n\}$$

defined by

$$\sigma(i) = i, \text{ if } i \in \{1, \dots, n\} \setminus \{f(m), n\}, \ \sigma(f(m)) = n, \text{ and } \sigma(n) = f(m).$$

A direct check shows that σ is a bijection, hence, by Lemma 2.4.2 the map

$$g := \sigma \circ f \colon \{1, \dots, m\} \to \{1, \dots, n\}$$

is surjective and

$$g(m) = \sigma(f(m)) = n.$$

Thus we may use g instead of f. We prove the result by induction on m. If $m = 1$ then $\{1, \dots, n\} = f(\{1\})$, whence $n = 1$, since f is surjective.Thus $f = id_{\{1\}}$ and the result follows. Assume $m > 1$ and the result true for $m - 1$. Let f' be the restriction of f to the set $Y := \{1, \dots, m-1\}$. Note first that $n \not\in f(Y)$, for otherwise, our inductive hypothesys would imply that $m - 1 = n$ whence $m > n$, against the assumption that $m \leq n$. Thus f' is actually a map from $\{1, \dots, m-1\}$ to $\{1, \dots, n-1\}$ and it is surjective, for if there were $j \in \{1, \dots, n-1\}$, such that

$$j \notin f'(\{1, \dots, m-1\}) = f(\{1, \dots, m-1\}),$$

then, since $f(m) = n \neq j$, we would have

$$j \notin f(\{1, \dots, m-1\}),$$

against the hypothesys that f is surjective. So f' maps surjectively $\{1, \dots, m-1\}$ onto $\{1, \dots, n-1\}$ and $m - 1 \leq n - 1$, since $m \leq n$. Hence, by the inductive hypothesys, $m - 1 = n - 1$ (so $m = n$) and the map

$$f' \colon \{1, \dots, m-1\} \to \{1, \dots, n-1\}$$

is bijective. Since $f(m) = f(n) = n$ while $f'(j) = f(j)$, for every $j \in \{1, \dots, m-1\}$, f' is injective and $n \notin f(\{1, \dots, m-1\}$, it follows, by a direct check, that f is injective. ∎

{uniquen}

Corollary 2.6.2 *Let n and m be natural numbers greater or equal to 1 and suppose there is a bijection f between $\{1, \dots, n\}$ and $\{1, \dots, m\}$. Then $n = m$*

PROOF. If $m \leq n$, then the resutl follows from Lemma 2.6.1, since a bijective map is in particular surjective . If $n \leq m$, just interchange the roles of m and n and apply again Lemma 2.6.1. ∎

Corollary 2.6.3 *Let X be a finite set, n and m in $\mathbb{N} \setminus \{0\}$ be such that there exist a bijection*

$$f \colon X \to \{1, \dots, n\}$$

and a bijection

$$g \colon X \to \{1, \dots, m\}.$$

Then $m = n$.

PROOF. Since g is bijective, by Lemma 2.4.1, its inverse correspondence $g^{-}1$ is also a bijective function from $\{1,\dots,m\}$ to X. Then, by Lemma 2.4.2, $f \circ g^{-}1$ is a bijection between

$$\{1,\dots,m\} \text{ and } \{1,\dots,n\},$$

whence $n = m$ by Corollary 2.6.2. ∎

Here are two obvious but important consequences:

{countcor}

Corollary 2.6.4 *Let X and Y be two finite sets. Then there exists a bijection from X to Y, if and only if $|X| = |Y|$.*

PROOF. Let

$$\delta_X \colon X \to \{1,\dots,|X|\} \text{ and } \delta_Y \colon Y \to \{1,\dots,|Y|\}$$

be bijections. Assume there exists a bijection $\phi \colon X \to Y$, then, by Lemma 2.4.1 and Lemma 2.4.2, the composition

$$\delta_Y \circ \phi \circ \delta_X^{-1}$$

is a bijective map from $\{1,\dots,|X|\}$ to $\{1,\dots,|Y|\}$, so the result follows by Lemma 2.6.2. Conversely assume $|X| = |Y|$, then

$$\delta_Y^{-1} \circ \delta_X$$

is a bijection between X and Y. ∎

{pig}

Corollary 2.6.5 *Let X be a finite set and Y be a subset of X. Then $X = Y$ if and only if $|X| = |Y|$.*

PROOF. Obviously, if $X = Y$, then also $|X| = |Y|$. Conversely, assume $|X| = |Y|$ and set $n = |X|$. We prove the result by induction on n. If $n = 0$, then $X = Y = \emptyset$. Assume $n \geq 1$ and the result true for sets of cardinality smaller than n. Let $\delta_Y \colon Y \to \{1,\dots,n\}$ be as in the proof of Corollary 2.6.4 and let $y \in Y$ be such that

$$\delta_Y(y) = n.$$

Consider the sets

$$X' := X \setminus \{y\} \text{ and } Y' := Y \setminus \{y\}$$

Then

$$Y' \subseteq X' \text{ and } |X'| = |Y'| = n - 1,$$

therefore, by the inductive hypothesys,

$$Y' = X'$$

whence

$$Y = Y' \cup \{y\} = X' \cup \{y\} = X'.$$

■

We can define an operation + on $\mathbb{N}$ (i.e. a map $+\colon \mathbb{N}\times\mathbb{N}\to\mathbb{N}$) by induction in the following way: for every $n\in\mathbb{N}$ define $n+0:=n$ and $n+1$, as in Equation (2.16). Assume now $m\in\mathbb{N}$ with $1\leq m$ and suppose you have defined $n+m$, say $m+n=t$. Then define

$$n+(m+1):=n+\{m,\{m\}\}:=\{n+m,\{n+m\}\}:=(n+m)+1 \qquad (2.17) \quad \{\texttt{sum2}\}$$

Let us make an example to understand what we have done:

Example 2.6.1 *Assume you want to define* $3+2$

1. *First you know, by definition (i.e. Equation (2.16)), that*

$$3+1=\{3,\{3\}\}=4,$$

which, written explicitly reads:

$$\{\emptyset,\{\emptyset\},\{\emptyset,\{\emptyset\}\}\}+\{\emptyset\}=\{\emptyset,\{\emptyset\},\{\emptyset,\{\emptyset\}\}\},\ \{\{\emptyset,\{\emptyset\},\{\emptyset,\{\emptyset\}\}\}\}.$$

2. *Now* $2=1+1$, *so* $3+2=3+(1+1)$
3. *Therefore, by Equation (2.17),* $3+2=3+(1+1):=(3+1)+1$.
4. *By Equation (2.16),* $3+1=4$,
5. *so* $3+2=(3+1)+1=4+1$,
6. *and finally, by Equation (2.16),* $4+1=5$, *so*

$$3+2=3+(1+1)=(3+1)+1=4+1=5.$$

No wonder it this sounds difficult, it usually takes years to understand this at the ground school.

The operation + is called *addition*. One can prove by induction the usual properties of the addition:

{assoc}

Lemma 2.6.6 (Associativity) *For every* $a,b,c\in\mathbb{N}$, $(a+b)+c=a+(b+c)$

Proof. Clearly, if $c=0$,

$$a+(b+0)=a+b=(a+b)+0.$$

Assume, by inductive hypothesis, that

$$a+(b+c)=(a+b)+c. \qquad (2.18) \quad \{\texttt{asso2}\}$$

We prove that it is also true for $c+1$. By the definition of addition $b+(c+1)=(b+c)+1$ and $((a+b)+c)+1=(a+b)+(c+1)$, whence, by Equation (2.18),

$$a+(b+(c+1))=a+(b+c)+1)=(a+(b+c))+1=((a+b)+c)+1=(a+b)+(c+1)$$

■

{simmet0}

Lemma 2.6.7 *For every $a \in \mathbb{N}$, $0 + a = a$*[16].

PROOF. By induction on a. If $a = 0$ or $a = 1$ the result is trivial: in the first case

$$0 + a = 0 + 0 = 0 = 0 + 0 = a + 0,$$

in the second case, by the definitions of $0 + 1$ and of $1 + 0$, we have

{sum3}
$$0 + a = 0 + 1 = 1 = 1 + 0 + 0 = a + 0. \tag{2.19}$$

Assume, by induction, that $0 + a = a + 0$, then, by Equations (2.17), 2.19), and the definition of the addition,

$$0 + (a + 1) = (0 + a) + 1 = a + 1 = (a + 1) + 0.$$

■

{simmet1}

Lemma 2.6.8 *For every $a \in \mathbb{N}$, $a + 1 = 1 + a$*

PROOF. If $a = 0$ the result follows by Lemma 2.6.7. Assume, by inductive hypothesis, that

{simmet12}
$$a + 1 = 1 + a. \tag{2.20}$$

Then, by Equation (2.6.7) and Lemma 2.6.12,

$$(a + 1) + 1 = (1 + a) + 1 = 1 + (a + 1).$$

■

{simmet}

Lemma 2.6.9 (SYMMETRY) *For every $a, b \in \mathbb{N}$, $a + b = b + a$*

PROOF. If $a = 0$, the result follows from Lemma 2.6.8. Assume, by induction, that

{simo}
$$a + b = b + a \tag{2.21}$$

for every $b \in \mathbb{N}$. Then, by Lemma 2.6.12, Lemma 2.6.8, and Equation (2.21),

$$(a + 1) + b = a + (1 + b) = a + (b + 1) = (a + b) + 1 = (b + a) + 1 = b + (a + 1).$$

■

The multiplication $\cdot$ is defined analogously: for every $n \in \mathbb{N}$

1. $0 \cdot n := 0$;

[16] Note that we have defined what $a + 0$ is, but we have not proven that $a + 0 = 0 + a$, nor, more generally, that $a + b = b + a$ for every pair of natural numbers (a, b).

2. $1 \cdot n := n$;

3. $(a+1) \cdot n := (a \cdot n) + n$

Using induction, one can prove the usual properties of the multiplication:

{assoc}

Lemma 2.6.10 (ASSOCIATIVITY) *For every $a, b, c \in \mathbb{N}$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$*

{simmet}

Lemma 2.6.11 (SYMMETRY) *For every $a, b \in \mathbb{N}$, $a \cdot b = b \cdot a$*

{assoc}

Lemma 2.6.12 (DISTIBUTIVITY) *For every $a, b, c \in \mathbb{N}$, $(a+b) \cdot c = (a \cdot c) + (b \cdot c)$*

{cancel}

Lemma 2.6.13 (CANCELLATION LAW[17]) *Let $a, a' \in \mathbb{N}$, then*

1. *For every $c, \in \mathbb{N}$,*

$$a + c = a' + c \text{ if and only if } a = a'$$

2. *For every $d, \in \mathbb{N} \setminus \{0\}$,*

$$a \cdot d = a' \cdot d \text{ if and only if } a = a'$$

3. *For every $b, e, f \in \mathbb{N}$ with $e \leq_{\mathbb{N}} f$,*

   (a) *$a \leq_{\mathbb{N}} b$ implies $a + e \leq_{\mathbb{N}} b + f$, and*

   (b) *$a + f \leq_{\mathbb{N}} b + e$ implies $a \leq_N b$.*

   (c) *If $0 < e \leq f$, then $a \leq_{\mathbb{N}} b$ implies $a \cdot e \leq_{\mathbb{N}} b \cdot f$, and*

   (d) *$a \cdot f \leq_{\mathbb{N}} b \cdot e$ implies $a \leq_N b$.*

The proofs are left as an exercise.

2.6.2 The integer numbers

The ground idea is to define the integer numbers as *differences* of natural numbers. From high school we know that, if a and b are natural number, then $a - b$ is an integer number and, conversely, if z is an integer, then either $z \in \mathbb{N}$, and $z = z - 0$ is a difference of natural numbers, or $-z \in \mathbb{N}$ and, again $z = 0 - (-z)$ is a difference of natural numbers[18]. Clearly different pairs of natural numbers can have the same difference, e.g. $5 - 4 = 8 - 7$, so we must *identify* all pairs of natural numbers that have the same difference. We'll do this using equivalences.

Let $+$ be the addition in $\mathbb{N}$, consider the Cartesian product $\mathbb{N} \times \mathbb{N}$ and define a relation $\cong$ on $\mathbb{N} \times \mathbb{N}$ in the following way:

$$(a, b) \cong (c, d) \text{ if and only if } a + d = b + c \tag{2.22}$$

{eqz}

[17] It is called "Cancellation" because the Lemma enables us to "cancel" c (resp. d) in the equations (resp. diesequations).

[18] We'll define the rational numbers in a similar way as quotients of integer numbers, the latter construction is probably well known from high school, so we advice the reader to keep that construction in mind in the sequel.

Lemma 2.6.14 *The relation $\cong$ is an equivalence relation on $\mathbb{N} \times \mathbb{N}$*

PROOF. We have to prove that $\cong$ is reflexive, symmetric, and transitive.

REFLEXIVITY $(a, b) \cong (a, b)$, since $a + b = a + b$.

SYMMETRY Assume $(a, b) \cong (c, d)$, then, by Equation (2.22), $a+d = b+c$, so also $b + c = a + d$, whence $(c, d) \cong (a, b)$.

TRANSITIVITY Assume $(a, b) \cong (c, d)$ and Assume $(c, d) \cong (e, f)$. Then $a + d = b + c$ and $c + f = d + e$, whence

$$a + c + f = a + d + e = b + c + e$$

which implies, by Lemma 2.6.13

$$a + f = b + e$$

that is $(a, b) \cong (e, f)$.

■

Define $\mathbb{Z}$ to be the factor set $(\mathbb{N} \times \mathbb{N})/_{\cong}$. So the elements of $\mathbb{Z}$ are the equivalence classes $[(a, b)]_{\cong}$ of pairs of elements of $\mathbb{N}$.

We now want to define the addition and the multiplication on $\mathbb{Z}$. We first define on $\mathbb{N} \times \mathbb{N}$ an addition $\oplus$ and a multiplication $\otimes$. Namely, define

{sumz}
$$(a, b) \oplus (c, d) := (a + c,\ b + d); \tag{2.23}$$

now let $\cdot$ be the multiplication in $\mathbb{N}$ and define

{multz}
$$(a, b) \otimes (c, d) := (a \cdot c + b \cdot d,\ a \cdot b + b \cdot c) \tag{2.24}$$

{properties}
Lemma 2.6.15 *Let (a, b), (c, d) and (e, f) be)elements of $\mathbb{N} \times \mathbb{N}$. Then*

1. (ASSOCIATIVITY)

   1.a $((a, b) \oplus (c, d)) \oplus (e, f) = (a, b) \oplus ((c, d) \oplus (e, f))$,

   1.b $((a, b) \otimes (c, d)) \otimes (e, f) = (a, b) \otimes ((c, d) \otimes (e, f))$;

2. (COMMUTATIVITY)

   2.a $(a, b) \oplus (c, d) = (c, d) \oplus (a, b)$,

   2.b $(a, b) \otimes (c, d) = (c, d) \otimes (a, b)$;

3. (DISTRIBUTIVITY) $((a, b) \oplus (c, d)) \otimes (e, f) = ((a, b) \otimes (e, f)) \oplus ((c, d) \otimes (e, f))$.

PROOF. These are all easy computations, we'll work out only *1.b* and leave the others as an exercise.

$$\begin{aligned}
& ((a,b)\otimes(c,d))\otimes(e,f)\\
= \;& (ac+bd,\ ad+bc)\otimes(e,f)\\
= \;& ((ac+bd)e+(ad+bc)f,\ (ac+bd)f+(ad+bc)e)\\
= \;& (ace+bde+adf+bcf,\ acf+bdf+ade+bce)\\
= \;& (ace+adf+bde+bcf,\ acf+ade+bdf+bce)\\
= \;& (a(ce+df)+b(de+cf),\ a(cf+de)+b(df+ce))\\
= \;& (a,b)\otimes((ce+df),\ (de+cf))\\
= \;& (a,b)\otimes((c,d)\otimes(e,f)).
\end{aligned}$$

■

The following Lemma gives a crucial property of these operations.

{eqcom}

Lemma 2.6.16 *The equivalence relation $\cong$ is compatible with $\oplus$ and $\otimes$, that is, for every $(a,b),(c,d),(a',b'),(c',d')\in\mathbb{N}\times\mathbb{N}$ such that*

$$(a,b)\cong(a',b')\ \textit{and}\ (c,d)\cong(c',d'),$$

we have

1. $((a,b)\oplus(c,d))\cong((a',b')\oplus(c',d'))$ *and*
2. $((a,b)\otimes(c,d))\cong((a',b')\otimes(c',d'))$.

PROOF. The first equivalence is easy: we have to prove that

$$(a+c)+(b'+d')=(b+d)+(a'+c') \tag{2.25}$$ {x1}

Since $(a,b)\cong(a',b')$ and $(c,d)\cong(c',d')$ we have

$$a+b'=b+a' \text{ and } c+d'=d+c', \tag{2.26}$$ {xa}

whence

$$\begin{aligned}
(a+c)+(b'+d') &= (a+b')+(c+d')\\
&= (b+a')+(d+c')=(b+d)+(a'+c').
\end{aligned}$$

The second equivalence is longer (though still elementary): begin with the equality

$$(a+b')d+b'c'=ad+b'd+b'c'=ad+b'(d+c'). \tag{2.27}$$ {t1}

By Equations (2.26) and (2.27), we get

$$(a'+b)d+b'c'=(a+b')d+b'c'=ad+b'd+b'c'=ad+b'(d'+c), \tag{2.28}$$ {t2}

whence

$$a'd + bd + b'c' = (a' + b)d + b'c' = ad + b'(d' + c) = ad + b'd' + b'c \quad (2.29) \quad \{t3\}$$

Now add $a'c'$ and bc to both members and obtain

$$a'c' + bc + a'd + bd + b'c' = a'c' + bc + ad + b'd' + b'c \quad (2.30) \quad \{t4\}$$

Collect a' in the first member and get

$$a'(c' + d) + bc + bd + b'c' = a'c' + bc + ad + b'd' + b'c \quad (2.31) \quad \{t5\}$$

By Equation (2.26) we get

$$a'(c' + d) = a'(c + d') \quad (2.32) \quad \{t6\}$$

Thus Equations (2.31) and (2.32) give

$$a'(c + d') + bc + bd + b'c' = a'c' + bc + ad + b'd' + b'c \quad (2.33) \quad \{t7\}$$

or, equivalently,

$$a'c + a'd' + bc + bd + b'c' = a'c' + bc + ad + b'd' + b'c \quad (2.34) \quad \{t8\}$$

Now collect c in the first member and get

$$(a' + b)c + a'd' + bd + b'c' = a'c' + bc + ad + b'd' + b'c \quad (2.35) \quad \{t9\}$$

Use again Equation (2.26) and get

$$(a + b')c + a'd' + bd + b'c' = a'c' + bc + ad + b'd' + b'c \quad (2.36) \quad \{t10\}$$

that is

$$ac + b'c + a'd' + bd + b'c' = a'c' + bc + ad + b'd' + b'c \quad (2.37) \quad \{t11\}$$

By Lemma 2.6.13 we can cancel $b'c$ from both members and get

$$ac + bd + a'd' + b'c' = ad + bc + a'c' + b'd',$$

that is

$$((a, b) \otimes (c, d)) \cong ((a', b') \otimes (c', d')),$$

as desired. ■

This result is important because it can be restated as follows:

Lemma 2.6.17 *For every* $(a, b), (c, d), (a', b'), (c', d') \in \mathbb{N} \times \mathbb{N}$ *such that*

$$(a, b) \cong (a', b') \text{ and } (c, d) \cong (c', d'),$$

we have

$$[(a, b) \oplus (c, d)]_{\cong} = [(a', b') \oplus (c', d') \text{ and } [(a, b) \otimes (c, d)]_{\cong} = [(a', b') \otimes (c', d')$$

This fact allows us to define an addition $+_{\mathbb{Z}}$ and a multiplication $\cdot_Z$ on the factor set $\mathbb{Z}$ as follows:

$$[(a,b)]_{\cong} +_{\mathbb{Z}} [(c,d)]_{\cong} := [(a,b) \oplus (c,d)]_{\cong} = [(a+c, b+d)]_{\cong}$$

and

$$[(a,b)]_{\cong} \cdot_{\mathbb{Z}} [(c,d)]_{\cong} := [(a,b) \otimes (c,d)]_{\cong} = [(ab+cd, ad+cb)]_{\cong}$$

and, by the above Lemma, these definitions do not depend on the chosen elements (a,b) and (c,d) of the classes $[(a,b)]_{\cong}$ and $[(c,d)]_{\cong}$.

To avoid heavy notation we shall use the symbols $+$ and $\cdot$ instead of $+_{\mathbb{Z}}$ and $\cdot_{\mathbb{Z}}$ and the symbol $[(a,b)]$ instead of $[(a,b)]_{\cong}$. Here are some basic properties of these operations:

Lemma 2.6.18 *For every $x, y, z \in \mathbb{Z}$,*

1. (ASSOCIATIVITY)

   1.a $x + (y + z) = (x + y) + z$

   1.b $x \cdot (y \cdot z) = (x \cdot y) \cdot z$

2. (COMMUTATIVITY)

   1.a $x + y = y + x$

   1.b $x \cdot y = y \cdot x$

3. (DISTRIBUTIVITY) $x + y) \cdot z = (x \cdot z) + (y \cdot z)$.

PROOF. All these properties follow immediately from the definition of $+$ and $\cdot$ in $\mathbb{Z}$ and the analogous properties of $\oplus$ and $\otimes$: e.g. if $x = [(a,b)]$, $y = [(c,d)]$, and $z = [(e,f)]$, with $a, b, c, d, e, f \in \mathbb{N}$, then

$$\begin{aligned} x + (y + z) &= [(a,b)] + ([(c,d)] + [(e,f)]) \\ &= [(a,b)] + [(c,d) \oplus (e,f)] \\ &= [(a,b) \oplus ((c,d) \oplus (e,f))] \\ &= [((a,b) \oplus (c,d)) \oplus (e,f)] \\ &= [(a,b) \oplus (c,d)] + [(e,f)] \\ &= ([(a,b)] + [(c,d)]) + [(e,f)] \\ &= (x + y) + z \end{aligned}$$

■

Let $0_{\mathbb{Z}} := [(0,0)]$ and $1_{\mathbb{Z}} := (1,0)$. Then

Lemma 2.6.19 (NEUTRAL ELEMENT AND IDENTITY) *For every $z \in \mathbb{Z}$*

1. (NEUTRAL ELEMENT) $z + 0 = z$;

2. (IDENTITY) $z \cdot 1 = z$*;*

3. $z \cdot 0 = 0$.

PROOF. Let $z := [(a, b)]$ with $a, b \in \mathbb{N}$. Then

$$z + 0_{\mathbb{Z}} = [(a, b)] + [(0, 0)] = [(a, b) \oplus (0, 0)] = [(a + 0,\ b + 0)] = [(a, b)]$$

Similarly

$$z \cdot 1_{\mathbb{Z}} = [(a, b)] \cdot [(1, 0)] = [(a, b) \otimes (1, 0)] = [(a \cdot 1 + b \cdot 0,\ a \cdot 0 + b \cdot 1)] = [(a, b)],$$

and, finally,

$$z \cdot 0_{\mathbb{Z}} = [(a, b)] \cdot [(0, 0)] = [(a, b) \otimes (0, 0)] = [(a \cdot 0 + b \cdot 0,\ a \cdot 0 + b \cdot 0)] = [(0, 0)].$$

■

Lemma 2.6.20 (OPPOSITE) *For every* $[(a, b)] \in \mathbb{Z}$*, with* $a, b \in \mathbb{N}$*,* $[(b, a)]$ *is the unique element of* $\mathbb{Z}$ *such that* $[(a, b)] + [(b, a)] = [(0, 0)]$.

PROOF. First note that

$$[(a, b)] + [(b, a)] = [(a + b,\ a + b)]$$

and, by the definition of the equivalence $\cong$,

$$(a + b, a + b) \cong (0, 0),$$

whence

$$[(a + b,\ a + b)] = [(0, 0)].$$

Conversely, suppose that $[(c, d)] \in \mathbb{Z}$ is such that

$$[(a, b)] + [(c, d)] = [(0, 0)].$$

Then

$$[(a + c,\ b + d)] = [(a, b)] + [(c, d)] = [(0, 0)],$$

whence

$$(a + c,\ b + d) \cong (0, 0).$$

But this means that

$$a + c = (a + c) + 0 = (b + d) + 0 = b + d,$$

that is

$$(b, a) \cong (c, d),$$

whence $[(c, d)] = [(b, a)]$. ■

If $z \in \mathbb{Z}$ with $z = [(a, b)]$, the element $[(b, a)]$ is usually denoted by $-z$ and is called the *opposite* of z.

We now want to define on $\mathbb{Z}$ the usual ordering $\leq_Z$ using the total ordering $\leq_{\mathbb{N}}$ on $\mathbb{N}$. In order to define $\leq_{\mathbb{Z}}$, we first need the following result

{apricot}

Lemma 2.6.21 *Let $(a,b),(a',b'),(c,d),(c',d')$ be pairs in $\mathbb{N}\times\mathbb{N}$ with*

$$(a,b)\cong(a',b') \text{ and } (c,d)\cong(c',d').$$

Then

$$a+c\leq_{\mathbb{N}} b+d \text{ if and only if } a'+c'\leq_{\mathbb{N}} b'+d'.$$

PROOF. Since $(a,b)\cong(a',b')$ and $(c,d)\cong(c',d')$, we have

$$a+b'=b+a' \text{ and } c+d'=d+c' \tag{2.38}$$ {apri}

Assume

$$a+c\leq_{\mathbb{N}} b+d.$$

Adding to both members $b'+d'$, we get

$$a+b'+c+d'\leq_{\mathbb{N}} b+b'+d+d'.$$

By Equation (2.38), it follows

$$a'+b+c'+d\leq_{\mathbb{N}} b+b'+d+d',$$

and, by Lemma 2.6.13, cancelling $b+d$ to both members, we obtain:

$$a'+c'\leq_{\mathbb{N}} b'+d'.$$

The converse is proven in the same way swapping the roles of a,b,c,d with those of a',b',c',d' ∎

As above Lemma 2.6.21 allows us to define a relation $\leq_Z$ in $\mathbb{Z}$ in the following way: For $[(a,b)]$ and $[(c,d)]$ in $\mathbb{Z}$ with $a,b,c,d\in\mathbb{N}$, set

$$[(a,b)]\leq_{\mathbb{Z}}[(c,d)] \text{ if and only if } a+d\leq_{\mathbb{N}} b+c.$$

Lemma 2.6.22 *The pair $(\mathbb{Z},\leq_{\mathbb{Z}})$ is a totally ordered set.*

PROOF. We first prove that $\leq_{\mathbb{Z}}$ is an order relation. It is obviously reflexive, since $a+b\leq_{\mathbb{N}} a+b$ implies $[(a,b)]\leq_{\mathbb{Z}}[(a,b)]$. Let $[(a,b)]$, $[(c,d)]$, and $[(e,f)]$ be integers. To prove that $\leq_{\mathbb{Z}}$ is antisymmetric, assume that

$$[(a,b)]\leq_{\mathbb{Z}}[(c,d)] \text{ and } [(c,d)]\leq_{\mathbb{Z}}[(a,b)],$$

then the definition of $\leq_{\mathbb{Z}}$ implies that

$$a+d\leq_{\mathbb{N}} c+b \text{ and } c+b\leq_{\mathbb{N}} a+d.$$

Since $\leq_{\mathbb{N}}$ is an order relation, this implies that

$$a+d=c+b,$$

whence $(a,b) \cong (c,d)$ or, equivalently

$$[(a,b)] = [(c,d)].$$

Now assume that $[(e,f)] \in \mathbb{Z}$ and

$$[(a,b)] \leq [(c,d)] \text{ and } [(c,d)] \leq [(e,f)].$$

Then

$$\begin{aligned} a + d &\leq_{\mathbb{N}} b + c \\ c + f &\leq_{\mathbb{N}} e + d \end{aligned} \tag{2.39}$$ {condor}

Whence, adding $c+f$ and $e+d$ in the first and second member of the Disequation (2.39), by Lemma 2.6.13, we get

$$a + d + c + f \leq_{\mathbb{N}} b + c + e + d.$$

Again, by Lemma 2.6.13, we can cancel $c+d$ in both members and get

$$a + f \leq_{\mathbb{N}} b + e,$$

whence $[(a,b)] \leq [(e,f)]$, proving transitivity. We finally prove that $\leq_{\mathbb{Z}}$ is a total ordering. But this is immediate, for if $[(a,b)]$ and $[(c,d)]$ are integers then

either $a + d \leq_{\mathbb{N}} c + b$, whence $[(a,b)] \leq_{\mathbb{Z}} [(c,d)]$,

or $c + b \leq_{\mathbb{N}} a + d$, whence $[(c,d)] \leq_{\mathbb{Z}} [(a,b)]$.

■

{cancelz}

Lemma 2.6.23 *All statements of Lemma 2.6.13 hold when $\mathbb{N}$ is substituted by $\mathbb{Z}$.*

Now consider the map

$$\begin{array}{ccc} \mathbb{N} & \longrightarrow & \mathbb{Z} \\ a & \mapsto & (a,0) \end{array}$$

This map is clearly injective, for if $a, b \in \mathbb{N}$ are such that $[(a,0)] = [(b,0)]$, then $(a,0) \cong (b,0)$ whence

$$a = a + 0 = b + 0 = b.$$

Also note that, for every $a, b \in \mathbb{N}$

1. $[(a,0)] +_{\mathbb{Z}} [(b,0)] = [(a+b,0)]$;
2. $[(a,0) \cdot_{\mathbb{Z}} [(b,0)] = [(a \cdot b,0)]$;
3. $[(a,0) \leq_{\mathbb{Z}} [(b,0)]$ if and only if $a \leq_{\mathbb{N}} b$;

4. for every $a, b \in \mathbb{N}$, $[(0,a)] \leq_{\mathbb{Z}} [(b,0)]$.

In other words, the set

$$\mathbb{Z}^{\geq 0} := \{[(a,0)] \mid a \in \mathbb{N}\}$$

behaves exactly in the same way as $\mathbb{N}$ does, so we may identify $\mathbb{N}$ with $\mathbb{Z}^{\geq 0}$ and thus consider $\mathbb{N}$ as a subset of $\mathbb{Z}$ (and this is what usually do). The set $\mathbb{Z}^{\geq 0}$ is also called the set of *non negative* integers. and the set

$$\mathbb{Z}^{>0} := \mathbb{Z}^{\geq 0} \setminus \{0\}$$

is called the set of *positive* integers.

Similarly, let

$$\mathbb{Z}^{\leq 0} := \{[(0,a)] \mid a \in \mathbb{N}\}$$

and

$$\mathbb{Z}^{<0} := \{[(0,a)] \mid a \in \mathbb{N}\} \setminus \{0\}.$$

Then $\mathbb{Z}^{\leq 0}$ is called the set of *non positive* integers and $\mathbb{Z}^{<0}$ is called the set of *negative* integers.

{obvi}

Lemma 2.6.24 *The following equalities hold:*

1. $\mathbb{Z}^{\leq 0} = \{-z \mid z \in \mathbb{Z}^{\geq 0}\}$;
2. $\mathbb{Z} = \mathbb{Z}^{\geq 0} \cup \mathbb{Z}^{<0} = \mathbb{Z}^{>0} \cup \mathbb{Z}^{<\leq 0} = \mathbb{Z}^{\geq 0} \cup \mathbb{Z}^{\leq 0}$;
3. $\emptyset = \mathbb{Z}^{\geq 0} \cap \mathbb{Z}^{<0} = \mathbb{Z}^{>0} \cap \mathbb{Z}^{<\leq 0} = \mathbb{Z}^{<0} \cap \mathbb{Z}^{>0}$;
4. $\{0\} = \mathbb{Z}^{\geq 0} \cap \mathbb{Z}^{\leq 0}$.

PROOF. Exercise ∎

Let z be an integer, define

$$|z| := \begin{cases} z \text{ if } z \text{ is non negative} \\ -z \text{ otherwise} \end{cases}$$

$|z|$ is called the *modulus* of z. So, for example $|3| = 3$ and $|-4| = -(-4) = 4$.

Lemma 2.6.25 THE EUCLIDEAN DIVISION *Let w and z be integers, with $z \neq 0$. Then there exist a unique pair (q,r) of integers such that*

1. $w = q \cdot z + r$
2. $0 \leq r < |z|$

PROOF. Consider the set

$$\mathcal{R} := \{w + t \cdot z \mid t \in \mathbb{Z}\} \cap \mathbb{Z}^{\geq 0}.$$

We prove that $\mathcal{R}$ is not empty. Assume first $0 \leq w$. Then

$$w = w + 0 \cdot z \in \mathcal{R}.$$

Assume $w < 0$. If $0 \leq z$, then, since $z \neq 0$, $1 \leq z$ by the well ordering of $\mathbb{N}$ (hence of $\mathbb{Z}^{\geq 0}$. So $-w \leq (-w) \cdot z$, whence, by Lemma 2.6.23

$$0 = w + (-w) \leq w + -w \cdot z \in \mathcal{R}.$$

Finally, if $z < 0$, then, $1 \leq -z$ and, again,

$$0 = w + (-w) \leq w + (-w)\cdot)(-z) = w + w \cdot z \in \mathcal{R}.$$

By the Minimum principle, $\mathcal{R}$ has a unique minimum element $r := w + sz$ with $s \in \mathbb{Z}$. We prove that $r < |z|$. Assume, by contradiction, that $z \leq r$. If $0 < z$, then $|z| = z$ and

$$0 \leq r + (-z) = (w + sz) + (-z) = w + (s + (-1))z \in \mathcal{R},$$

but since $0 < z$, $r + (-z) < r$ which is a contradiction against the minimality of r. So assume $z < 0$, then $|z| = -z$ and

$$0 \leq r + z = (w + s \cdot z) + z = w + (s + 1) \cdot z \in \mathcal{R},$$

and again we get a contradiction against the minimality of r, since $r + z \leq r$. Now let $q := -s$ then

$$w = ((-s) \cdot z) + w + (s \cdot z) = q \cdot z + r.$$

Finally assume q' and r' are integers such that

1. $w = q' \cdot z + r'$ and
2. $0 \leq r' \leq |z|$

Assume, by contradiction that $r' \neq r$. By the minimality of r, we have $r < r'$, whence

$$0 < r' + (-r) < r'$$

but

$$r' + (-r) = (q + (-q')) \cdot z$$

the uniqueness of r and the cancellation law imply the uniqueness of q. ∎

If w and z are integers, we say that z is a *multiple* of w if there is an integer k such that $z = kw$. In this case we also say that w is a *divisor* of z(or w *divides* z). A *prime integer* is an integer p such that the set of of its divisors is $\{1, -1, z, -z\}$.

2.7 Exercises

{invinj}

Exercise 2.7.1 *Prove that a map*

$$f\colon A \longrightarrow B$$

is injective if and only if there is a map

$$h\colon B \longrightarrow B$$

such that $h \circ f = id_A$.

{invinj}

Exercise 2.7.2 *Prove that a map*

$$f\colon A \longrightarrow B$$

is surjective if and only if there is a map

$$h\colon B \longrightarrow B$$

such that $f \circ h = id_B$.

{comp1}

Exercise 2.7.3 *Let ρ be a correspondance between the sets A and B and let δ be a correspondence between the sets B and C. Define the composition of δ with ρ so that, in the particular case when ρ and δ are maps, this is precisely the composition of maps.*

{comp0}

Exercise 2.7.4 *Prove Lemma 2.6.13*

{comploj}

Exercise 2.7.5 *Prove Lemma 2.6.23*

{comp01}

Exercise 2.7.6 *Prove Lemma 2.6.24*

{comp2}

Exercise 2.7.7 *Let $w, z \in \mathbb{Z}$ prove that*

1. $(-1) \cdot z = -z$;
2. *if* $w, z \in Z^{geq0}$, *then* $w \cdot z \in Z^{geq0}$;
3. *if* $w, z \in Z^{leq0}$, *then* $w \cdot z \in Z^{geq0}$;
4. *if* $w \in Z^{geq0}$ *and* $z \in Z^{leq0}$, *then* $w \cdot z \in Z^{leq0}$;

{comp3}

Exercise 2.7.8

{comp4}

Exercise 2.7.9

{comp25}

Exercise 2.7.10

Chapter 3

Symmetry

3.1 Introduction

"Simmetry is often a symptom of decadence." With these words my father once replied my objection that the façades of the early Venetian palaces were not symmetrical.

Indeed, there seems to be a general agreement among aesteticians and psychologists of perception that symmetry is on one side associated to beauty, harmony, balance, steadyness, order but also, on the other side, to lack of fantasy, rigidity, monotony, stillness, and boredom. No wonder that the dynamical and lively merchant Venetian community never paid much attention to have symmetrical palaces: the façades of many buildings on the Grand Canal, such from the Ca' d'Oro, to the waterfront of the Doge's are asymmetrical. Even the hull of a gondola developed eventually into an asymmetrical shape.

In the merchant houses, such as the Ca' d'Oro, the asymmetrical design was motivated by functional reasons: the water entrance and the *porteghi* occupying two thirds on the left of the main façade and the offices (*mezzanino*) and the rooms on the remaining third. This typology is recalled in the waterfront façade of the Doge's Palace, where the two big windows on the righ are slightly lower than the others.

Similarly, the asymmetrical design of a gondola is needed to counterbalance the momentum induced by the asymmetrical Venetian way of rowing (the only way to get along the narrow canals of Venice).

Usually the symmetry of an object is described as the set of transformations that leave invariant that object. For example a square is invariant for the four rotations around its center (of 0, 90, 180 and 270 degrees), the two reflections

Figure 3.1: The Ca' d'Oro

around its diagonals and the two reflections around its middle axes, making eight transformations in total. You may enjoy yourself trying to prove that there are no more.

In this sense, the awareness of symmetry can be of extreme importance (e.g. the awareness of the yearly temporal symmetry of the seasons was vital for our hunter-gatherer ancestors) and the chase for symmetry is a major ingredient of scientifical research: the goal of performing a scientifical experiment several times under different conditions is to determine under which spatial, temporal, environmental etc. transformations the outputs remain invariant, that is determine the symmetry of that experiment.

Maybe this could explain our attraction towards symmetry: survival issues made us naturally inclined to looking for symmetry, so that now we feel somehow rewarded when we detect it. And maybe this could explain Escher's success amongst mathematicians: they eventually detect something they know in Escher's drawings and thus feel rewarded by that.

Now look at the following picture

They reproduce the same square (upper left) under different points of view and, as such, each one of the shapes has to be invariant under eight transformations corresponding to the eight transformations of the square (we will see in

Figure 3.2: The waterfront façade of the Doge's Palace

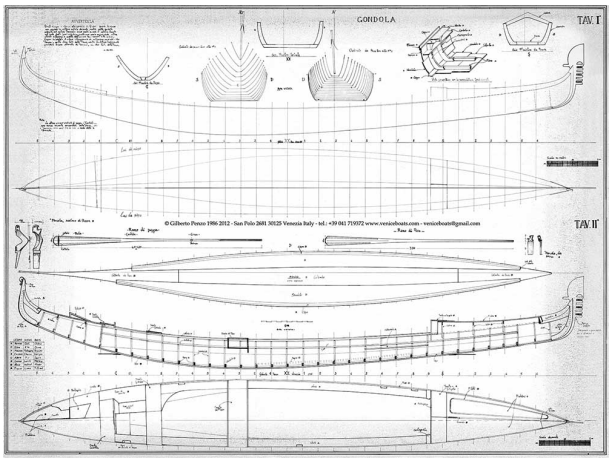


Figure 3.3: The plans of a Gondola *(courtesy of Gilberto Penzo)*

the next chapters how these transformations can be computed). Nevertheless, to the unacccustomed eye, only the picture un the upper left appears "fully" symmetrical, in the sense that all the eight tranformations leave the square invariant. This shows that symmetry is relative, in the sense that we decide which transformations, leaving an object invariant, are important for us.[1].

[1]A word has to be said about the title of the series: actually the pictures of the square have nothing to do with non Euclidean geometries in the strict sense. Nevertheless quoting non Euclidean geometries does impress mathematically ignorant observers who do not like to admit their ignorance. This is a perverse use of mathematics which I believe should be avoided (in a certai way it is analogous to the political use of symmetry in architecture or the political use of the archistars I'll discuss below). You have to be suspicious when an artist talks about mathematical concepts such as non Euclidean geometries, exotic topologies, fractals, hyperbolic paraboloids, etc.: in most cases mathematics is a cover to hide the mediocrity of the artwork.

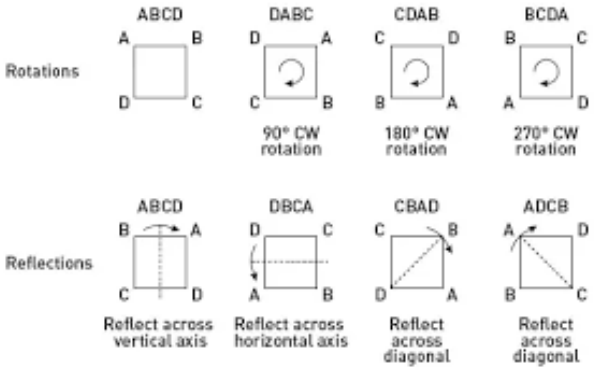


Figure 3.4: The eight symmetries of a square

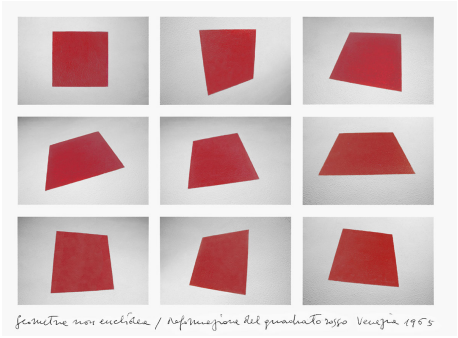


Figure 3.5: Mario Cresci, Alterazione del quadrato, dalla serie Geometria non euclidea, Venezia 1965.

3.1.1 Symmetry in military design

It might be interesting to analyse the use of symmetry in military design: here functionality is the absolute must, all the rest is useless, if not harmful. It is remarkable that eventually these designs turn out to be also aesthetically valid: rephrasing Gardella's assertion, my father used to tell me that functionality is a way to produce beauty, and beauty a confirmation that the functionality issues have been properly solved.

Ships

Generally the hull of a ship is symmetrical along its vertical and longitudinal plan: that is, if we cut a ship from bow to stern with a vertical mirror, the mirrored image coincides with the image of the ship behind the mirror (with the exceptions of the gondola and the proa). If we do not consider the sail nor its appendages, the hull of a submarine has even more symmetries for each

cross section is round. In both cases the symmetrical design is clearly due to hydrostatic (water pressure in a submarine) or/and hydrodynamic purposes.

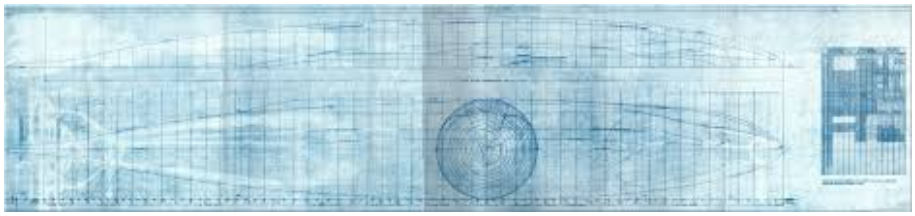


Figure 3.6: USS Albacore blueprints

Figure 3.7: The vipera

There are also examples of bow stern symmetries. A remarkable one is the *vipera* a Venetian boat used by the police during the Austrian occupation to chase poachers along the canals of the lagoon: due its bow-stern almost symmetric design its direction could be quickly inverted just by letting the rowers turn around while leaving the boat still.

The deck of a battleship is symmetrical along its longitudinal axis. This is justified by the fact that the guns are attached to the ship and at gun's relatively short range the enemy can happen to be on any side. An asymmetrical design could produce a weaker and a stronger side, making the ship vulnerable on the weaker side. The bow stern asymmetry, with two main batteries on the front and only one on the back, might be explained by the fact that usually a battleship sails *towards* the enemy.

On the opposite side, the major offending and defending devices of an aircraft carrier (the airplanes) act completely independently of the ship (apart from

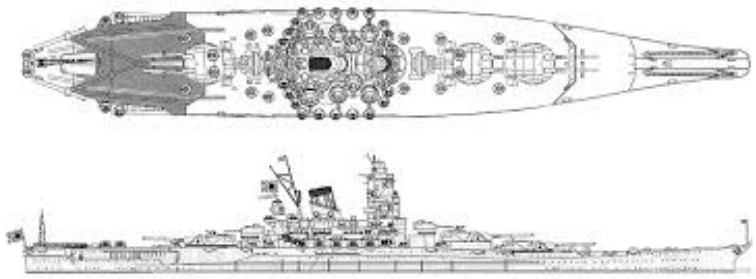


Figure 3.8: The Yamato battleship

take off, landing and storage) so that there is no need for symmetry, while the asymmetrical design (with e.g. the island on the right and the angled flight deck) make an aircraft carrier more efficient operationally.

Emblematically, the symmetric Yamato battleship sank after an air attack by bombers that took off from USS aircraft carriers' (Enterprise, Yorktown, and Intrepid) asymmetrical decks.

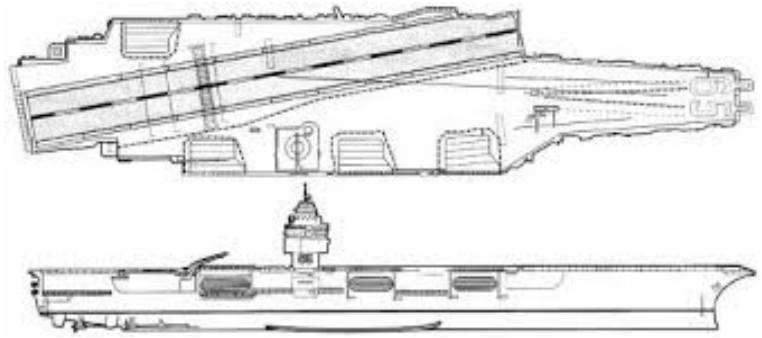


Figure 3.9: USS Enterprise

Fortresses

Here we have a similar situation, we just give two examples in which the functionality of an asymmetrical (Candia) and symmetrical (Palmanova) designs are evidently given by the place: The wall of Candia (Heraklion) in Crete: the general structure is fully asymmetrical due to the position with the sea on the northern side and the mainland on the southern. On the opposite side, the fortress of Palmanova, located in the midle of a flatland has a perfectly 9-gonal symmetry.

Figure 3.10: The Fortress of Candia

The Pentagon

We should also mention that symmetry may turn out to be disatrous. In his beautiful book *Symmetry* published first in 1951, the mathematician Hermann Weyl wrote: "Now of course we have the Pentagon building in Washington. By its size and distinctive shape, it provides an attractive landmark for bombers" [12, p.66]. We sadly experienced that Weyl's words have been tragically confirmed on September 11th, 2001.

3.1.2 Political use of architectonic symmetry

Generally, the occasions where symmetry in architecture or, more generally, in design, is justified are much less than we are accostumed to expect, for the conditions (light, exposure, use etc...) on one side of a building are very seldom the same as those on the other sides. Nevertheless symmetry in architecture can still have important uses:

Impress people

The aforementioned sense of balance, steadyness, order given by symmetry has been often exploited by regimes (and not only dictatorial) to gain consent or to express mightyness and power. This has been widely used until the recent years (now the use of hiring archistars has proved to be an even better and, apparenlty, more politically correct way to get people's approval, so that the use of symmetry is now confined usually among dictatorships in lesser developed countries). There are hundreds of examples of such a use of symmetry: the Stalinist Gothic,

Figure 3.11: The Fortress Palmanova

the already mentioned Pentagon building, The White House, The Presidential Complex in Ankara, most royal palaces (exercise: find an asymmetrical one built after the XVII century), the pyramids, the greek temples, churches such as St. Peter's Basilica in the Vatican, Notre Dame in Paris, St. Stephen cathedral in Vienna, the magnificient Hagia Sophia, or the mathematically perfect Blue Mosque in Istanbul.

Modularity

By this we mean the possibility of interchanging parts of a building (e.g. rooms) without changing its structure. This kind of symmetry can facilitate the designing process (it has been widely used by Andrea Palladio for the design of his Villas in the Venetian countryside) and can be useful also for political purposes. Nowadays it is common in administrative buildings, schools etc. and it is a must when designing a university department: to prevent envy and discord, all professor's offices need to have the same shape, view and exposition (this was not the case when we moved to our new department and I experienced one of my colleagues measuring all offices to be sure his was the largest). Maybe this was also a motivation for the highly symmetrical structure of the Lomonosov University in Moscow about which a Russian mathematician once hyperbolically told me that any finite simple group has a faithful representation as a group of symmetry of that building (this is obviously not true since there are infinitely many finite simple groups).

Figure 3.12: The Moscow State University

3.1.3 Symmetry and macaroni

3.2 Groups

Groups are the mathematical tool to measure symmetry. It is remarkable that this concept was introduced and proved to be effective only at the beginning of the 19th century by a young French mathematician: Évariste Galois (1811-1831)...yes he died, in a duel, at the age of 21, but his ideas had an enormous influence on mathematics. In the last 200 years the theory of groups developed immensely and groups are now ubiquitous in many disciplines. We'll discuss some basics of this theory in the sequel.

3.2.1 Operations

At the basis of the concept of group lies the concept of operation. So let's think about what we are used to call an operation. For example consider the addition between integer numbers: when we add two integers, say e.g. 5 and 7, we obtain a new integer, 12 in this case. What the addition does is simply to associate to a pair of numbers a number. In other words, the addition is a map

$$\begin{array}{rccc} +: & \mathbb{Z}\times\mathbb{Z} & \to & \mathbb{Z} \\ & (a,b) & \mapsto & a+b \end{array}$$

from the cartesian product $\mathbb{Z}\times\mathbb{Z}$ to $\mathbb{Z}$. So, e.g., 12 is the image of the pair $(5,7)$ via the addition map. Similarly the multiplication or difference between

two integer numbers are maps

$$\begin{array}{rccc} \cdot\colon & \mathbb{Z}\times\mathbb{Z} & \to & \mathbb{Z} \\ & (a,b) & \mapsto & a\cdot b \end{array} \quad \text{resp.} \quad \begin{array}{rccc} -\colon & \mathbb{Z}\times\mathbb{Z} & \to & \mathbb{Z} \\ & (a,b) & \mapsto & a-b \end{array}$$

from the cartesian product $\mathbb{Z}\times\mathbb{Z}$ to $\mathbb{Z}$. In this case $35(=5\cdot 7)$ is the image of the pair $(5,7)$ via the multiplication and $-2(=5-7)$ is the image of $(5,7)$ via the subtraction.

More generally, given a set G, an *operation* on G is a map

$$*\colon G\times G\to G$$

The symbols that are normally used to denote an operation are

$$\cdot, \circ. \times, +, *, -, :\ .$$

Further, for a pair a, b in $G\times G$, its image via the operation $*$ is denoted by

$$a*b \text{ instead of } *(a,b).$$

Examples of operations are the

- addition and the multiplication in $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$;
- the subtraction in $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ (but not in $\mathbb{N}$ for, e.g. the image of $5-7$ is not defined in $\mathbb{N}$);
- the division in $\mathbb{Q}\setminus\{0\}$, $\mathbb{R}\setminus\{0\}$, $\mathbb{C}\setminus\{0\}$ (we need to exclude 0, since division by 0 is not defined);
- the composition of maps in the set X^X of all maps of a set X into itself.
- the composition of permutations in the subset S_X of X^X consisting of all permutations of a set X.

The last example is most important and it is probably the one from which the concept of group had its origins and raison d'être. We shall discuss it more deeply in the sequel.

An operation $*$ on a set G is said to be *commutative* if for every a and b in G

{comm}
$$a*b=b*a \tag{3.1}$$

For example, addition and multiplication are commutative:

$$a+b=b+a \text{ and } a\cdot b=b\cdot a.$$

While subtraction and division are not:

$$5-7=-2\neq 2=7-5 \text{ and } 4\colon 2=2\neq\frac{1}{2}=2\colon 4.$$

Similarly if a set X contains at least 3 elements, the composition in X^X is not commutative: take three distinct elements x, y, and z in X and consider the permutation $\sigma_{x,y}$ that swaps x and x and fixes all other elements of X and the permutation $\sigma_{y,z}$ that swaps y and z and fixes all other elements of X. Now let's prove that

$$\sigma_{x,y} \circ \sigma_{y,z} \neq \sigma_{y,z} \circ \sigma_{x,y}.$$

Since two maps are equal if and only if they coincide element by element, we just need to find an element $t \in X$ such that

$$(\sigma_{x,y} \circ \sigma_{y,z})(t) \neq (\sigma_{y,z} \circ \sigma_{x,y})(t).$$

We show that the above inequality holds if we choose t to be x: since x is fixed by $\sigma_{y,z}$ and is sent to y by $\sigma_{x,y}$, we have

$$\sigma_{y,z}(x) = x \text{ and } \sigma_{x,y}(x) = y,$$

hence

$$(\sigma_{x,y} \circ \sigma_{y,z})(x) = \sigma_{x,y}(\sigma_{y,z}(x)) = \sigma_{x,y}(x) = y.$$

On the other hand, since

$$\sigma_{y,z}(y) = z,$$

we have

$$(\sigma_{y,z} \circ \sigma_{x,y})(x) = (\sigma_{y,z}(\sigma_{x,y}(x)) = \sigma_{y,z}(y) = z \neq y.$$

Magmas and semigroups

A nonempty set G endowed with an operation $*$ (or, more formally a pair $(G, *)$ where G is a nonempty set and $*$ is an operation on G) is called a *magma*.

A magma $(G, *)$ is called a *semigroup* if the operation $*$ is *associative*, that is, for every $a, b, c \in G$,

$$a * (b * c) = (a * b) * c.$$

Examples:

- The addition and the multiplication in $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are associative, so, for all $G \in \{\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}\}$ and $* \in \{+, \cdot\}$, $(G, *)$ is a semigroup.

- The composition $\circ$ in the set X^X is associative (see Proposition 2.4.3), so $(X^X, \circ)$ is a semigroup.

- Subtraction and division are not associative: e.g. $8 - (7 - 2) = 8 - 5 = 3$ but $(8 - 7) - 2 = 1 - 2 = -1$, and similarly for the division, so $(\mathbb{Z}, -)$ and $(\mathbb{Q} \setminus \{0\}, :)$ are magmas but not semigroups.

Monoids

A semigroup $(G, *)$ is called a *monoid* if there exist elements $e \in G$ such that, for every $a \in G$

$$e * a = a * e = a. \tag{3.2}$$ {monoid}

Note that, if such an elemet exists, it must be unique:

Lemma 3.2.1 *Let $(G, *)$ be a semigroup and assume e_1 and e_2 are elements of G such that, for every $a \in G$,*

$$e_1 * a = a * e_1 = a \tag{3.3}$$ {e1}

and

$$e_2 * a = a * e_2 = a. \tag{3.4}$$ {e2}

Then $e_1 = e_2$.

PROOF. By Equation (3.4), taking $a = e_1$, we have

$$e_1 = e_1 * e_2$$

By Equation (3.3), taking $a = e_2$, we get

$$e_1 * e_2 = e_2.$$

So $e_1 = e_1 * e_2 = e_2$. ∎

The unique element e in a monoid $(G, *)$ that satisfies Equation (3.2) for every a in G is usually denoted by 1 and it is called the *identity* of the monoid $(G, *)$. In case the operation is denoted by the symbol + (the so-called *additive* notation) we use the symbol 0 and it is called the *zero* or the *neutral element* of $(G, +)$.

Examples:

- The semigroups $(\mathbb{N}, +)$, $(\mathbb{Z}, +)$, etc. are monoids and the identity is the number 0.
- The semigroups $(\mathbb{N}, \cdot)$, $(\mathbb{Z}, \cdot)$ etc. are monoids and the identity is the number 1.
- The semigroup $(X^X, \cdot)$ is a monoid and the identity is the identity map id_X on X that is the permutation of X that fixes every element of X:

$$\begin{array}{rccc} id_x\colon & X & \to & X \\ & x & \mapsto & x \end{array}$$

- $(\mathbb{N} \setminus 0, +)$ is a semigroup, but not a monoid.

Actually there is an easy way to obtain a monoid from a semigroup: just add to a semigroup a new element, say e, and impose it behaves like the identity element.

Given a monoid $(G, *)$ with identity 1, an element $a \in G$ is said to be *invertible* if there exist elements $b \in G$ such that

{ab}
$$a * b = b * a = 1 \tag{3.5}$$

Lemma 3.2.2 *Let $(G, *)$ be a monoid with identity 1 and $a \in G$. Suppose b_1 and b_2 are elements of G such that*

$$ab_1 = b_1 a = 1 \tag{3.6}$$ {b1}

and

$$ab_2 = b_2 a = 1. \tag{3.7}$$ {b2}

Then $b_1 = b_2$.

PROOF. By Equation (4.7.4),

$$b_1 = b_1 * 1 = b_1 * (a * b_2)$$

and, by associativity and Equation (4.11),

$$b_1 * (a * b_2) = (b_1 * a) * b_2 = 1 * b_2 = b_2.$$

■

So, given a monoid $(G, *)$ with identity 1, if an element a is invertible, the unique element $b \in G$ such that Equation (3.5) is satisfied is called *inverse* (*opposite* in the additive notation) and is denoted by a^{-1} ($-a$ in the additive notation). Note that in a monoid $1^{-1} = 1$, since $1 = 1 * 1$ (similarly $-0 = 0$ in the additive notation). Examples:

- In $(N, +)$ only 0 is invertible while in $(\mathbb{Z}, +)$ every element is invertible.
- In $(N, \cdot)$ only 1 is invertible while in $(\mathbb{Z}, \cdot)$ the invertible elements are 1 and -1.
- In $(X^X, \circ)$ the elements that are invertible are precisely the permutations.
- In particular, in $(S_X, \circ)$ every element is invertible.
- In $(\mathbb{Q}, \cdot)$ and in $\mathbb{R}, \cdot$ every element different from the number 0 is invertible.

Note that Further, if a and b are invertible in G, then also $a * b$ is invertible and its inverse is

$$(a * b)^{-1} = b^{-1} a^{-1} \tag{3.8}$$ {Weyl}

(in his book [12], Herman Weyl gave a nice example of this fact: when you dress, you first put on your sockets and then your shoes. When you undress you first take off your shoes and then your sockets. That is, when you undress you do the reverse operations you did when dressing *in the reverse order*).

Groups

A *group* is a monoid in which all elements are invertible.

Groups are the mathematical tools for measuring symmetry in a similar way natural numbers measure finite quantities. Examples of groups are $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, $(\mathbb{C}, +)$, $(\mathbb{Q} \setminus 0, \cdot)$, $(\mathbb{R} \setminus 0, \cdot)$, $(\mathbb{C} \setminus 0, \cdot)$, and $(S_X, \circ)$. Similarly, the set of the eight symmetries of the square in the previous section is a group. Other examples of a group are the set of sequences of moves that lead from one configuration of the Rubik cube to another.

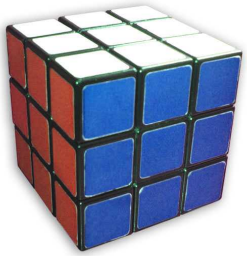


Figure 3.13: The Rubik's cube

Or, similarly, the set of sequences of moves in the 15-Puzzle that lead to one configuration with the empty tile in the lower right corner to another.

{15Puzzle}

Figure 3.14: The 15-Puzzle

If a group $(G, *)$ is *finite*, that is, G has a finite number of elements, say n. Let $g_1, g_2, g_3, \ldots, g_n$ be these elements and assume for convenience that $g_1 = 1_G$ is the identity of G. We can construct the *Pythagorean table* of G. This is a generalisation of the Pithagorean table we learned in the ground school: we first

list all the elements $g_1, g_2, g_3, \ldots, g_n$ of G, starting by the identity, in the first row and the first column of the table and, for every i and j in $\{1, \ldots, n\}$ the entry in the intersection of the row beginning with x_i and the column beginning with x_j is the product $x_i * x_j$:

1_G	g_2	$\cdots$	g_j	$\cdots$	g_n
g_2	$g_2 * g_2$	$\cdots$	$g_2 * g_j$	$\cdots$	$g_2 * g_n$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
g_i	$g_i * g_2$	$\cdots$	$g_i * g_j$	$\cdots$	$g_i * g_n$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
g_n	$g_n * g_2$	$\cdots$	$g_n * g_j$	$\cdots$	$g_n * g_n$

Note that choosing $g_1 = 1_G$, as we did, we have $g_i = g_i * g_1$ and $g_j = g_1 * g_j$, so that also the entries in the first row and the first column are consistent with the above rule.

As an exercise, let us compute the Pithagorean table of $(S_3, \circ)$, the group of all permutations of the set $\{1, 2, 3\}$ where the operation $\circ$ is the composition of permutations. One can see easily (and we shall do it later in full generality) that S_3 contains exactly six elements. If σ is a permutation of $\{1, 2, 3\}$ we can represent σ with a *matrix* with two rows and three colums in such a way that in the first row we list the elements 1, 2, and 3 and in the second row we list their images $\sigma(1)$, $\sigma(2)$, $\sigma(3)$:

$$\begin{pmatrix} 1 & 2 & 3 \\ \sigma(1) & \sigma(2) & \sigma(3) \end{pmatrix}$$

With this notation, the six elements of S_3 are

$$id_{\{1,2,3\}} := \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \ \rho_1 := \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \ \rho_2 := \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \quad (3.9) \quad \{\texttt{S3}\}$$

$$\sigma_1 := \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \ \sigma_2 := \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \ \sigma_3 := \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} \quad (3.10)$$

We first list the elements of S_3 beginning with the identity map $id_{\{1,2,3\}}$, which will be denoted by 1.

Now let us begin computing $\rho_1 \circ \rho_1$: By equation 3.9, we get

$$\rho_1(1) = 2, \ \ \rho_1(2) = 3, \text{ and } \rho_1(3) = 1,$$

hence

$$\rho_1 \circ \rho_1(1) = \rho_1(2) = 3, \ \ \rho_1 \circ \rho_1(2) = \rho_1(3) = 1, \text{ and } \rho_1 \circ \rho_1(3) = \rho_1(1) = 2.$$

So

$$\rho_1 \circ \rho_1 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} = \rho_2.$$

Thus we put ρ_2 in the entry in the intersection of the second line with the second column (i.e. those beginning with ρ_1.

Next we go on computing $\rho_1 \circ \rho_2$. This time we have, by equation 3.9:

$$\rho_2(1) = 3, \ \ \rho_2(2) = 1, \text{ and } \rho_2(3) = 2,$$

$$\rho_1 \circ \rho_2(1) = \rho_1(3) = 1, \ \ \rho_1 \circ \rho_2(2) = \rho_1(1) = 2, \text{ and } \rho_1 \circ \rho_2(3) = \rho_1(2) = 3.$$

So $\rho_1 \circ \rho_2$ is the identity map, and we put 1 in the entry in the intersection of the second row with the third column.

Going on like that (the computations are left to the reader), we get that the Pythagorean table of S_3 is

			$\downarrow$			
	1	ρ_1	ρ_2	σ_1	σ_2	σ_3
	ρ_1	ρ_2	1	σ_3	σ_1	$\sigma 2$
	ρ_2	1	ρ_1	σ_2	σ_3	$\sigma 1$
	σ_1	σ_2	σ_3	1	ρ_1	ρ_2
$\rightarrow$	σ_2	σ_3	σ_1	ρ_2	1	ρ_1
	σ_3	σ_1	σ_2	ρ_1	ρ_2	1

For example, the product $\sigma_2 \circ \rho_2$ is the entry σ_1 lying in the intersection of the row whose first entry is σ_2 (marked by the horizontal arrow) with the column whose first entry is ρ_2 (marked by the vertical arrow). Note that in this case the table is not symmetric with respect to the *main diagonal* (going upper left to lower right) because in S_3 the composition of permutations is not commutative: for example

$$\sigma_2 \circ \rho_2 = \sigma_1 \neq \sigma_2 = \rho_2 \circ \sigma_2.$$

A group $(G, *)$ in which the operation $*$ is commutative, is said *abelian*[2].

When we do not need to specify the operation $*$ of a group $(G, *)$ we simply write G instead of $(G, *)$ and, except for the additive notation, for two elements a and b in G, we denote their *product* $a * b$ simply by ab.

There are four fundamental concepts associated to a group: subgroups, cosets, quotients, and homomorphisms closely related to each other, which I'll deal in the next subsections.

3.2.2 Subgroups

Let G be a group. A subset H of G is called a *subgroup* if

S1 H is not empty;

S2 for every $a \in H$, $a^{-1} \in H$;

[2]After the Norwegian mathematician Niels Henrik Abel, 1802-1829.

S3 for every a and b in H, $ab \in H$.

Examples:

- **Cyclic subgroups**: Let $(G, *)$ be a group. for an element $a \in G$ and an integer z we set:

  $$a^0 := 1;$$
  $$a^z := \overbrace{a * a * \cdots * a}^{z \text{ times}} \text{ if } z > 0$$
  $$a^z := \overbrace{a^{-1} * a^{-1} * \cdots * a^{-1}}^{-z \text{ times}} \text{ if } z < 0$$

  The set $\langle a \rangle := \{a^z | z \in \mathbb{Z}\}$ is a subgroup of G, for it is clearly nonempty $1 = a^0 \in \langle a \rangle$, if a^{z_1} and a^{z_2} are elements of $\langle a \rangle$, then $(a^{z_1})^{-1} = a^{-z_1} \in \langle a \rangle$ and $a^{z_1} * a^{z_2} = a^{z_1+z_2} \in \langle a \rangle$. The subgroup $\langle a \rangle$ is called the *cyclic subgroup* of G *generated* by a. In the additive notation (that is using the symbol $+$ for the operation), it is customary to write za instead of a^z. Note that it may well happen that, for two distinct integers z_1 and z_2, $a^{z_1} = a^{z_2}$, for example, in the cyclic subgroup of S_3 generated by ρ_1, we have

  $$\rho_1 = \rho_1^4 = \rho_1^{-2} = \rho_1^7 = \rho_1^{-5} = ... = \rho_1^{3z+1}$$

  for every integer z.

- **Subgroups of** $(\mathbb{Z}, +)$: For every integer n the set

  $$n\mathbb{Z} := \{nz | z \in \mathbb{Z}\}$$

  of the multiples of n is a subgroup of $(\mathbb{Z}, +)$. This is clear for n is a multiple of n so $n\mathbb{Z}$ is not empty. Moreover, if nz is a multiple of n (with $z \in \mathbb{Z}$), its opposite $-(nz) = n(-z)$ is also a multiple of n, since $-z$ is still in $\mathbb{Z}$ (note that we are using the additive notation). Finally if nz_1 and nz_2 are multiples of n (with z_1 and z_2 in $\mathbb{Z}$), then also $nz_1 + nz_2 = n(z - 1 + z_2)$ is a multiple of n (since $z_1 + z_2$ is an integer). One can actually prove that these are the only subgroups of $(\mathbb{Z}, +)$: let H be a subgroup of $(\mathbb{Z}, +)$, then either $H \neq \{0\} = 0\mathbb{Z}$, or H contains elements different from 0. Let a be such an element, then either a or $-a$ is positive, So H contains positive integers. Let n be the minimal positive integer containbed in H then $H = n\mathbb{Z}$. In fact it is easy to see that $n\mathbb{Z} \subseteq H$, conversely, if $l \in H$, then we divide l by n and get

  $$l = nq + r \tag{3.11}$$ {rest}

  for a pair of integers q and r with $0 \leq r < n$ (r is the remainder of the division of l by n). Equation (3.11) can be rewritten as

  $$r = l - nq \tag{3.12}$$ {rest1}

  The righthand side of Equation (3.12) is an element of H, since

  $$l \in H$$

and

$$qn = \in H.$$

This implies that $r \in H$. Since $0 \leq r < n$, the minimal choice of n implies that $r = 0$ whence $l = nq \in n\mathbb{Z}$.

- **Subgroups of S_3**: By inspection of the Pythagorean table of S_2 we see that the only subgroups of S_3 are

$$S_3, A_3 := \{1, \rho_1, \rho_2\}, \{1, \sigma_1\}, \{1, \sigma_2\}, \{1, \sigma_3\}, \{1\}$$

(I leave the reader verify this).

- **Centralisers** If G is a group and $a \in G$ the set

$$C_G(a) := \{g \in G | g^{-1}ag\}$$

is a subgroup of G. Clearly $1 \in C_G(a)$, since $1^{-1}a1 = 1a1 = a1 = a$. assume $g \in C_G(a)$, then

{ag}
$$a = g^{-1}ag \tag{3.13}$$

whence, multiplying both sides by g on the left and by g^{-1} on the right, we get

$$gag^{-1} = g(g^{-1}ag)g^{-1} = (gg^{-1})a(g^{-}1g) = 1a1 = a,$$

that is $g^{-1} \in C_G(a)$. Finally if g_1 and g_2 are in $C_G(a)$, then, by Equation (3.8),

$$(g_1g_2)^{-1}a(g_1g_2) = (g_2^{-1}g_1^{-1})a(g_1g_2) = g_2^{-1}(g_1^{-1}ag_1)g_2 = g_2^{-1}ag_2 = a,$$

that is $g_1g_2 \in C_G(a)$. The subgroup $C_G(a)$ is called the *centraliser* of a in G.

- **Stabilisers** More generally, if G is a subgroup of the permutation group S_X of a set X and $x \in X$, the set

$$G_x := \{g \in G | g(x) = x\}$$

is a subgroup of G. The reader can easily prove this by himself.

{idsub}
Lemma 3.2.3 *In fact, if H is a subgroup of G, then H contains the identity 1_G of G.*

PROOF. By S1, H is non empty, so it must contain an element a. By S2 it must also contain the inverse a^{-1} of a and by S3 it must contain $aa^{-1} = 1$. ∎

{intsubg}
Lemma 3.2.4 *Let H and K be subgroups of a group G. Then $H \cap K$ is a subgroup of G.*

PROOF. Since $1 \in H \cap K$, we have $H \cap K \neq \emptyset$. Assume $a, b \in H \cap K$ then $a, b \in H$ and $a, b \in K$. Since both H and K are subgroups, ab and a^{-1} are elements of H and ab and a^{-1} are elements of K, whence ab and a^{-1} are also elements of $H \cap K$. ∎

Warning: in general it is not true that the union of two subgroups is a subgroup:

Lemma 3.2.5 *Let H and K be subgroups of a group G. Then $H \cup K$ is a subgroup of G if and only if either $H \subseteq K$ of $K \subseteq H$.*

PROOF. Obviously, if $H \leq K$, then $H \cup K = K$ which is a subgroup of G. In the same way, if $K \leq H$, $H \cup K = H$ which is again a subgroup of G. Thus assume

$$H \not\subseteq K \text{ and } K \not\subseteq H. \tag{3.14}$$ {sub}

We prove that $H \cup K$ is not closed under the multiplication. The key point is that, by (3.14), there are two elements

$$h \in H \setminus K \text{ and } k \in K \setminus H,$$

so that their product cannot be an element of $H \cup K$. In fact, assume by contradiction that

$$hk \in H \cup K.$$

Then, either $hk \in H$ or $hk \in K$. But, if $hk = h_0 \in H$, then $k = h^{-1}h_0 \in H$, against the choice of k. Similarly, if $hk = k_0 \in K$, then $h = k_0k^{-1} \in K$, again against the choice of h. ∎

Nevertheless there is still a minimal subgroup of G that contains H and K. This is the intersection, denoted by $\langle H, K \rangle$, of all subgroups that contain H and K. $\langle H, K \rangle$ is called the subgroup of G *generated by* H and K. Its elements are all the products

$$h_1^{r_1} k_1^{s_1} h_2^{r_2} k_2^{s_2} \cdots h_n^{r_n} k_n^{s_n},$$

where $n \in \mathbb{N} \setminus \{0\}$, and, for every $i \in \{1, \ldots, n\}$, $h_i \in H$, $k_i \in K$ and r_i and s_i are integers.

If X is a subset og a group G denote by X^{-1} the set

$$X^{-1} := \{x^{-1} | x \in X\}$$

of all inverses of elements of X. Note that

{inverse}

Lemma 3.2.6 *If X is a nonempty subset of G, then X is closed under the inverse if and only if $X = X^{-1}$. In particular, for every subgroup H of G, we have $H = H_{-1}$*

PROOF. The proof follows immediately from the definitions ■

If H and K are nonempty subsets of a group G, denote by

$$HK := \{h \cdot k | h \in H \text{ and } k \in K\}$$

This is the set of all elements of G that can be written as the product of an element of H and an element of K in that order. In general, if the operation $\cdot$ is not commutative it is not true that $HK = KH$. More precisely:

{HK}

Lemma 3.2.7 *If H and K are subgroups of a group G, $HK = (KH)^{-1}$*

PROOF. Let $h \cdot k \in HK$ with $h \in H$ and $k \in K$. Since H and K are subgroups, $h^{-1} \in H$ and $k^{-1} \in K$. Thus

$$h \cdot k = (h^{-1})^{-1} \cdot (k^{-1})^{-1} = (k^{-1} \cdot h^{-1})^{-1} \in (KH)^{-1}.$$

Conversely, if $(k \cdot h)^{-1} \in (KH)^{-1}$, with $k \in K$ and $h \in H$ then

$$(k \cdot h)^{-1} = h^{-1} \cdot k^{-1} \in HK$$

■

The following result shows the precise condition for HK to be a subgroup.

{prod}

Lemma 3.2.8 *Lat H and K be sugroups of a group G. Then the following assertions are equivalent:*

1. $HK = KH$
2. HK *is a subgroup of* G
3. $\langle H, K \rangle = HK$

PROOF. Assume $HK = KH$. We prove that HK is closed under the multiplication and the inverse. Let $h_1 \cdot k_1$ and $h_2 \cdot k_2$ in HK with $h_1, h_2 \in H$ and $k_1, k_2 \in K$. Since $k_1 \cdot h_2 \in KH$ and $KH = HK$ there exist $h_3 \in H$ and $k_3 \in K$ such that

{ha3}

$$k_1 \cdot h_2 = h_3 \cdot k_3 \tag{3.15}$$

So, by associativity and Equation (3.15),

$$\begin{aligned}(h_1 \cdot k_1) \cdot (h_2 \cdot k_2) &= h_1 \cdot (k_1 \cdot h_2) \cdot k_1 = h_1 \cdot (h_3 \cdot k_3) \cdot k_1 \\ &= (h_1 \cdot h_3) \cdot (k_3 \cdot k_1) \in HK,\end{aligned}$$

showing that HK is closed under the product. Further,

$$(h_1 \cdot k_1)^{-1} = k_1^{-1} \cdot h_1^{-1} \in KH = HK$$

so HK is closed also under the inverse, proving *(2)*.

Now assume HK is a subgroup. Since $H = \{h \cdot 1_G | h \in H\}$ and $1_G \in K$, we have

$$H \leq HK, \text{ and similarly } K \leq HK$$

so

$$\langle H, K\rangle \leq HK$$

since $\langle H, K\rangle$ is the smallest subgroup of G containing H and K. Conversely, $\langle H, K\rangle$ is closed under the multiplications, so it has co contain all elements $h \cdot k$ where $h \in H$ and $k \in K$, whence

$$HK \leq \langle H, K\rangle$$

and *(3)* follows.

Finally assume $\langle H, K\rangle = HK$. Then, by Lemma 3.2.7 and Lemma 3.2.6

$$KH = (HK)^{-1} = \langle H, K\rangle^{-1} = \langle H, K\rangle = HK,$$

giving *(1)*. ∎

3.2.3 Cosets

Let G be a group and H a subgroup of G. Define a relation $\sim_H$ on G in the following way: for every a and b in G

$$a \sim_H b \text{ if and only if } ab^{-1} \in H \tag{3.16}$$ {cost}

Lemma 3.2.9 *With the above notation $\sim_H$ is an equivalence relation on G.*

PROOF. We have to prove that $\sim_H$ is reflexive, symmetric and transitive.

Reflexivity: For every $a \in G$ $aa^{-1} = 1_G$ and $1_G \in H$ by Lemma 3.2.3. So $a \sim_H a$.

Symmetry: Assume $a \sim_H b$. Then $ab^{-1} \in H$. Since H is a subgroup, and $ab^{-1} \in H$, H must contain also the inverse $(ab^{-1})^{-1}$ of ab^{-1}. Since, by Equation 3.8,

$$(ab^{-1})^{-1} = (b^{-1})^{-1}a^{-1} = ba^{-1},$$

it follows that $ba^{-1} \in H$, that is $b \sim_H a$.

Transitivity: Assume a, b, and c are elements of G such that $a \sim_H b$ and $b \sim_H c$. Then $ab^{-1} \in H$ and $bc^{-1} \in H$. Since H is a subgroup, we have also $(ab^{-1})(bc^{-1}) \in H$. But then, by the associativity,

$$ac^{-1} = a(1_G)c^{-1} = a(bb^{-1})c^{-1} = (ab^{-1})(bc^{-1}) \in H,$$

whence $a \sim_H c$

■

For $a \in G$, the equivalence class $[a]_{\sim_H}$ has a nice description: set

$$Ha := \{ha | h \in H\}.$$

Then

$$[a]_{\sim_H} = Ha \tag{3.17}$$ {deesqH}

Indeed assume $b \in Ha$ we prove that $b \in [a]_{\sim_H}$. Since $b \in Ha$, by the definition of Ha, there is an element $h \in H$ such that

$$b = na.$$

multiplying both sides by h^{-1} on the left and by b^{-1} on the right we obtain

$$h^{-1} = h^{-1}bb^{-1} = h^{-1}hab^{-1} = ab^{-1}.$$

Since $h \in H$ also $h^{-1} \in H$ whence $ab^{-1} \in H$. So $a \sim_H b$, whence $b \in [a]_{\sim_H}$.

Conversely, assume $b \in [a]_{\sim_H}$. Then $ab^{-1} \in H$, whence $ba^{-1} \in H$ and, by the associativity,

$$b = b(a^{-1}a) = (ba^{-1})a \in Ha.$$

Ha is called the *right coset* of H with *representative* a. You can figure Ha as the set of the elements of H "shifted" by the element a of G.

By Lemma 2.5.2, Now consider the factor set $G/\sim H$. This is the set of the right cosets Ha where a ranges through the elements of G. By Lemma 2.5.2, $G/\sim H$ is a partition of G, so G is the disjoint union of the right cosets of H in G and each one of these cosets has as many elements as H (possibly infinitely many). This fact is of enormous importance for the study of finite groups.

{4}

Lemma 3.2.10 *For every $a \in G$, the map*

$$\begin{array}{rccc} \phi\colon & H & \to & Ha \\ & h & \mapsto & ha \end{array}$$

is bijective.

Proof. The map ϕ is surjective by the definition of right coset of H in G. Let us prove that ϕ is injective: assume h_1 and h_2 are two elements of H such that $\phi(h_1) = \phi(h_2)$. Then

$$\begin{array}{rcl} h_1 & = & h_1 1 = h_1(aa^{-1}) = (h_1 a)a^{-1} = \phi(h_1)a^{-1} \\ & = & \phi(h_2)a^{-1} = (h_2 a)a^{-1} = h_2(aa^{-1}) = h_2 1 = h_2. \end{array}$$

So ϕ is also injective, whence bijective. ■ The number of the right cosets of a

subgroup H of a group G is called the *index* of H in G and it is denoted by $|G : H|$.

{Lagrange}

Corollary 3.2.11 (LAGRANGE'S THEOREM)[3] *If G is a finite group and H is a subgroup of G, then*

$$|G| = |G : H| \cdot |H|.$$

In particular the order of H and its index in G divide the order of G.

PROOF. By Lemma **??** and Lemma **??**, G is the disjoint union of the right cosets of H in G. Assume these cosets are

$$Ha_1,\ Ha_2,\ \ldots,\ Ha_n$$

with $n = |G : H|$. Then we can count the elements of G first by counting the elements of each one of the right cosets of H and sum all $|Ha_i|$'s :

$$|G| = |Ha_1| + |Ha_2| + \cdots + |Ha_n|. \tag{3.18}$$ {x}

By Lemma 3.2.10 $|Ha_1| = |H|$ for every $i \in \{1, \ldots, n\}$, so the sum on the left of Equation (3.18), is equal to $|H| \cdot n = |H| \cdot |G : H|$. ∎

Here are some examples:

- **The right cosets of A_3 in S_3.** A_3 has precisely two right cosets in S_3. One is A_3 itself the other is $A_3\sigma_1$. So $|S_3 : A_3| = 2$ In fact, if we check the Pythagorean table of S_3, we get immediately

  $$A_3 = \{1,\ \rho_1,\ \rho_2\} = A_3\rho_1 = A_3\rho_2$$

  and

  $$A_3\sigma_1 = \{1 \circ \sigma_1,\ \rho_1 \circ \sigma_1,\ \rho_2 \circ \sigma_1\} = \{\sigma_1,\ \sigma_3,\ \sigma_2\} = A_3\sigma_2 = A_3\sigma_3$$

- **The right cosets of $\{1, \sigma\}$ in S_3.** Set $B := \{1\sigma\}$. Again from the Pythagorean table of S_3 we get that B has three right cosets in S_3 and they are

  $$B,\ B\rho_1 = \{\rho_1, \sigma_3\} = B\sigma_3, \text{ and } B\rho_2 = \{\rho_2, \sigma_2\} = B\sigma_2.$$

  So $S_3 : \{1, \sigma\}| = 3$.

- **The right cosets of $n\mathbb{Z}$ in $\mathbb{Z}$.** Clearly every integer t can be written as

  1. a multiple of n: $t = qz$ ($q \in \mathbb{Z}$), or
  2. 1 plus a multiple of n: $t = 1 + qz$, or
  3. 2 plus a multiple of n: $t = 2 + qz$, or
  4. ...

[3] Joseph-Louis Lagrange, born Giuseppe Luigi Lagrangia in 1736 in Turin, died in 1813 in Paris

5. $n-1$ plus a multiple of n: $t = (n-1) + qz$

Thus the right cosets of $n\mathbb{Z}$ in $\mathbb{Z}$ are

$$n\mathbb{Z},\ \ n\mathbb{Z}+1,\ \ n\mathbb{Z}+2,\ \ \ldots,\ \ n\mathbb{Z}+(n-1)$$

and they are all disjoint (whence distinct): assume $0 \le b \le a < n$ and

$$t \in (n\mathbb{Z}+a) \cap (n\mathbb{Z}+b)$$

then there exist q_1 and q_2 in $\mathbb{Z}$ such that $t = q_1 n + a$ and $t = q_2 n + b$, so

$$a - b = (t - q_1 n) - (t - q_2 n) = (q_1 n) - (q_2 n) = (q_1 - q_2)n$$

Since $(q_1 - q_2)n$ is a multiple of n and $0 \le a - b \le a < n$, we get $a - b = 0$, that is $a = b$. So $|\mathbb{Z} : n\mathbb{Z}| = n$.

3.2.4 Normal Subgroups and Factor Groups

{sec:normfac}

A subgroup N of a group G is called *normal* if

{defnorm}
$$\textit{for every } n \in N \textit{ and } g \in G,\ \ g^{-1}ng \in N. \tag{3.19}$$

The element $g^{-1}ng$ is called the *conjugate* of n by g. So N is normal in G if N contains all conjugates of its elements by all elements of G. Note that, since $g = (g^{-1})^{-1}$ for every $g \in G$, (3.19) could be equivalently stated as

{defnorm1}
$$\textit{or every } n \in N \textit{ and } g \in G,\ \ gng^{-1} \in N. \tag{3.20}$$

Examples:

Trivial normal subgroups: In every group G, G and $\{1\}$ are obviously normal subgroups these are called *trivial* normal subgroups. All other normal subgroups are called *proper normal subgroups*.

Normal subgroups in abelian groups: In an abelian group all subgroups are obviously normal. If the group is not abelian this is not more true in general, and actually the class of groups in which every subgroup is normal is very close to the class of abelian groups.

Normal subgroups of S_3: A_3 is the unique proper normal subgroup of S_3 (check this using Lagrange's Theorem and the multiplication table of S_3).

Thus being normal for a subgroup is quite exceptional.

If N is a normal subgroup of a group G we can multiply two right cosets of N in G and get another coset, that is we can treat the right cosets of N in G (which are subsets of G) as elements of a new group G/N.

More precisely, if N is a normal subgroup of a group G, then, for every $a, b \in G$ and for every $a' \in Na$ and $b' \in Nb$, we have

{normal}
$$Nab = Na'b'. \tag{3.21}$$

In fact, if $a' \in Na$ and $b' \in Nb$, then there are elements n_a and n_b in N such that

$$a' = n_a a \text{ and } b' = n_b b.$$

Since N is normal in G, we have $a^{-1}n_b a \in N$, whence

$$a'b' = n_a a n_b b = n_a a n_b 1 b = n_a a n_b (a^{-1}a) b = n_a (a n_b a^{-1}) ab \in Nab.$$

We can therefore define an operation on the set G/N of the right cosets of N in G by setting:

$$Na \cdot Nb := Nab. \tag{3.22}$$ {quotprod}

It is immediate to check that G/N is a group with respect to this operation.

Note that, if N is not normal, Equation (3.22) does not define an operation. E.g. Let $G = S_3$, let N be the sugbroup $\{1, \sigma_1\}$, $a = 1$, $a' = \sigma_1$, $b = \rho_1$. So $Nb = \{\rho_1, \sigma_2\}$. Set $b' := \sigma_2$. Then

$$N = Na = Na' \text{ and } Nb = Nb'.$$

But

$$Nab = N\rho_1 = Nb$$

while

$$Na'b' = N\sigma_1\rho_1 = N\sigma_2 \neq Nb$$

The problem is that we have defined the product of two cosets using two representatives of these cosets. The fact that N is normal guarantees that, by choosing different representatives, we get the same result.

When N is a normal subgroup of G, the group G/N is called the *factor group* of G *modulo* its normal subgroup N.

Examples:

The factor group S_3/A_3: Looking at the multiplication table of S_3 one can easily verify that A_3 is a normal subgroup of S_3. We have seen that A_3 has two right cosets in S_3:

$$A_3 = \{1, \rho_1, \rho_2\} \text{ and } A_3\sigma_1 = \{\sigma_1, \sigma_2, \sigma_3\}$$

Checking at the multiplication table we see that the multiplication table of the factor group S_3/A_3 is the following

A_3	$A_3\sigma_1$
$A_3\sigma_1$	A_3

The factor group $\mathbb{Z}/6\mathbb{Z}$: Since $\mathbb{Z}$ is abelian, every subgroup of $\mathbb{Z}$ is normal. In particular $6\mathbb{Z}$ is a normal subgroup of $\mathbb{Z}$. The right cosets of $6\mathbb{Z}$ in $\mathbb{Z}$ are

$$\overline{0} := 6\mathbb{Z},\ \ \overline{1} := 6\mathbb{Z}+1,\ \ \overline{0}2 := 6\mathbb{Z}+2,\ \ \overline{3} := 6\mathbb{Z}+3,\ \ \overline{4} := 6\mathbb{Z}+4,\ \ \overline{5} := 6\mathbb{Z}+5.$$

The multiplication table of $\mathbb{Z}/6\mathbb{Z}$ is the following:

$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$
$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$	$\overline{0}$
$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$	$\overline{0}$	$\overline{1}$
$\overline{3}$	$\overline{4}$	$\overline{5}$	$\overline{0}$	$\overline{1}$	$\overline{2}$
$\overline{4}$	$\overline{5}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$
$\overline{5}$	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$

Note that we can choose the representative of the class $\overline{a}+\overline{b}$ as $\overline{r}$ where r is the remainder of the division of $a+b$ by 6.

Clearly, the above argument, which is true for the integer 6, can be easily generalised to any integer n.

A group G different from $\{1\}$ without proper normal subgroups is called *simple*. By Lagrange's theorem, if the order of a group G is a prime number, then G has no proper subgroups, so no proper normal subgroups either whence it is simple. E.g. $\mathbb{Z}/2\mathbb{Z}$, $\mathbb{Z}/3\mathbb{Z}$, $\mathbb{Z}/5\mathbb{Z}$, $\mathbb{Z}/7\mathbb{Z}, \ldots$ etc. are simple. These are the only simple abelian groups. There are however simple nonabelian groups whose order is not a prime. These groups have been classified very recently (2004) and the proof of this classification is extremely long (about 15.000 pages spread on nearly 500 scientific papers). This is one of the greatest achievements of contemporary mathematics. For an historical account see [10].

3.2.5 Homomorphisms of groups

{homg}

There is a particularly interesting class maps between groups. The maps of this class are called homomorphisms and they are characterised by the fact that they preserve the operations. Precisely a map $: A \to B$ between two groups $(A, \cdot)$ and $(B, *)$ is called a *homomorphism* if, for every a_1, a_2 in A the following identity holds:

{hom}

$$f(a_1 \cdot a_2) = f(a_1) * f(a_2) \tag{3.23}$$

Before giving some examples, we list some of the elementary properties of the homomorphisms. We warmly suggest the reader to try to prove the results by him(her)self before reading the proofs given here.

In the following lemmas we assume that $(A, \cdot)$ and $(B, *)$ are groups and

$$f\colon A \to B$$

is a homomorphism of groups

{comphom}

Lemma 3.2.12 *The composition of two homomorphisms is a homomorphism*

PROOF. Assume $(C, \star)$ is another group and $g\colon B \to C$ is a homomorphism of groups and consider the composition

$$\begin{array}{rccc} g \circ f\colon & A & \to & C \\ & a & \mapsto & g(f(a)) \end{array}$$

We have to prove that, for every a_1 and a_2 in A,

$$(g \circ f)(a_1 \cdot a_2) = (g \circ f)(a_1) \star (g \circ f)(a_2) \tag{3.24}$$ {gcircf}

Since f and g are homomorphisms

$$f(a_1 \cdot a_2) = f(a_1) * f(a_2) \text{ and } g(f(a_1) * f(a_2)) = g(f(a_1)) \star g(f(a_2)) \tag{3.25}$$ {f}

It follows that

$$\begin{array}{rcl} (g \circ f)(a_1 \cdot a_2) & = & g(f(a_1 \cdot a_2)) = g(f(a_1) * f(a_2)) = g(f(a_1)) \star g(f(a_2)) \\ & = & (g \circ f)(a_1) \star (g \circ f)(a_2). \end{array}$$

■

A bijective homomorphism of groups is called an *isomorphism.*

{idhom}

Lemma 3.2.13 *The map*

$$\begin{array}{rccc} id_A\colon & A & \to & A \\ & a & \mapsto & a \end{array}$$

is a isomorphism

PROOF. This is obvious: if a_1 and a_2 are elements of A, then

$$id_A(a_1) = a_1,\ \ id_A(a_2) = a_2, \text{ and } id_A(a_1 \cdot a_2) = a_1 \cdot a_2,$$

whence

$$id_A(a_1 \cdot a_2) = a_1 \cdot a_2 = id_A(a_1) \cdot id_A(a_2)$$

■

{inversehom}

Lemma 3.2.14 *If $f\colon A \to B$ is an isomorphism, then its inverse function $f^{-1}\colon B \to A$ is also an isomorphism.*

PROOF. f^{-1} is bijective, since it is the inverse of a bijective function. So we just have to prove that f^{-1} is a homomorphism. Let b_1 and b_2 be two elements of B. Since f is bijective there exist a_1 and a_2 in A such that

$$f(a_1) = b_1 \text{ and } f(a_2) = b_2.$$

Then, since f is a homomorphism and $f^{-1} \circ f = id_A$,

$$\begin{aligned} f^{-1}(b_1 * b_2) &= f^{-1}(f(a_1) * f(a_2)) = f^{-1}(f(a_1 \cdot a_2)) = (f^{-1} \circ f)(a_1 \cdot a_2) \\ &= a_1 \cdot a_2 = f^{-1}(b_1) \cdot f^{-1}(b_2)) \end{aligned}$$

■

A homomorphism of groups between A and A itself is called an *endomorphism*. If it is also bijective it is called an *automorphism*. The set of all automorphisms of a group A is denoted by $Aut(A)$.

{AutA}

Proposition 3.2.15 *$(Aut(A), \circ)$ is a group*

PROOF. By Lemma 3.2.12 and Lemma 2.4.2 the composition of two automorphisms of A is an automorphism of A, so $\circ$ is an operation on $Aut(A)$. It is associative by Lemma 2.4.3. The map id_A is the identity of $(Aut(A), \circ)$ and, by Lemma 3.2.14, every element g in $Aut(A)$ has an inverse g^{-1} in $Aut(A)$. ■

For every group A, the group $Aut(A)$ is called the *automorphism group* of the group A. It can be thought as the group of all symmetries of the group A.

Denote by 1_A and by 1_B the identities of the groups A and B respectively.

{image1}

Lemma 3.2.16 *$f(1_A) = 1_B$*

PROOF. Since f is an homomorphism and $1_A \cdot 1_A = 1_A$, we have

$$f(1_A) = f(1_A \cdot 1_A) = f(1_A) * f(1_A)$$

whence, multiplying both sides by $(f(1_A))^{-1}$ on the left , we get

$$\begin{aligned} 1_B &= (f(1_A))^{-1} * f(1_A) = (f(1_A))^{-1} * (f(1_A) * f(1_A)) \\ &= ((f(1_A))^{-1} * f(1_A)) * f(1_A)) = 1_B * f(1_A) = f(1_A) \end{aligned}$$

■

{invhom}

Lemma 3.2.17 *For every a in A, $f(a^{-1}) = (f(a))^{-1}$*

PROOF. By Lemma 3.2.16

$$1_B = f(1_A) = f(a \cdot a^{-1}) = f(a) * f(a)^{-1}),$$

whence, multiplying both sides on the left by $(f(a))^{-1}$, we get

$$\begin{aligned} (f(a))^{-1} &= (f_(a))^{-1} * 1_B = (f(a))^{-1} * (f(a) * f(a^{-1})) \\ &= ((f(a))^{-1} * f(a)) * f(a^{-1}) = 1_B * f(a^{-1}) = f(a^{-1}). \end{aligned}$$

■

{subgimage}

Lemma 3.2.18 *If H is a subgroup of A, then $f(H)$ is a subgroup of B*

PROOF. Since H is a subgroup of A, $1_A \in H$, so, by Lemma 3.2.16, $1_B = f(1_A) \in f(H)$, whence $f(H)N \neq \emptyset$. We prove now that $f(H)$ is closed under the operation $*$. Assume b_1 and b_2 are elements of $f(H)$. Then there are two elements a_1 and a_2 in H such that

$$f(a_1) = b_1, \text{ and } f(a_2) = b_2.$$

Since H is a subgroup, $a_1 \cdot a_2 \in H$, so

$$b_1 * b_2 = f(a_1) * f(a_2) = f(a_1 \cdot a_2) \in f(H).$$

Finally, assume $b \in f(H)$, then there exists $a \in H$ such that $b = f(a)$. Since H is a subgroup, $a^{-1} \in H$, whence, by Lemma 3.2.17,

$$b^{-1} = (f(a))^{-1} = f(a^{-1}) \in f(H),$$

so $f(H)$ contains all inverses of its elements, whence $f(H)$ is a subgroup. ∎

{invimagesub}

Lemma 3.2.19 *If K is a subgroup of B, then $f^{-1}(K)$ is a subgroup of A. Moreover, if K is normal in B then also $f^{-1}(K)$ is normal in A.*

PROOF. Since K is a subgroup, by Lemma 3.2.16, $f(1_A) = 1_B \in K$, so $1_A \in f^{-1}(K)$ and $f^{-1}(K) \neq \emptyset$. Assume a_1 and a_2 are elements of $f^{-1}(K)$. Then $f(a_1)$ and $f(a_2)$ are elements of K. Since f is a homomorphism and K is a subgroup,

$$f(a_1 \cdot a_2) = f(a_1) * f(a_2) \in K,$$

whence $a_1 \cdot a_2 \in f^{-1}(K)$. Finally assume $a \in f^{-1}(K)$, then, by Lemma 3.2.17,

$$f(a^{-1}) = (f(a))^{-1} \in K,$$

so also $a^{-1} \in f^{-1}(K)$.

Now assume K is normal in B. Let $a \in f^{-1}(K)$ and $g \in A$, then, since $f(a) \in K$b and K is normal in B, by Lemma 3.2.17, we have

$$f(g^{-1}ag) = f(g)^{-1}f(b)f(g) \in K.$$

Thus $g^{-1}ag \in f^{-1}(K)$, whence $f^{-1}(K)$ is normal in K. ∎

Since $\{1_B\}$ is trivially a normal subgroup of B, by Lemma 3.2.19 it follows that $f^{-1}(1_B)$ is a normal subgroup of A (see laso Exercise 3.4.2). This subgroup is called the *kernel* of f and is denoted by $\ker(f)$. By definition $\ker(f)$ is the set of the elements of A that are mapped to the identity of B.

{equivkern}

Remark 3.2.20 *Note that the equivalence $\sim_f$ associated to the homomorphism f and the equivalence $\sim_{ker(f)}$ associated to the subgroup* $\ker(f)$ *are actually the same. In fact, for every a, b in G,*

$$\begin{aligned} a \sim_f b \quad &\iff \quad f(a) = f(b) \iff f(a)f(b)^{-1} = 1 \\ &\iff \quad f(ab^{-1}) = 1 \iff ab^{-1} \in \ker(f) \iff a \sim_{\ker(f)} b \end{aligned}$$

{ker}

Lemma 3.2.21 *The homomorphism f is injective if and only if* $\ker(f) = 1_A$.

PROOF. Assume f is injective. Clearly $\{1_A\} \subseteq \ker(f)$. Conversely if $a \in \ker(f)$. Then, by Lemma 3.2.16, $f(a) = 1_b = f(1_A$, whence $a = 1_A \in \{1_A\}$ so $\ker(f) \subseteq \{1_A\}$, whence $\ker(f) = \{1_A\}$. Now assume $\ker(f) = \{1_A\}$ and let a_1 and a_2 be two elements of A such that $f(a_1) = f(a_2)$. Then

$$f(a_1 \cdot a_2^{-1}) = f(a_1) * f(a_2^{-1}) = f(a_1) * (f(a_2))^{-1} = f(a_1) * (f(a_1))^{-1} = 1B,$$

whence

$$a_1 \cdot a_2^{-1} \in \ker(f) = \{1_A\}.$$

Thus $a_1 \cdot a_2^{-1} = 1_A$, that is $a_1 = a_2$, so f is injective. ∎

Examples:

- **The exponential:** Let $\mathbb{R}^+$ be the set of positive real numbers. Then it is easy to see that $(\mathbb{R}^+, \cdot)$ is a group (here $\cdot$ is the usual multiplication in $\mathbb{R}$).

  $$\begin{array}{rccc} exp_2\colon & \mathbb{Z} & \to & \mathbb{R}^+ \\ & z & \mapsto & 2^z \end{array}$$

  is a homomorphism between the (additive) group $(\mathbb{Z}, +)$ and the (multiplicative) group $(\mathbb{R}^+, \cdot)$. Indeed, for every pair of integers a and b

  $$exp_2(a + b) = 2^(a + b) = 2^a \cdot 2^b = exp_2(a) \cdot exp_2(b).$$

- **The logarithm:** If we extend the domain of exp_2 to the additive group of all real numbers, we get that $exp_2\mathbb{R} \to \mathbb{R}^+$ is also an isomorphism whose inverse is the logarithm in basis 2:

  $$\begin{array}{rccc} log_2\colon & \mathbb{R}^+ & \to & \mathbb{R} \\ & x & \mapsto & log_2(x) \end{array}$$

The following lemmas show more examples which will be important in the sequel.

{conjugation}

Lemma 3.2.22 (Conjugation by an element in a group) *Let $(G, \cdot)$ be a group and $g \in G$. Define*

$$\begin{array}{rccc} \gamma_g\colon & G & \to & G \\ & a & \mapsto & g \cdot a \cdot g^{-1} \end{array}$$

Then γ_g is an automorphism of G.

PROOF. For every a and b in G,

$$\begin{aligned}\gamma_g(a\cdot b) &= g\cdot(a\cdot b)\cdot g^{-1}\\ &= (g\cdot a)\cdot(b\cdot g^{-1})\\ &= (g\cdot a)\cdot 1_G\cdot(b\cdot g^{-1})\\ &= (g\cdot a)\cdot(g^{-1}\cdot g)\cdot(b\cdot g^{-1})\\ &= (g\cdot a\cdot g^{-1})\cdot(g\cdot b\cdot g^{-1})\\ &= \gamma_g(a)\cdot\gamma_g(b)\end{aligned}$$

So γ_g is an endomorphism of G. But it is also bijective, its inverse being $\gamma_{g^{-1}}$. ∎

The automorphism γ_g is called the *inner automorphism* induced by conjugation by g. Now, we have seen that if G is a group, $(Aut(G),\circ)$ is also a group.

Lemma 3.2.23 (The action of a group on itself by conjugation) *Let $(G,\cdot)$ be a group. Consider the map* {actconj}

$$\begin{array}{rccc}\gamma\colon & G & \to & Aut(G)\\ & g & \mapsto & \gamma_g\end{array}$$

Then γ is a homomorphism of groups.

PROOF. We have to prove that, for every $g,h\in G$,

$$\gamma_{g\cdot h}=\gamma_g\circ\gamma_h, \tag{3.26}$$ {cra}

(since, by definition, $\gamma(g\cdot h)=\gamma_{g\cdot h}$, $\gamma(g)=\gamma_g$, and $\gamma(h)=\gamma_h$). Both sides of Equation 3.26 are automorphisms of G, in particular they have the same domain and the same codomain. Thus, to prove that they are equal, we have to prove that, for every $a\in G$

$$\gamma_{g\cdot h}(a)=(\gamma_g\circ\gamma_h)(a).$$

Now

$$\begin{aligned}\gamma_{g\cdot h}(a) &= (g\cdot h)\cdot a\cdot(g\cdot h)^{-1})\\ &= (g\cdot h)\cdot a\cdot(h^{-1}\cdot g^{-1})\\ &= g\cdot(h\cdot a\cdot h^{-1})\cdot g^{-1}\\ &= g\cdot\gamma_h(a)\cdot g^{-1}\\ &= \gamma_g(\gamma_h(a))\\ &= (\gamma_g\circ\gamma_h)(a)\end{aligned}$$

So γ is a homomorphism. ∎ The image of γ is the set $Inn(G)$ of the inner automorphisms of G. By Lemma 3.2.18 $Inn(G)$ is a subgroup of $Aut(G)$. The kernel of γ is the set

$$\{g\in G|\gamma_g=id_G\},$$

that is the set

$$\{g \in G | gag^{-1} = a \text{ for every } a \in G\}$$

or equivalently, the set

$$\{g \in G | ga = ag \text{ for every } a \in G\}$$

of the elements g of G that *commute* with every other element a of G. This set is a subgroup of G by Lemma 3.2.19, it is called the *center* of G and denoted by $Z(G)$. Clearly G is abelian if and only if $G = Z(G)$. In this case γ maps every element of G to the identity map id_G of G.

{canproj} **Lemma 3.2.24 (The projection on a factor group)** *Let $(G, \cdot)$ be a group, N be a normal subgroup of G and let G/N be the factor group of G modulo N. Let*

$$\begin{array}{rccc} \pi\colon & G & \to & G/N \\ & g & \mapsto & Ng \end{array}$$

be the canonical projection, then π is a (surjective) homomorphism of groups.

PROOF. We have already seen in Subsection 2.5.3 that π is surjective . Now assume a and a' are elements of G then, by Equation 3.21,

$$\pi(aa') = Naa^prime = NaNa' = \pi(a)\pi(a').$$

■

{firstisog} **Theorem 3.2.25** (FIRST HOMOMORPHISM THEOREM FOR GROUPS *Let $f\colon G \to H$ be a homomorphism of groups. $\pi\colon G \to G \to \ker(f)$ the canonical projection of G onto the factor group $G/\ker(f)$. Then there is a unique injective group homomorphism $\bar{f}\colon G/\ker(f) \to H$ such that $\bar{f} \circ \pi = f$. In particular $im(\bar{f}) = im(f)$.*

PROOF. By Remark 3.2.20 the equivalence $\sim_f$ associated to the homomorphism f coincides with the equivalence
$sim_{\ker(f)}$ associated to the subgroup $\ker f$. So the existence, the uniqueness, the injectivity of $\bar{f}$ and the fact that $im(\bar{f}) = im(f)$ follow from Theorem 2.5.4. We are left to prove that $\bar{f}$ is a homomorphism. Let $N = \ker(f)$, Na and Na' in G/N. Then, since f is a homomorphism of groups,

$$\bar{f}(NaNa') = \bar{f}(Naa') = f(aa') = f(a)f(a') = \bar{f}(Na)\bar{f}(Na').$$

■

Assume G is a group of permutations of a set X and let Y be a *G-invariant* subset of X, that is $g(y) \in Y$ for every $g \in G$ and $y \in Y$. Then every element of G induces *by restriction* a permutation $g_{|_Y}$ of the set Y defined by

$$\begin{array}{rccc} g_{|_Y}\colon & Y & \longrightarrow & Y \\ & y & \mapsto & g(y) \end{array}$$

{restsub} **Lemma 3.2.26 (The restriction to an invariant subset)** *With the above notation, The map*

$$\begin{array}{rccc} {}_{|Y}\colon & G & \longrightarrow & S_Y \\ & g & \mapsto & g_{|Y} \end{array}$$

Is a homomorphism of groups.

PROOF. We have to prove that, if $a, b \in G$,

$$(a \circ b)_{|Y} = a_{|Y} \circ b_{|Y}.$$

Again both maps are permutations of Y, so we just need to prove that, for every $y \in Y$,

$$(a \circ b)_{|Y}(y) = (a_{|Y} \circ b_{|Y})(y).$$

But, by definition of ${}_{|Y}$,

$$(a \circ b)_{|Y}(y) = (a \circ b)(y) = a(b(y)) = a(b_{|Y}(y)) = a_{|Y}(b_{|Y}(y)) = (a_{|Y} \circ b_{|Y})(y).$$

■

3.2.6 The classification of the cyclic groups

We apply some of the above result to obtain a classification, up to isomorphism, of all cyclic groups.

{classfin}

Theorem 3.2.27 *Let G be a cyclic group, then there exists a subgroup K of $(\mathbb{Z}, +)$ such that G is isomorphic to $\mathbb{Z}/K$.*

PROOF. Suppose g is a generator of G, then, by definition of a cyclic group,

$$G = \{g^z \ Z \in \mathbb{Z}\}$$

(note that we use the multiplicative notation for G and the additive notation for $\mathbb{Z}$). Now let $f\colon \mathbb{Z} \to G$ be the map defined, for every $z \in \mathbb{Z}$, by

$$z \mapsto g^z.$$

Since, for every z_1 and z_2 in $\mathbb{Z}$, $g^{z_1+z_2} = g^{z_1} g^{z_2}$ it follows that

$$f(z_1 + z_2) = g^{z_1+z_2} = g^{z_1} g^{z_2} = f(z_1) f(z_2),$$

that is ϕ is a homomorphism of groups and it is surjective since G is cyclic generated by g. By the First Homomorphism Theorem for groups (Theorem 3.2.25), it follows that $G = im(f) \cong \mathbb{Z}/\ker(f)$ whence the thesis with $K = \ker(f)$. ■

3.3 Permutation groups

Given a set X, a subgroup G of the group S_X of all permutations of X is called a *permutation group* on X. Actually, for every group G one can construct a set X such that G is isomorphic to a permutation group on X (see exercise 3.4.5), so every group can be regarded as a permutation group. It is however convenient to have at hand the (passive) set X in order to understand how the (active) group G works. The principal concepts to have in mind when dealing with a permutation group G on a set X are

1. the orbits (which are subsets of X),
2. the G-invariant subsets of X, and
3. the stabilisers in G of the points of X (these are subgroups of G).

In the sequel, let G be a permutation group on a set X.

3.3.1 Orbits

Let $x \in X$. We set

$$G(x) := \{g(x) | g \in G\}$$

$G(x)$ is called the G-*orbit* of x. It is the set of all possible images of x via the elements of G.

Example:

- **The orbits for a subgroup of S_3**
  - If $G = S_3$

    $G(1) = \{1_G(1), \rho_1(1), \rho_2(1), \sigma_1(1), \sigma_2(1), \sigma_3(3)\} = \{1, 2, 3, 1, 3, 2\} = \{1, 2, 3\}$.

    The same holds also for $G(2)$ and $G(3)$.
  - Similarly, if $H = \langle \rho \rangle$,

    $$H(1) = \{1_G(1), \rho_1(1), \rho_2(1)\} = \{1, 2, 3\}$$

    and, as above this set is also equal to $H(2)$ and $H(3)$.
  - If $K = \langle \sigma_1 \rangle$,

    $K(1) = \{1_G(1), \sigma_1(1)\} = \{1, 1\} = \{1\}$ and $K(2) = \{1_G(2), \sigma_1(2)\} = \{2, 3\}$

    .
  - The reader can now easily complete the list with all the other subgroups of S_3

We say that G is *transitive* on a set X if X is a G-orbit, or, equivalently, for every $x, y \in X$, there is $g \in G$, such that $g(x) = y$. In the above examples G and H are transitive, while K is not (there is no element of K that maps 1 to 2).

{orb1}

Lemma 3.3.1 *If x and y are elements of X and $y \in G(x)$, then $G(x) = G(y)$*

PROOF. if $y \in G(x)$ then there is an element g of G such that

$$y = g(x), \text{ whence } x = g^{-1}(x).$$

If $z \in G(y)$, then there is an element $h \in G$ such that $z = h(y)$, whence

$$z = h(y) = h(g(x)) = (h \circ g)(x) \in G(x),$$

which proves that $G(y) \subseteq G(x)$.

Conversely assume $t \in G(x)$, then there is an element $k \in G$ such that $t = k(x)$, whence

$$t = k(x) = k(g^{-1}(x)) = (k \circ g^{-1})(y) \in G(y),$$

which proves that $G(x) \subseteq G(y)$. ∎

{orb}

Lemma 3.3.2 *The set of the G-orbits is a partition of X.*

PROOF. Clearly for every $x \in X$

$$x = 1_G(x) \in G(x),$$

so every element of X lies in a G-orbit and X is therefore the union of all its G-orbits. Now assume $G(x)$ and $G(y)$ have a non empty intersection and let z be an element of $G(x) \cap G(y)$. Then, by Lemma 3.3.1,

$$G(x) = G(z) = G(y).$$

∎

Note that, by definition, an orbit is a G-invariant subset of X: Indeed, if $y \in G(x)$ then there is an element $h \in G$ such that $y = h(x)$. Thus, for every $g \in G$,

$$g(y) = g(h(x)) = (g \circ h)(x) \in G(x).$$

Conversely a minimal nonempty G-invariant subset Y of X is an orbit: if $y \in Y$, then $G(y) \leq Y$ because Y is G-invariant. So $Y = G(y)$ by the minimal choice of Y. Note also that every non empty G-invariant subset is the disjoint union of its G-orbits.

If $g(y) = y$ for every $g \in G$, or equivalently $\{y\}$ is a G-orbit, we say that y is a *fixed point* for G. If $G = \langle g \rangle$ we shall also say that y is a fixed point for g.

3.3.2 Stabilisers

For $x \in X$ denote by G_x the set

$$G_x := \{g \in G | g(x) = x\}$$

G_x is called the *stabiliser* of x in G.

Lemma 3.3.3 *G_x is a subgroup of G.*

PROOF. Since $id_X(x) = x$, we have $id_X \in G_x$. Let $a, b \in G_x$, then $a(x) = b(x) = x$, so

$$(a \circ b)(x) = a(b(x)) = a(x) = x,$$

and

$$a^{-1}(x) = a^{-1}(a(x)) = (a^{-1} \circ a)(x) = id_X(x) = x,$$

whence $a \circ b$ and a^{-1} are elements of G_x and so G_x is a subgroup. ∎

Example:

- **The stabiliser of a point in S_3** Let $G := S_3$ and choose $x = 1$. We want to determine the subgroup G_1 of the permutations of G that fix 1. If a permutation σ of the set $\{1, 2, 3\}$ fixes 1, then it has to permute the elements 2 and 3. Conversely any permutation of the set $\{2, 3\}$ can be extended to a permutation of $\{1, 2, 3\}$ that fixes 1. So

  $$G_1 = \{1_G, \sigma_1\}$$

  (warning: do not confuse the element 1 of the set $\{1, 2, 3\}$ with the identity 1_G of G, which is the identiy map on $\{1, 2, 3\}$)

  There is also a more elegant way to see this: Since $\{2, 3\}$ is G_1-invariant the restriction

  $$|_{\{2,3\}}: \begin{array}{ccc} G_1 & \longrightarrow & S_{\{2,3\}} \\ \sigma & \mapsto & \sigma_{|_{\{2,3\}}} \end{array}$$

  is a homomorphism of groups and, as one sees easily, it is bijective.

The most important result about finite permutation groups is the following

{indexstab}

Theorem 3.3.4 *Let G be a finite permutation group on a set X and let $x \in X$. Then*

$$|G(x)| = |G : G_x|$$

PROOF. We prove the theorem by constructing a bijection between the set $|G/\ |G_x$ of the right cosets of G_x in G and the orbit $G(x)$. First of all, note that if $h, g \in G$ with $h \in G_x \circ g$ of G_x in G, then there is $k \in G_x$, such that $g = k \circ h$, so

$$g^{-1}(x) = (k \circ h)^{-1} = (h^{-1} \circ k^{-1})(x) = h^{-1}(k^{-1}(x))$$

but k and k^{-1} are elements of G_x so they fix x, that is $k^{-1}(x) = x$, and so

$$h^{-1}(k^{-1}(x) = h^{-1}(x).$$

This shows that, for every element $h \in G_x \circ (g)$,

$$h^{-1}(x) = g^{-1}(x).$$

We can therefore associate to each coset $G_x(g)$ of G_x in G the element $g(x)$ and this element does not depend on the choce of the representative g of $G_x \cdot g$, but only on the coset. So we have a map

$$\begin{array}{rccc} \phi\colon & G/\sim_{G_x} & \longrightarrow & G(x) \\ & G_x \circ g & \mapsto & g(x) \end{array}$$

Since every element of G is contained in a right coset of G_x in G, this map is surjective. We prove that it is also injective, hence bijective. Assume a and b are elements of G such that

$$\phi(G_x \circ a) = \phi(G_x \circ b)$$

Then, by definition of ϕ,

$$a^{-1}(x) = b^{-1}(x)$$

whence, applying a to both sides, we get

$$x = id_X(x) = (a \circ a^{-1})(x) = a(a^{-1}(x)) = a(b^{-1}(x)) = (a \circ b^{-1})(x),$$

that is

$$a \circ b^{-1} \in G_x.$$

But this means that

$$a = a \circ 1_G = a \circ (b^{-1} \circ b) = (a \circ b^{-1})b \in G_x \circ b,$$

whence $G_x(a) = G_x(b)$, by Lemma **??**. So ϕ is bijective and the result follows. ∎

3.3.3 The Frattini Argument

We conclude this section with a simple though very important result: the Frattini Argument[4] which shows under what conditions you can reach a configuration of a set X via certain permutations of a subgroup G of S_X that leave fixed some subset of X. This is actually what enables you to solve the 15-Puzzle (or the Rubik's cube): you start putting the tile #1 in its right position (say upper left) and then try to put the tile #2 in the next position in the first row with permutations that leave fixed tile #1, that is with permutations that lie in the stabiliser of tile #1, and go on iterating this procedure using permutations that lie in the stabiliser of the tiles that are already in the right position.

Here's the statement of the Frattini Argument and its proof.

[4] After Giovanni Frattini (1852-1925) even though the argument is also attributed to Alfredo Capelli (1855-1910)

Theorem 3.3.5 (The Frattini Argument) *Let G be a permutation group of a set X and let H be a subgroup of G. If H is transitive on X, then*

$$G = G_x H$$

for every $x \in X$,

Proof. Let $x \in X$. Then, for every $g \in G$, x and $g(x)$ are elements of X. Since H is transitive on X, there is an element $h \in H$ such that

{Frattini}
$$g(x) = h(x) \tag{3.27}$$

taking the images of both sides under h^{-1} we have

$$(h^{-1} \circ g)(x) = h^{-1}(g(x)) = h^{-1}(h(x)) = x,$$

that is

$$h^{-1} \circ g \in G_x.$$

But then, by associativity,

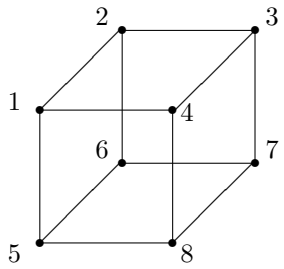
$$g = g \circ 1_G = g \circ (h^{-1} \circ h) = (g \circ h^{-1}) \circ h \in G_x H$$

■

3.3.4 Applications

Theorem 3.3.4 together with Lagrange's Theorem (Corollary 3.2.11 is extremely useful for computing the symmetry groups of certain structures.

The symmetry group of a cube: Consider a cube and label the vertices with the numbers $1, 2, 3, 4, 5, 6, 7, 8$ as in the picture below. Call two vertices *adjacent* if they lie in the same edge (so, e.g. the vertices adjacent to 1 are 2, 4 and 5, while the others are not adjacent to 1). Let G be the symmetry group of the cube. This can be regarded as the set of all permutations σ of the vertices of the cube that preserve adjacency, that is if two vertices a and b are adjacent, then also $\sigma(a)$ and $\sigma(b)$ are adjacent. (so, eg. if $\sigma(1) = 8$, then, since 1 and 2 are adjacent, $\sigma(2)$ has to be in the set $\{5, 4, 7\}$, for the other vertices are not adjacent to 8).



We want to show how Theorem 3.3.4 and Lagrange's Theorem can be used to compute the order of G. Clearly any rotation around a central axis is a symmetry of the cube and, since the vertex 1 can be sent to any of the eight vertices, we have that

$$G(1) = \{1, 2, 3, 4, 5, 6, 7, 8\},$$

so, by Theorem 3.3.4

$$|G : G_1| = |G(1)| = 8 \tag{3.28}$$ {G_1}

We are thus reduced to compute the order of G_1, the stabiliser in G of the vertex 1. To do so we compute the length of the G_1 orbit of the vertex 2. If σ is an element of G_1, then σ has to send the vertex 2 into another vertex adjacent to 1, i.e. one in the set $\{2, 4, 5\}$. It is easy to see that the three rotations of 120°, 240° and 360° around the diagonal through the vertices 1 and 7 are in G_1 and send 2 to 5, 4, and 2. So

$$G_1(2) = \{2, 4, 5\}$$

and

$$|G_1 : G_{1,2}| = |G_1(2)| = 3. \tag{3.29}$$ {G_12}

where $G_{1,2}$ is the stabiliser of the vertex 2 in G_1. We now compute the order of $G_{1,2}$, as above we compute the length of the $G_{1,2}$-orbit of 4. If $\tau \in G_{1,2}$ then, since 4 is adjacent to 1, τ has to send 4 to a vertex adjacent to 1 and different from 2 (why?), that is $G_{1,2}(4) \subseteq \{4, 5\}$. Clearly the identity sends 4 to 4 and is an element of $G_{1,2}$. But it is possible also to sent 4 to 5 by the element of $G_{1,2}$ that turns the cube inside out. So $G_{1,2}(4) = \{4, 5\}$ and

$$|G_{1,2} : G_{1,2,4}| = |G_{1,2}(4)| = 2. \tag{3.30}$$ {G-124}

At this stage we are done for if ρ is an element of $G_{1,2,4}$, then ρ has also to stabilise 5, for this is the unique point adjacent to 1 left. But then it has also to fix 6 for 6 is the unique point adjacent to 5 and 2 other than 1. Similarly it has to fix 3 and 8, but then it must fix also 7, for there are no more points to send 7. Thus ρ is the identity map and $G_{1,2,4} = \{1_G\}$. Now, by Lagrange's Theorem,

$$\begin{aligned} |G| &= |G : G_1| \cdot |G_1| = |G : G_1| \cdot |G_1 : G_{1,2}| \cdot |G_{1,2}| \\ &= |G : G_1| \cdot |G_1 : G_{1,2}| \cdot |G_{1,2} : G_{1,2,4}||G_{1,2,4}| = 8 \cdot 3 \cdot 2 \cdot 1 = 48 \end{aligned}$$

That is the cube has precisely 48 symmetries.

The isometries of the Euclidean plane: We will define properly what the Euclidean plane is in the next chapters. For the moment just think of it as the set E of points and lines you have studied at high school, endowed with the usual distance between two points.

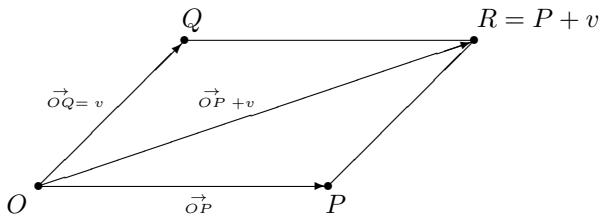
Fix a distinguished point O, called the *origin* of the plane, and define the *space of vectors* V_E associated to E as the set of all *vectors* with *origin* in P, i.e. the set of all pairs (O, P) with $P \in E$. Note that the map $P \mapsto (O, P)$

defines a bijection between E and V_E. Usually the vector (O, P) is denoted by $\overrightarrow{OP}$. Vectors can be summed with the usual parallelogram rule and, endowed with this operation, V_E is a commutative group with identity $\overrightarrow{OO}$.

An *isometry* (or *rigid motion*) of the Euclidean plane is a permutation σ of the points that sends lines to lines and preserve the distance between points, that is: if L is a line, then also $\sigma(L)$ is a line and for any two points P and Q the distance between P and Q is the same as the distance between their images $\sigma(P)$ and $\sigma(Q)$. The set of isometries of the Euclidean plane E will be denoted by $\mathcal{I}(E)$.

Examples of isometries:

- **Translation along a vector**: If P is a point in E and $v := \overrightarrow{OQ}$ is a vector in V_E, we define $P + v$ as the point R of E that corresponds to the vector $\overrightarrow{OP} + v$ as in the picture below:



  It can be shown (and we shall do that once we have formally defined the Euclidean plane) that, given a vector v, the map

$$\begin{array}{rccc} \tau_v\colon & E & \to & E \\ & P & \mapsto & P + v \end{array}$$

  is an isometry of E. This map is called the *translation along* the vector v. It shifts all points of the Euclidean plane parallely to the line through O and Q for the same length of the vector v. To understand it, stick a piece of white paper to your drawing table and mark some points on it. Then lay over the sheet of paper another sheet of paper (preferably translucent) and mark on the second paper the points you marked on the first sheet. Finally shift the second paper in one direction without rotating it and fix the second paper in the new position: you can do that by sticking the second paper to a set square that can slide on a parallel straightedge (you heard about those instruments? They were the architect's working tools until some decades ago). Then you will see that all the points marked (but actually all the points) in the second paper are the points on the first paper shifted by the same vector.

Denote by $\mathcal{T}(E)$ the set of all translations of the Euclidean plane E. Note that if v and w are vectors in V_E,

$$\tau_{v+w} = \tau_v \circ \tau_w,$$

hence the map

$$\begin{array}{rccc} \tau\colon & V_E & \to & \mathcal{I}(E) \\ & v & \mapsto & \tau_v \end{array}$$

is a homomorphism (actually an isomorphism!) of groups and $\mathcal{T}(E) = \tau(V_E)$ is a subgroup of $\mathcal{I}(E)$. Note that $\mathcal{T}(E)$ is transitive on the set of points of E. Indeed if P is a point of E, set $v := \overrightarrow{OP}$. Then

$$P = O + v = \tau_v(O),$$

so every point P in E belongs to the $\mathcal{T}(E)$-orbit of O.

- **Rotation by an angle α around a point**: At present this is difficult to define formally, but easy to describe: Take again two sheets of paper, one matt and one translucent. Stick the matt one to your drawing table and mark again some points. Then put the translucent paper over it and fix it with a pin in the center (so that the second sheet may rotate around the pin) and mark the same points on the second paper. Then, after rotating the second around the pin by a certain angle, say α, fix it. Again you will see that the point where you have put the pin is left fixed and all the other points in the second sheet lie in the same circles as the corresponding points on the first sheet, but all rotated at the same angle α. The rotation around the origin by the angle α will be denoted by ρ_α and the set of rotations around O will be denoted by $\mathcal{R}(E)$. Now we can identify the set of angles with the (additive) factor group $(\mathbb{R}/360\mathbb{R}, +)$ of $(\mathbb{R}, +)$ modulo its subgroup $360\mathbb{R}$ and Again, for two angles α and β, we have

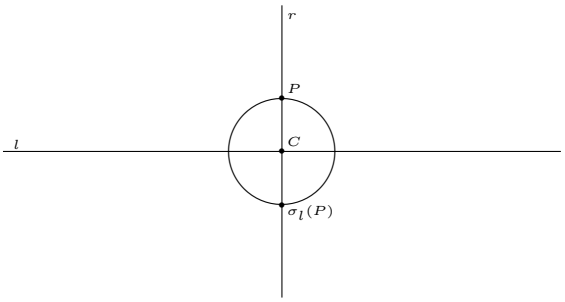
$$\rho_{\alpha+\beta} = \rho_\alpha \circ \rho_\beta.$$

so the map

$$\begin{array}{rccc} \tau\colon & \mathbb{R}/360\mathbb{R} & \to & \mathcal{R}(E) \\ & \alpha & \mapsto & \rho_\alpha \end{array}$$

is an isomorphism of groups, so $\mathcal{R}(E)$ is a subgroup of $\mathcal{I}(E)$. This group is by no means transitive on E. the $\mathcal{R}(E)$ orbits of a point P is the set of all points that have the same distance as P from O, i.e. the points lying in the circle through P with center O.

- **Reflections through a line**. Let l be a line in E and P a point. Consider the line r through P orthogonal to l, let C be the unique point in $l \cap r$ and $\mathcal{C}$ the circle through P and center C and let $\sigma_l(P)$ be the intersection of $\mathcal{C}$ with r other than P, as in the picture below:



Denote by σ_l the map from E to E that maps each point P of E to the point $\sigma_l(P)$. Then σ_l is an isometry and it is called the *reflection* with *axis* l. The axis l is the set of fixed points of σ_l. You can think of l as a sort of one dimensional mirror so that every element of E is sent by σ_l to his mirrored image. Note that $\sigma_l \circ \sigma_l$ leaves every point fixed (the mirrored image of the mirrored image of a point P is the point P). Thus $\sigma_l^{-1} = \sigma_l$ and $\mathcal{S}_l := \{id_E, \sigma_l\}$ is a subgroup with two elements of $\mathcal{I}(E)$.

Now we apply the Frattini Argument to the group $\mathcal{I}(E)$. Since $\mathcal{T}(E)$ is transitive on the set of all points of E, by the Frattini Argument, we have

{Frat1}
$$\mathcal{I}(E) = \mathcal{I}(E)_O \mathcal{T}(E). \tag{3.31}$$

On the other hand given a point P different from O, the circle through P with center in O is invariant by the action of the stabiliser $\mathcal{I}(E)_O$) of the origin O in $\mathcal{I}(E)$ and $\mathcal{R}(E)$ is transitive on this circle. thus, again by the Frattini Argument,

{Frat2}
$$\mathcal{I}(E)_O = \mathcal{I}(E)_{O,P} \mathcal{R}(E) \tag{3.32}$$

But now one can see easily that $\mathcal{I}(E)_{O,P}$ leaves fixed pointwise the line l through O and P and that the only isometries that do so are the identity of E and the reflection with axis l, so

{Frat3}
$$\mathcal{I}(E)_{O,P} = \mathcal{S}_l(E) \tag{3.33}$$

Thus

$$\mathcal{I}(E) = \mathcal{S}_l(E)\mathcal{R}(E)\mathcal{T}(E)$$

in other words,

{isometries}

Theorem 3.3.6 *Every isometry of the Euclidean plane is the product of a reflection, a rotation, and a translation.*

3.3.5 The finite symmetric groups

In this subsection we shall show some basic properties of the symmetric group S_X of all permutations of a set X, not necessarily finite for the moment. The first remark is that if X and Y are two sets of the same cardinality (i.e. there is a bijection between X and Y), then the groups S_X and S_Y are isomorphic.

{isoperm1}

Lemma 3.3.7 *Let X and Y be sets and $f\colon X \to Y$ a bijection. Then, for every $\sigma \in S_X$ the map $f \circ \sigma \circ f^{-1}$ is a permutation of Y*

PROOF. Let $y \in Y$ then $f^{-1}(y) \in X$. Since $\sigma \in S_X$, $\sigma \circ f^{-1}(y) \in X$ whence $f \circ \sigma \circ f^{-1}(y) \in Y$. So $f \circ \sigma \circ f^{-1}$ is a map from Y to Y and it is bijective by Lemma 2.4.2, since it is the composition of bijective maps. Hence $f \circ \sigma \circ f^{-1}$ is a permutation of Y. ∎

If $f\colon X \to Y$ is a bijection between the sets X and Y, denote by γ_f the map

$$\begin{array}{rccc} \gamma_f\colon & S_X & \to & S_Y \\ & \sigma & \mapsto & f \circ \sigma \circ f^{-1} \end{array}$$

{isoperm2}

Theorem 3.3.8 *Let X and Y be sets and $f\colon X \to Y$ a bijection. Then the map γ_f is an isomorphism of groups whose inverse map γ_f^{-1} is $\gamma_{f^{-1}}$*

PROOF. By Lemma 3.3.7, γ_f is a map from S_X to S_Y. Assume σ and τ are permutations in S_X. Then, by Lemma 2.4.3, Lemma 2.4.4, and Lemma 2.4.5,

$$\begin{aligned} \gamma_f(\sigma \circ \tau) &= f \circ (\sigma \circ \tau) \circ f^{-1} = f \circ (\sigma \circ f^{-1} \circ f \circ \tau) \circ f^{-1} \\ &= (f \circ \sigma \circ f^{-1}) \circ (f \circ \tau \circ f^{-1}) = \gamma_f(\sigma) \circ \gamma_f(\tau). \end{aligned}$$

So γ_f is a homomorphism of groups. To prove that γ_f is bijective we use Lemma 2.4.6. Indeed, consider the map $\gamma_{f^{-1}}$ then, for every $\sigma \in S_X$, by Lemma 2.4.4, Lemma 2.4.5, and the associativity of the composition of functions (Lemma 2.4.3) we have

$$\begin{aligned} \gamma_{f^{-1}} \circ \gamma_f(\sigma) &= \gamma_{f^{-1}}(\gamma_f(\sigma)) = \gamma_{f^{-1}}(f \circ \sigma \circ f^{-1}) \\ &= f^{-1} \circ (f \circ \sigma \circ f^{-1}) \circ f = (f^{-1} \circ f) \circ \sigma \circ (f^{-1} \circ f) \\ &= id_X \circ \sigma \circ id_X = \sigma = id_{S_X}(\sigma) \end{aligned}$$

and, for every $\rho \in S_Y$,

$$\begin{aligned} \gamma_f \circ \gamma_{f^{-1}}(\rho) &= \gamma_f(\gamma_{f^{-1}}(\rho)) = \gamma_f(f^{-1} \circ \rho \circ f) \\ &= f \circ (f^{-1} \circ \rho \circ f) \circ f^{-1} = (f \circ f^{-1}) \circ \rho \circ (f \circ f^{-1}) \\ &= id_Y \circ \rho \circ id_Y = \rho = id_{S_Y}(\rho) \end{aligned}$$

and the result follows by Lemma 2.4.6. ∎

The importance of Theorem 3.3.8 is that, in order to identify the group S_X up to isomorphism, the only parameter we need to consider is the cardinality

$|X|$ of the set X. In particular, if $|X|$ is finite, say n, for $n \in \mathbb{N} \setminus \{0\}$, then (by definition of cardinality!) there is a bijection between X and the set $\{1, \ldots, n\}$ and so the groups S_X and $S_{\{1,\ldots,n\}}$ are isomorphic. Thus we can reduce ourselves to investigating the group $S_{\{1,\ldots,n\}}$. For brevity denote by S_n the group $S_{\{1,\ldots,n\}}$, for every $n \in \mathbb{N} \setminus \{0\}$.

The order of S_n

The first important information about a finite group is its cardinality, for example one can prove that a group of even order has an element of order 2 (see Exercise 3.4.13). In a slightly more difficult way one can prove that if the order of G is $p^t \cdot m$ for a prime p, and positive integers t and m, with m coprime to p, then G has subgroups of order p^t (Sylow's Theorem [**?**], a very important theorem for finite groups). Finally the celebrated Theorem of Feit and Thompson [**?**] (whose proof is extremely difficult) states that every finite group G of odd order is either abelian and its order is a prime number or has a proper (i.e. different from $\{1\}$ and G itself) normal subgrup.

In order to compute the order of S_n we need to define the *factorial* $n!$ of a non zero natural number n. This is just the product of the first positive integers up to n. So

$1! = 1,$

$2! = 2 \cdot 1 = 2,$

$3! = 3 \cdot 2 \cdot 1 = 6,$

$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24,$

and so on

Formally $n!$ is defined inductively setting

$1! := 1$ and,

inductively, for every $n \in \mathbb{N} \setminus \{0\}$, $(n+1)! := (n+1) \cdot (n!)$.

{cardSn}

Theorem 3.3.9 *Let $n \in \mathbb{N} \setminus \{0\}$, then*

$$|S_n| = n!.$$

Proof. We give a sketch of the proof: the idea is to count in how many different ways we can choose a permutation σ of the set $\{1, \ldots, n\}$. This is equivalent to count all the possibilities for

$$\sigma(1),\ \sigma(2), \ldots, \sigma(n).$$

We do this by first counting all the possible choices for $\sigma(1)$ then, once $\sigma(1)$ has been fixed, count all the possible choices for $\sigma(2)$ and so on up to $\sigma(n)$ Now for $\sigma(1)$ we can choose any element of $\{1, \ldots, n\}$, so we have

$$n \text{ choices for } \sigma(1).$$

Assume $\sigma(1)$ has been fixed, we can choose $\sigma(2)$ to be any element of $\{1, \ldots, n\}$ except for $\sigma(1)$ (otherwise $\sigma(1)$ would be equal to $\sigma(2)$ which is forbidden, since a permutation is bijective). So we have

$$n - 1 \text{ choices for } \sigma(2).$$

Again assume $\sigma(1)$ and $\sigma(2)$ have been fixed and look for the possible choices for $\sigma(3)$. As before we can chhose $\sigma(3)$ to be any element of $\{1, \ldots, n\}$ except for $\sigma(1)$ and $\sigma(2)$. So we have

$$n - 2 \text{ choices for } \sigma(3).$$

We go on this way up to $\sigma(n-1)$ and $\sigma(n)$. Assume $\sigma(1), \sigma(2), \ldots \sigma(n-2)$ have been fixed. Note that they are all distinct, so $\{\sigma(1), \sigma(2), \ldots \sigma(n-2)\}$ is a subset of $\{1, \ldots, n\}$ with $n-2$ elements. Thus we can choose $\sigma(n-1)$ to be any element of $\{1, \ldots, n\} \setminus \{\sigma(1), \sigma(2), \ldots \sigma(n-2)\}$ that is in a set of two elements. Thus we have

$$2 \text{ choices for } \sigma(n-1).$$

And finally

$$\text{just one choice for } \sigma(n),$$

namely the unique element of $\{1, \ldots, n\} \setminus \{\sigma(1), \sigma(2), \ldots \sigma(n-2), \sigma(n-1)\}$. Multiplying all the possible choices, we get that the possible choices for σ are

the possible choices for $\sigma(1)$

times

the possible choices for $\sigma(2)$

times

$\vdots$

the possible choices for $\sigma(n-1)$

times

the possible choices for $\sigma(n)$,

that is

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

■

3.3.6 A closer look at permutations

Now let σ be a permutation of a set X. Observe that if $|X| = 1$, S_X is the trivial group whose unique element is id_X and there's nothing much to say about it. Thus we shall assume in the sequel that X contains at least two elements. Denote by $Mov(\sigma)$ the set of elements of X that are *moved* by σ, that is the elements $x \in X$ such that $\sigma(x) \neq x$. Further denote by $Fix(\sigma)$ the set of the elements of X that are fixed by σ. Obviously an element of X can either be moved or left fixed by σ and no element that is moved by σ is left fixed by σ. Therefore

{movfix}

Lemma 3.3.10 *Let σ be a permutation of a set X. Then*

$$X = Mov(\sigma) \cup Fix(\sigma) \text{ and } Mov(\sigma) \cap Fix(\sigma) = \emptyset$$

{mov}

Lemma 3.3.11 *Let σ be a permutation of a set X. If $x \in Mov(\sigma)$ then also $\sigma(x)$ and $\sigma^{-1}(x)$ lie in $Mov(\sigma)$.*

PROOF. Assume, by means of contradiction, that $\sigma(x) \notin Mov(\sigma)$. Then $\sigma(x)$ is fixed by σ, hence

$$\sigma(\sigma(x)) = \sigma(x) = \sigma(\sigma(x),$$

that is x and $\sigma(x)$ have the same image. Since σ is a permutation, hence injective we have $x = \sigma^n(x)$, against the hypothesis that $x \in Mov(\sigma)$). The proof that $\sigma^{-1}(x) \in Mov(\sigma)$ is similar. ∎

{mov>1}

Corollary 3.3.12 *Let σ be a permutation of a set X. Then $|Mov(\sigma)| \geq 2$.*

PROOF. if $x \in Mov\sigma$ then, by Lemma 3.3.11, also $\sigma(x) \in Mov(\sigma)$, and $x \neq \sigma(x)$. ∎

Corollary 3.3.13 *Let σ be a permutation of a set X and $n \in \mathbb{Z}$. If $x \in Mov(\sigma)$ (resp. $x \in Fix(\sigma)$), then also $\sigma^n(x) \in Mov(\sigma)$ (resp. $\sigma^n(x) \in Fix(\sigma)$).*

PROOF. If $x \in Fix(\sigma)$ the result is obvious, since $\sigma^n(x) = x$ for every $n \in \mathbb{Z}$. If $x \in Mov(\sigma)$, the result follows by easy induction from Lemma 3.3.11. ∎

Obviously any two powers of permutation σ commute:

$$\sigma^n \circ \sigma^t = \sigma^{n+t} = \sigma^{t+n} = \sigma^t \circ \sigma^n.$$

However, in general, we cannot expect that two randomly chosen permutations commute (we have already seen this in S_3, The next lemma gives an important sufficient condition for two permutations to commute.

{movcomm}

Lemma 3.3.14 *Let σ and τ be permutations of a set X such that $Mov(\sigma) \cap Mov(\tau) = \emptyset$. Then*

$$\sigma \circ \tau = \tau \circ \sigma$$

PROOF. Since $Mov(\sigma) \cap Mov(\tau) = \emptyset$ every element of X that is movced by σ is left fixed by τ and every element that is moved by τ is left fixed by σ. We prove that, for every $x \in X$

$$\sigma \circ \tau(x) = \tau \circ \sigma(x).$$

We distinguish two cases:

Case 1: $x \in Mov(\sigma)$, so $\tau(x) = x$ and, by the above remark, $\tau(\sigma(x) = \sigma(x)$. Then

$$\sigma \circ \tau(x) = \sigma(\tau(x)) = \sigma(x) = \tau(\sigma(x) = \tau \circ \sigma(x).$$

Case 2: $x \notin Mov(\sigma)$. If also $x \notin Mov(\tau)$ then $x = \sigma(x) = \tau(x)$, whence

$$\sigma \circ \tau(x) = \sigma(\tau(x)) = \sigma(x) = x = \tau(x) = \tau(\sigma(x)) = \tau \circ \sigma(x).$$

If $x \in Mov(\tau)$, then proceed as in case 1 swapping σ and τ.

■

Two permutations σ and τ such that $Mov(\sigma) \cap Mov(\tau) = \emptyset$ are called disjoint.

Cycles

Let X be a set and n be an integer with $2 \leq n \leq |X|$. A permutation γ in S_X is called a *cycle of length* n if there is an element $x \in X$ such that

1. x, $\gamma(x)$, $\gamma^2(x)$, ..., $\gamma^{n-2}(x)$, and $\gamma^{n-1}(x)$ are all distinct and
2. $Mov(\gamma) = \{x, \gamma(x), \gamma^2(x), \ldots, \gamma^{n-2}(x), \gamma^{n-1}(x)\}$.

It is also convenient to define a *cycle of length* 1 to be the identity map on X (cfr. Corollary 3.3.12).

{cycles1}

Lemma 3.3.15 *Let n be an integer and γ a cycle of length n on a set X. Then γ has order n*

PROOF. If $n = 1$ γ is the identity map which has order 1 as an element of S_X. Assume $n \geq 2$ and let x be as in the above definition. Since the elements x, $\gamma(x)$, $\gamma^2(x)$, ..., $\gamma^{n-2}(x)$, $\gamma^{n-1}(x)$ are all distinct, $\gamma^t \neq id_X$ for every $t \in \{1, \ldots, n-1\}$, so γ has order at least n. Conversely, we prove that $\gamma^n = id_X$. By Lemma 3.3.11,

$$\gamma^n(x) \in Mov(\gamma) = \{x, \gamma(x), \gamma^2(x), \ldots, \gamma^{n-2}(x), \gamma^{n-1}(x)\},$$

so there exists $s \in \{0, \ldots, n-1\}$ such that

$$\gamma^n(x) = \gamma^s(x) \tag{3.34}$$

{gamman}

(note that $\gamma^0 = id_X$, so $x = \gamma^0(x)$). Since $s \in \{0, \ldots, n-1\}$, $n - s \in \{1, \ldots, n\}$. Thus, composing with γ^{-s}, both members of Equation 3.34 we get

$$\gamma^{n-s}(x) = \gamma^{-s} \circ \gamma^n(x) = \gamma^{-s} \circ \gamma^s(x) = x.$$

Again, since the elements Since the elements x, $\gamma(x)$, $\gamma^2(x)$, ..., $\gamma^{n-2}(x)$, $\gamma^{n-1}(x)$ are all distinct, and $n - s \in \{1, \ldots, n\}$, it follows that the unique possibility is that $s = 0$ and so $\gamma^n(x) = 0$. Now, for every $r \in \{0, \ldots, n-1\}$, we have

$$\gamma^n \circ \gamma^t(x) = \gamma^{n+t}(x) = \gamma^t \circ \gamma^n(x) = \gamma^t(x)$$

So, if $y \in Mov(\gamma)$, then $y \in Fix(\gamma^n)$ and, obviously, if $y \in Fix(\gamma)$ then $y \in Fix(\gamma)^n$, so, by Lemma 3.3.10, γ^n fixes all the elements of X, hence $\gamma^n = id_X$. ■

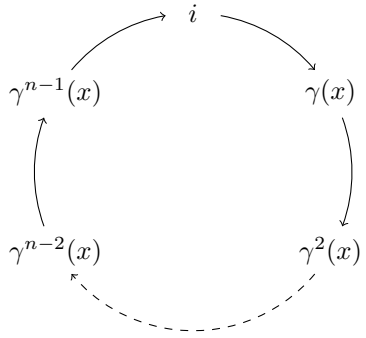
{cycles2}

Corollary 3.3.16 *Let n be a positive integer and let γ be a cycle of length n on a set X. Then*

$$\gamma^{kn+t}(y) = \gamma^t(y)$$

for all y in X and all $k, t \in \mathbb{Z}$.

Here's a way to visualize how a cycle γ of length n operates: draw the elements x, $\gamma(x)$, $\gamma^2(x)$, ..., $\gamma^{n-2}(x)$, $\gamma^{n-1}(x)$ as the edges of an n-gon as below. Then the cycle γ permutes the edges by a rotation of $360/n$ degrees.



This suggest a convenient way to denote a cycle: let γ be as above, then we can identify γ with the n-tuple

$$(x, \gamma(x), \gamma^2(x), \ldots, \gamma^{n-2}(x), \gamma^{n-1}(x)).$$

But be warned that this n-tuple is not unique, since for any $t \in \mathbb{Z}$ also the n-tuple

$$(\gamma^t(x), \gamma^{t+1}(x), \ldots, \gamma^{t+(n-2)}(x), \gamma^{t+(n-1)}(x).$$

also identifies γ.

{cycdec}

Lemma 3.3.17 *Let X be a finite set. Then every permutation of X different from the identity can be written uniquely as a product (composition) of disjoint cycles.*

PROOF. We will not give a formal proof, but just a practical method to detect the cycles. Of course we may assume $|X| > 1$, otherwise the result is obvious since there are no permutations different from the identity. Let σ be a permutation of X, and let $Mov(\sigma) = \{x_1, \ldots, x_n\}$. Start with x_1 and consider its orbit O_1 under $\langle\sigma\rangle$:

$$\{x_1, \sigma(x_1), \sigma^2(x_1), \ldots, \sigma^{t_1}(x_1)\}.$$

Set

$$\gamma_1 = (x_1, \sigma(x_1), \sigma^2(x_1), \ldots, \sigma^{t_1}(x_1)).$$

If $O_1 = Mov(x)$ $\sigma = \gamma_1$ is a cycle and we are done (see Exercise 3.4.14). Otherwise choose $y \in Mov(x) \setminus O_1$, let

$$O_2 := \{\{y, \sigma(y), \sigma^2(y), \ldots, \sigma^{t_2}(y)\}$$

be the orbit of y under $\langle\sigma\rangle$ and let γ_2 be the cycle

$$(y, \sigma(y), \sigma^2(y), \ldots, \sigma^{t_2}(y)).$$

Again either $Mov(\sigma) = O_1 \cup O_2$, or there is an element $w \in Mov(\sigma) \setminus (O_1 \cup O_2)$. In the latter case consider again the orbit O_3 of w under $\langle\sigma\rangle$ and the associated cycle γ_3 as above. Since X is finite, this procedure must exhaust all the elements of $Mov(x)$ after a finite number of steps, say r, so you eventually get r cycles $\gamma_1, \gamma_2, ..., \gamma_r$. Now it is immediate to see that

$$\sigma = \gamma_1 \circ \gamma_2 \circ \cdots \circ \gamma_r$$

for if $z \in X$ then either $z \in Fix(\sigma)$, so $z \in Fix(\gamma_i)$ for every $i \in \{1, \ldots, r\}$ or there is an orbit O_l such that $z \in O_l$. So $\sigma(z) = \gamma_l(z)$ and, fro every $j \in \{1, \ldots, r\} \setminus l$, γ_j fixes z and $\gamma_l(z)$. But then

$$\sigma(z) = \gamma_l(z) = \gamma_1 \circ \gamma_2 \circ \cdots \circ \gamma_l \circ \cdots \circ \gamma_r(z).$$

■

Example: Write the permutation

$$\sigma := \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 2 & 3 & 1 & 4 & 7 & 6 & 5 & 9 & 8 \end{pmatrix}$$

of the set $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ as a product of cycles. We have

$$Mov(\sigma) = \{1, 2, 3, 5, 7, 8, 9\}$$

(4 and 6 are left fixed). First consider the orbit O_1 of 1 under $\langle\sigma\rangle$:

$$\{1, \sigma(1), \sigma^2(1)\} = \{1, 2, 3\}$$

and let

$$\gamma_1 \text{ be the cycle } (1, 2, 3).$$

Now $Mov(\sigma) \setminus O_1 = \{5, 7, 8, 9\}$, so consider the orbit O_2 of 5 under $\langle\sigma\rangle$:

$$\{5, \sigma(5)\} = \{5, 7\}$$

and let

$$\gamma_2 \text{ be the cycle } (5, 7).$$

Again we have $Mov(\sigma) \setminus (O_1 \cup O_2) = \{8, 9\}$. Thus consider the orbit O_3 of 8 under $\langle\sigma\rangle$:

$$\{8, \sigma(8)\} = \{8, 9\}$$

and let

$$\gamma_3 \text{ be the cycle } (8, 9).$$

Now, since $Mov(\sigma) = O_1 \cup O_2 \cup O_3$, we are finished and

{3cycles}
$$\sigma = \gamma_1 \circ \gamma_2 \circ \gamma_3 \tag{3.35}$$

(we leave the reader to check directly this equality).

Note that, since disjoint permutations commute with each other, we do not need to care in which order the cycles γ_1, ..., γ^r are composed: Equation 3.35 could be equivalently be written as

$$\sigma = \gamma_2 \circ \gamma_3 \circ \gamma_1$$

or

$$\sigma = \gamma_3 \circ \gamma_2 \circ \gamma_1.$$

We shall denote the length of a cycle γ by $l(\gamma)$.

Let

$$\sigma = \gamma_1 \circ \gamma_2 \circ \cdots \circ \gamma_t$$

be a decomposition of a permutation σ as a product of disjoint cycles γ_1, γ_2 $ldots\gamma_t$. By the above remark, we may choose the indices $\{1, \ldots, t\}$ in such a way that

$$l(\gamma_1) \leq l(\gamma_2) \leq \cdots \leq l(\gamma_t).$$

The $t - uple$

$$(l(\gamma_1), l(\gamma_2), \cdots, l(\gamma_t)$$

is called the *type* of σ.

For example, we have seen that the permutation

$$\sigma := \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 2 & 3 & 1 & 4 & 7 & 6 & 5 & 9 & 8 \end{pmatrix}$$

can be written as the product of the disjoint cycles

$$(5,7),/,/,/,(8,9), \text{ and } (1,2,3).$$

whose lengths are respectively

$$2,\ 2,\ \text{and } 3.$$

Therefore the type of σ is the triple

$$(2,2,3).$$

{cyccon}

Lemma 3.3.18 *Let n be an integer greater than 2 and X be a finite set with $n \leq |X|$. Let*

$$\gamma := (x_1, x_2, \ldots, x_n)$$

be a cycle of length n in S_X. Then, for every permutation σ in S_X,

$$\sigma \circ \gamma \circ \sigma^{-1} = (\sigma(x_1), \sigma(x_2), \ldots, \sigma(x_n)).$$

In particular $\sigma \circ \gamma \circ \sigma^{-1}$ is also a cycle of length n.

PROOF. If $y \notin \{\sigma(x_1), \ldots, \sigma(x_n)\}$, then $\sigma^{-1}(y) \notin \{x_1, \ldots, x_n\}$. Thus

$$Mov(\sigma \circ \gamma \circ \sigma^{-1}) \subseteq \{\sigma(x_1), \sigma(x_2), \ldots, \sigma(x_n)\}.$$

On the other hand, since $\sigma^{-1} \circ \sigma = id_X$, for each $i \in \{1, \ldots, n-1\}$, we have

$$\sigma \circ \gamma \circ \sigma^{-1}(\sigma(x_i)) = \sigma \circ \gamma \circ \sigma^{-1} \circ \sigma(x_i) = \sigma \circ \gamma \circ (x_i) = \sigma(x_{i+1})$$

and

$$\sigma \circ \gamma \circ \sigma^{-1}(\sigma(x_n)) = \sigma \circ \gamma \circ \sigma^{-1} \circ \sigma(x_n) = \sigma \circ \gamma \circ (x_n) = \sigma(x_1),$$

proving the assertion. ∎

Corollary 3.3.19 *Let ϕ and ψ be two permutations of the finite set X. Then there exists a permutation $\sigma \in S_X$ such that $\phi = \sigma \circ \psi \circ \sigma^{-1}$ if and only if ϕ and ψ have the same type.*

PROOF. By Lemma refconjugation, the conjugation γ_σ by σ is an automorphism of S_X and, by Lemma 3.3.18 γ_σ preserves the length of the cycles, so ϕ and ψ have the same type by Lemma 3.3.17, since an automorphism of S_X sends disjoint permutations to disjoint permutations. Conversely, assume ϕ and ψ have the same type, say $(t_1, \ldots, t_r)$. Let

$$\phi := (x_{1,1}, \ldots, x_{1,t_1}) \circ (x_{2,1}, \ldots, x_{2,t_2}), \circ \cdots \circ (x_{r,1}, \ldots, x_{r,t_r})$$

and

$$\psi := (y_{1,1}, \ldots, y_{1,t_1}) \circ (y_{2,1}, \ldots, y_{2,t_2}), \circ \cdots \circ (y_{r,1}, \ldots, y_{r,t_r})$$

be the respective decompositions as a product of disjoint cycles. Since the cycles are disjoint, the $x_{i,j}$'s are all different and the same holds for the $y_{i,j}$'s. Therefore there is a permutation σ in S_X such that

$$\sigma(x_{i,j}) = y_{i,j} \text{ for every } j \in \{1, \dots, r\} \text{ and every } i \in \{1, \dots, t_j\}.$$

In other words, the restriction of σ to $Mov(\sigma)$ is the permutation

$$\begin{pmatrix} x_{1,1} & \dots & x_{1,t_1} & x_{2,1} & \dots & x_{2,t_2} & \cdots & x_{r,1} & \dots & x_{r,t_r} \\ y_{1,1} & \dots & y_{1,t_1} & y_{2,1} & \dots & y_{2,t_2} & \cdots & y_{r,1} & \dots & y_{r,t_r} \end{pmatrix}.$$

Then, by Lemma 3.3.18 and Lemma 3.2.22, we have $\sigma\phi\sigma^{-1} = \psi$. ∎

Transpositions

Cycles of length 2 are called *transpositions*. They play a fundamental role when investigating the structure of a symmetric group. One reason is that every permutation can be written as a product of transpositions. This is an immediate consequence of the following

{cyctrans}

Lemma 3.3.20 *Let X be a finite set containing at least two elements, n be a positive integer (with $2 \leq n \leq |X|$), and γ be a cycle of length n in S_X. Then γ is a product of $n-1$ transpositions.*

PROOF. Let $\gamma := (x_0, x_1, x_2, \dots, x_{n-1})$ Then a direct computation shows that

$$\gamma = (x_0, x_1) \circ (x_1, x_2) \circ \cdots \circ (x_{n-3}, x_{n-2}) \circ (x_{n-2}, x_{n-1})$$

(simply compare the images of $x_0, \dots x_{n-1}$ under γ and under $(x_0, x_1) \circ (x_1, x_2) \circ \cdots \circ (x_{n-3}, x_{n-2}) \circ (x_{n-2}, x_{n-1})$.) ∎

{transgen}

Corollary 3.3.21 *For every finite set X with $|X| > 1$, the symmetric group S_X is generated by transpositions.*

PROOF. Let σ be a permutation in S_X. If $\sigma = id_X$ then, for every transposition τ in S_X, $\sigma = \tau \circ \tau$. If $\sigma \neq id_x$, then, by Lemma 3.3.17, σ is a product of cycles and, by Lemma 3.3.21, every cycle of X is a product of transpositions. In both cases σ is a product of transpositions. ∎

The sign of a permutation

It is important to point out that, as a difference to the decomposition of a permutation as a product of disjoint cycles, **the decomposition of a permutation as a product of transpositions is not unique**.

For example, if $X = \{1, \dots, n\}$

$$(1,2) = (1,2) \circ (1,2) \circ (1,2) = (1,2) \circ (1,3) \circ (1,3)$$

On the other hand, there is still a feature that is invariant in all decompositions, namely:

{sign}

Theorem 3.3.22 *Let X be a finite set and σ be a permutation of X. If σ is the product of an even number of transpositions, then every decomposition of σ as a product of transpositions contains an even number of factors.*

In order to prove this theorem we first define a map, called *sign*,

$$sgn\colon S_X \to \{1, -1\}$$

in the following way: If σ is a permutation of type $(t_1, \ldots, t_r)$ we set

$$sgn(\sigma) := (-1)^{(t_1+t_2+\cdots+t_r)-r}$$

Let us first understand what this means: If γ is a cycle of length n then γ has type (n) whence $r = 1$ and

$$sgn(\gamma) = (-1)^{n-1}.$$

That is **cycles of odd length have sign equal to** 1 **and cycles of even length have sign equal to** -1. **In particular, all transpositions have sign** -1.

Let σ be as above. Since

$$(t_1 + t_2 + \cdots + t_r) - r = (t_1 - 1) + (t_2 - 1) + \cdots + (t_r - 1)$$

we have

$$\begin{aligned}(-1)^{(t_1+t_2+\cdots+t_r)-r} &= (-1)^{(t_1-1)+(t_2-1)+\cdots+(t_r-1)} \\ &= (-1)^{t_1-1} \cdot (-1)^{t_2-1} \cdot \cdots \cdot (-1)^{t_r-1}\end{aligned}$$

the latter is precisely the product of the signs of the cycles appearing in the decomposition of σ as a product of disjoint cycles. So **the sign of a permutation σ is the product of the signs of the cycles appearing in its decomposition as a product of disjoint cycles**. As a consequence we get immediately the following

{disperm}

Lemma 3.3.23 *Let X be a finite set and let ρ and σ be two disjoint permutations of X. Then*

$$sgn(\rho \circ \sigma) = sign(\rho) \cdot sgn(\sigma).$$

Proof. Let

$$\rho := \gamma_1 \circ \gamma_2 \circ \cdots \circ \gamma_r$$

and

$$\sigma := \delta_1 \circ \delta_2 \circ \cdots \circ \delta_s$$

be the decompositions of ρ and σ as a product of disjoint cycles. Since ρ and σ are disjoint,

$$\rho \circ \sigma = \gamma_1 \circ \gamma_2 \circ \cdots \circ \gamma_r \circ \delta_1 \circ \delta_2 \circ \cdots \circ \delta_s$$

is the decomposition of $\rho \circ \sigma$ as a product of disjoint cycles, hence the result follows by the definition of the function sgn. ∎

When we compose a transposition τ with a permutation σ the sign of σ changes. To prove this we shall follow [6, pag49-50] and distinguish three cases, depending whether the intersection $Mov(\tau) \cap Mov(\sigma)$ has order 0, 1, or 2. The first case should be obvious:

{case0}

Lemma 3.3.24 *Let σ be a permutation on a set X and τ be a transposition such that $Mov(\tau) \cap Mov(\sigma) = \emptyset$. Then*

$$sgn(\tau \circ \sigma) = -sgn(\sigma).$$

PROOF. Let $\gamma_1, \ldots, \gamma_r$ be the disjoint cycles such that

$$\sigma = \gamma_1 \circ \cdots \circ \gamma_r.$$

Since $Mov(\tau) \cap Mov(\sigma) = \emptyset$, $\tau, \gamma_1, \ldots, \gamma_r$ are also disjoint cycles and obviously

$$\tau \circ \sigma = \tau \circ \gamma_1 \circ \cdots \circ \gamma_r.$$

Since $sgn(\tau) = -1$, it follows that

$$sgn(\tau \circ \sigma) = sgn(\tau) \cdot sgn(\gamma_1) \cdot \cdots \cdot sgn(\gamma_r) = sgn(\tau) \cdot sgn(\sigma) = -sgn(\sigma).$$

∎

{basform1}

Lemma 3.3.25 *Let X be a set, n an integer with $2 \geq n \geq |X|$, $y, x_0, x_1, \ldots x_n$ distinct elements of X. Then*

$$(x_0, y) \circ (x_0, \ldots, x_n) = (x_0, x_1, \cdots, x_n, y)$$

PROOF. Just compute the the images of $y, x_0, \ldots, x_n$ on both sides of the above equation. ∎

{case1}

Corollary 3.3.26 *Let σ be a permutation on a set X and τ be a transposition such that $|Mov(\tau) \cap Mov(\sigma)| = 1$. Then*

$$sgn(\tau \circ \sigma) = -sgn(\sigma).$$

PROOF. Let $\tau := (a, b)$ and $\gamma_1, \ldots, \gamma_r$ be the disjoint cycles such that

$$\sigma = \gamma_1 \circ \cdots \circ \gamma_r.$$

Since $|Mov(\tau) \cap Mov(\sigma)| = 1$ there is a unique cycle $\gamma := (x_0, \dots, x_n)$ in the decomposition of σ such that $|Mov(\tau) \cap Mov(\gamma)| = 1$. Since disjoint cycles commute, we may choose the indices $1, \dots, r$ in such a way that $\gamma = \gamma_1$. Thus τ is disjoint with $\gamma_2, \dots, \gamma_r$. By Lemma 3.3.25 we have

{cluc}
$$sgn(\tau \circ \gamma) = (-1)^{n+1} = -(-1)^n = -sgn(\gamma). \tag{3.36}$$

Since $(\tau \circ \gamma_1)$ and $(\gamma_2 \circ \cdots \circ \gamma_r)$ are disjoint permutations, by Lemma 3.3.23 and Equation 3.36, we have

$$\begin{aligned} sgn(\tau \circ \sigma) &= sgn(\tau \circ \gamma_1 \circ \gamma_2 \circ \cdots \circ \gamma_r) \\ &= sgn((\tau \circ \gamma_1) \circ (\gamma_2 \circ \cdots \circ \gamma_r)) \\ &= sgn(\tau \circ \gamma_1) \cdot sgn(\gamma_2) \cdot \cdots \cdot sgn(\gamma_r) \\ &= -sgn(\gamma_1) \cdot sgn(\gamma_2) \cdot \cdots \cdot sgn(\gamma_r) \\ &= -sgn(\sigma). \end{aligned}$$

■

{basform}

Lemma 3.3.27 *Let X be a set, n an integer with $2 \geq n \geq |X|$, $x_0, x_1, \dots x_n$ distinct elements of X and $i \in \{1, \dots, n\}$. Then*

$$(x_0, x_i) \circ (x_0, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) = (x_i, x_{i+1}, \dots, x_n) \circ (x_0, x_1, \cdots, x_{i-1}) \tag{3.37}$$

{equno1}

and

$$(x_0, x_i) \circ (x_i, x_{i+1}, \dots, x_n) \circ (x_0, x_1, \cdots, x_{i-1}) = (x_0, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n). \tag{3.38}$$

{equdue}

In particular, multiplication by the ransposition (x_1, x_i) changes the sign of the permutations $(x_0, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n)$, resp. $(x_i, x_{i+1}, \dots, x_n) \circ (x_0, x_1, \cdots, x_{i-1})$.

PROOF. As above, just compute the the images of $x_0, \dots, x_n$ on both sides of the above equations (note also that, since $(x_0, x_1) = (x_0, x_1)^{-1}$, Equation 3.38 is obtained from Equation 3.37 multiplying both sides on the left by the transposition (x_0, x_i)). The final remark follows by the definition of the function sign. ■

{case2}

Corollary 3.3.28 *Let σ be a permutation on a set X and τ be a transposition such that $Mov(\tau) \subseteq Mov(\sigma)$. Then*

$$sgn(\tau \circ \sigma) = -sgn(\sigma).$$

PROOF. Let $\tau := (a, b)$ and let

$$\sigma = (x_{1,1}, \dots, x_{1,t_1}) \circ (x_{2,1}, \dots, x_{2,t_2}), \circ \cdots \circ (x_{r,1}, \dots, x_{r,t_r})$$

be the decomposition of σ as a product of disjoint cycles and set

$$\gamma_i := (x_{i,1}, \ldots, x_{i,t_i}).$$

We distinguish two cases: first assume there is a cycle γ in the decomposition of σ such that $\{a, b\} \subseteq Mov(\gamma)$. Since disjoint cycles commute with each other, we may choose the indices $1, \ldots, r$ in such a way that $\gamma = \gamma_1$ hence we may assume that $a = x_{1,1}$ and $b = x_{1,j}$ for a proper $j \in \{1, \ldots, t_1\}$. By Lemma 3.3.27 it follows that

{clac}
$$sign(\tau \circ \gamma_1) = (-1)^{n-1} = -(-1)^n = -sign(\gamma_1). \tag{3.39}$$

Since $(\tau \circ \gamma_1)$ and $(\gamma_2 \circ \cdots \circ \gamma_r)$ are disjoint permutations, by Lemma 3.3.23 and Equation 3.40, we have

$$\begin{aligned} sign(\tau \circ \sigma) &= sign(\tau \circ \gamma_1 \circ \gamma_2 \circ \cdots \circ \gamma_r) \\ &= sign((\tau \circ \gamma_1) \circ (\gamma_2 \circ \cdots \circ \gamma_r)) \\ &= sign(\tau \circ \gamma_1) \cdot sgn(\gamma_2 \circ \cdots \circ \gamma_r)) \\ &= -sign(\gamma_1) \cdot sgn(\gamma_2) \cdot \cdots \cdot sgn(\gamma_r) \\ &= -sign(\sigma) \end{aligned}$$

Next assume there are two distinct cycles γ and δ in the decomposition of σ such that $a \in Mov(\gamma)$ and $b \in Mov(\delta)$. Since disjoint cycles commute with each other, we may choose the indices $1, \ldots, r$ in such a way that $\gamma = \gamma_1$ and $\delta = \gamma_2$, hence we may also assume that $a = x_{1,1}$ and $b = x_{2,1}$. By Equation 3.38 in Lemma 3.3.27 it follows that

{clac}
$$sign(\tau \circ \gamma_1\gamma_2) = -sign(\gamma_1 \circ \gamma_2). \tag{3.40}$$

Since $(\tau \circ \gamma_1 \circ \gamma_2)$ and $(\gamma_3 \circ \cdots \circ \gamma_r)$ are disjoint permutations, by Lemma 3.3.23 and Equation 3.40, we have

$$\begin{aligned} sign(\tau \circ \sigma) &= sign(\tau \circ \gamma_1 \circ \gamma_2 \circ \gamma_3 \cdots \circ \gamma_r) \\ &= sign((\tau \circ \gamma_1 \circ \gamma_2) \circ (\gamma_3) \cdots \circ \gamma_r)) \\ &= sign(\tau \circ \gamma_1 \circ \gamma_2) \cdot sgn(\gamma_3 \circ \cdots \circ \gamma_r)) \\ &= -sign(\gamma_1 \circ \gamma_2) \cdot sgn(\gamma_3) \cdot \cdots \cdot sgn(\gamma_r) \\ &= -sign(\sigma) \end{aligned}$$

■

The next theorem gathers all the results we just proved and gives some consequences.

{sign}
Theorem 3.3.29 *Let X be a finite set with $|X| \geq 2$.*

1. *If τ is a transposition and σ is a permutation in S_X, then*

$$sgn(\tau \circ \sigma) = -sgn(\sigma).$$

2. *If a permutation σ in S_X can be written as a product of an even (resp. odd) number of transpositions, then the number of transpositions appearing in any other decomposition of σ as a product of transpositions is even (resp. odd).*

3. *If ρ and σ are premutations in S_x, then*

$$sgn(\rho \circ \sigma) = sgn(\rho) \cdot sgn(\sigma)$$

4. *In particular the map sgn is a homomorphism of groups from S_X to the multiplicative subgroup $\{1, -1\}$ of $(\mathbb{Q} \setminus \{0\}, \cdot)$.*

5. *The kernel of sgn is the set of all permutations that can be written as a product of an even number of permutations. In particular, this set is a normal subgroup of index 2 of S_X.*

PROOF. The first assertion follows from Lemma 3.3.24, Corollary 3.3.26, and Corollary 3.3.28. This obviously implies that the sign of a product of an even (resp. odd) number of transpositions is 1 (resp. -1), giving the second and third assertions. The fourth assertion is a reformulation of the third . Since the permutations that are products of an even number of transpositions have sign 1 it follows that $\ker(sgn)$ is the set of these permutations. Being the kernel of a homomorphism $\ker(sgn)$ is also a normal subgroup of S_X. Let γ be a transposition and σ a permutation in S_X, then either σ is a product of an even number of transpositions (that is $\sigma \in \ker(sgn)$), or σ is a product of an odd number of transpositions. In the latter case $\sigma \circ \tau$ is a product of an even number of transpositions, that is σ lies in the coset $\ker(sgn)\tau$ of $\ker(sgn)$. So there are precisely two cosets of $\ker(sgn)$ in S_X, that is $\ker(sgn)$ and $\ker(sgn)\tau$, whence $\ker(sgn)$ has index 2 in S_X, proving the last assertion. ∎

With the notation of Theorem 3.3.29, the kernel of *sgn* is denoted by A_X (or A_n if $X = \{1, \ldots, n\}$) and is called *alternating group of X*. By Lagrange's Theorem (Corollary 3.2.11),

$$|A_n| = \frac{n!}{2}.$$

A consequence of this fact is that, if we swap two tiles in the 15-Puzzle, it is impossible to return to the original position: e.g. if the tiles are as in Figure 3.2.1, it is impossible to get the configuration

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

It can be proved that, if $|X| \geq 5$ the group A_5 is simple and non abelian. A consequence, whose proof , if $|X| \geq 5$, the group A_X is a non abelian simple group. A consequence of this fact is the impossibility of solving by radicals

certain polynomial equations of degree greater than 5. The proof of this fact is beyond the scope of this book, the interested reader can look at Chapter 4 of [6]. We just try to give a sketch of the problem.

We all know how to solve polyinomial equations of degree 1 (linear equations): if

{gal1}
$$ax + b = 0 \tag{3.41}$$

is a linear equation with indeterminate x and $a, b \in \mathbb{R}$, with $a \not= 0$, then the solution is

$$\frac{-b}{a},$$

Meaning that substitutig x with $\frac{-b}{a}$, Equation 3.41 becomes an identity. We should also have learned at school how to solve an equation of second degree, Namely, given a polynomial equation

{gal2}
$$ax^2 + bx + c = 0 \tag{3.42}$$

of degree 2 (i.e. $a \neq 0$) the possible solutions ξ_i, $i \in \{1, 2\}$ are given by the well known formula

$$\xi_i := \frac{-b + (-1)^i \sqrt{b^2 - 4ac}}{2a}.$$

This means that if we have a programmable calculator that can do the four operations $+, -, \cdot, :$ and the extraction of square roots, we can compile a program that produces the solutions ξ_1 and ξ_2 of Equation 3.42, once it is given as input its coefficients a, b, c. Similar, but more complicated formulas were discovered during the Reneissance by Scipione del Ferro (1465-1526), Niccol Tartaglia (1499-1557), and Niccolò Cardano (1501-1576) also for equations of third, resp. fourth, degree:

$$ax^3 + bx^2 + cx + d = 0, \text{ resp. } ax^4 + bx^3 + cx^2 + dx + e = 0$$

on the indeterminate x with coefficients a, b, c, d, e in $\mathbb{R}$. But when trying to attack the equations of fifth degree, all previous methods failed. So that Lagrange first conjectured the impossibility of such a formula. This problem was eventually solved by Paolo Ruffini (1765-1822) (with some gaps), by Niels Henrik Abel, and, in full generality and with a new dramatic proof, by Évariste Galois. Essentially Galois proof uses certain symmetries of polynomials (yes also polynomials can have simmetries!) and shows that the solutions of a polynomial equation of degree n can be expressed from its coefficients by a formula that involves only the four operations and the extraction up to the nth root if and only if its symmetry group does not contain any subgroup admitting a factor which is a non abelian simple group. In particular he showed that there are certain polynomial equations of degree n greater or equal than 5 whose symmetry groups are isomorphic to the whole group S_n which has A_n as a non abelian simple subgroup, hence for these polynomials no formula exists.

3.4 Exercises

{norm1}

Exercise 3.4.1 *Let $f\colon A \to B$ be a homomorphism between the groups A and B. If H is a normal subgroup of A, then $f(H)$ is a normal subgroup of $f(A)$.*

{norm2}

Exercise 3.4.2 *Let $f\colon A \to B$ be a homomorphism between the groups A and B. If K is a normal subgroup of B, then $f^{-1}(K)$ is a normal subgroup of A.*

{kern}

Exercise 3.4.3 *Prove that the kernel of a homomorphism between groups is a normal subgroup of the domain*

{innorm}

Exercise 3.4.4 *Let G be a group, prove that $Inn(G)$ is a normal subgroup of $Aut(G)$.*

{Cayley}

Exercise 3.4.5 (Cayley's Theorem) *Let $(G, \cdot)$ be a group.*

1. *Prove that, for every element g of G, the map*

$$\begin{array}{rccc} \mu_g\colon & G & \to & G \\ & a & \mapsto & g \cdot a \end{array}$$

   is a permutation of the set G (hint: the quickest way is to prove that $\mu_{g^{-1}}$ is the inverse map of μ_g).

2. *Prove that the map*

$$\begin{array}{rccc} \mu\colon & G & \to & S_G \\ & g & \mapsto & \mu_g \end{array}$$

   is a homomorphism of groups between $(G, \cdot)$ and $(S_G, *)$.

3. *prove that μ is injective, hence G is isomorphic to the subgroup $\mu(G)$ of S_G, which is a permutation group on G (as a set).*

{S_X1}

Exercise 3.4.6 *Let X be a set $x \in X$, and $G = S_X$. Prove that $G_x = S_Y$ where $Y := X \setminus \{x\}$.*

{football}

Exercise 3.4.7 *Compute the order of the symmetry group of a soccer ball (Telstar type).*

{octahedron}

Exercise 3.4.8 *Compute the order of the symmetry group of a octahedron (if you are smart, you can get it from the symmetry group of a cube)*

{dodecahedron}

Exercise 3.4.9 *Compute the order of the symmetry group of a dodecahedron.*

{icosahedron}

Exercise 3.4.10 *Compute the order of the symmetry group of a icosahedron (if you are smart, you can get it from the symmetry group of a dodecahedron) .*

{translation}

Exercise 3.4.11 *Prove that $\mathcal{T}(E)$ is a normal subgroup of $\mathcal{I}(E)$.*

{translation}

Exercise 3.4.12 *Prove that $\mathcal{R}(E)$ is a normal subgroup of $\mathcal{I}(E)_O$.*

{evenorder}

Exercise 3.4.13 *Prove that is G is a finite group of even order, then G has elements of order 2.*
(Hint, first note that an element g of G has order 2 if and only if $g \neq 1_G$ and $g = g^{-1}$. So g has order 2 if and only if $g \neq 1$ and the set $\{g, g^{-1}\}$ contains only one element. Now consider all the sets $\{g, g^{-1}\}$ where g ranges through all the elements of G. Clearly G is the union of these sets and if $h, g \in G$ with $\{h, h^{-1}\} \cap \{g, g^{-1}\} \neq \emptyset$ then (prove) that $\{h, h^{-1}\} = \{g, g^{-1}\}$, so G is the disjoint union of these sets whence $|G|$ is the sum of the orders of all these sets. Now there is at least one of these sets that contains only one element, namely the set $\{1, 1\}$. If there were no element of order 2, then all the other sets would contain two elements, and so the order of G would be odd, against the hypothesis)

{orbcyc}

Exercise 3.4.14 *Let n be a positive integer. Prove that a permutation σ on a finite set X is a cycle of length n if and only if $\langle\sigma\rangle$ has a unique orbit of length n.*

{cycdecex}

Exercise 3.4.15 *Give a formal proof of Lemma 3.3.17 (Hint: use induction on $|Mov(\sigma)|$).*

{3trans}

Exercise 3.4.16 *Let σ and τ be two transpositions on a set X. prove that $\sigma \circ \tau$ is either the identity, a permutation of type(2,2), or a cycle of length 3.*

3.5 Not enough?

The best introductions on finite groups are [11], [2], and [9]. The first one brings you smoothly to the deeper concepts of the theory, the last two ones are pretty dense.

Chapter 4

Point and Line to Hyperplane

{Kandinsk}

The geometric point is an invisible thing. Therefore, it must be defined as an incorporeal thing. Considered in terms of substance, it equals zero.

Hidden in this zero, however, are various attributes which are "human" in nature. We think of this zero-the geometric point-in relation to the greatest possible brevity, i.e., to the highest degree of restraint which, nevertheless, speaks.

Thus we look upon the geometric point as the ultimate and most singular **union of silence and speech**.

The geometric point has, therefore, been given its material form, in the first instance, in writing. It belongs to language and signifies silence.

In the flow of speech, the point symbolizes interruption, non-existence (negative element), and at the same time it forms a bridge from one existence to another (positive element). In writing, this constitutes its **inner significance**

Externally, it is merely a sign serving a useful end and carries with it the element of the "practical useful", with which we have been acquainted since childhood. The external sign becomes a thing of habit and veils the inner sound of the symbol.

The inner becomes walled-up through the outer.

The point belongs to the more confined circle of habitual everyday phenomena with its traditional sound, which is mute.

The sound of that silence customarily connected with the point is so emphatic that it overshadows the other characteristics.

> All appearances that are traditionally familiar because of their singular expression, become mute to us. We no longer react to their appeal and are surrounded by silence; so we succumb to the deadly grip of "practical efficiency"
>
> *Wassily Kandinsky* [7]

4.1 Introduction

In a 19th century textbook of mathematics for the officers of the (at that time) Royal Italian Army, I found the following "definition" of a point:

"A point is a place in the space without any extension."

Unfortunately this is not a definition but, as in case of sets, just a description in vague terms of our idea of what a point should be. But, having already sets at hand, there's no need to introduce the concept of a point as another *primitive* one, since we can use the axioms of set theory to define the concept of point. We remark, however, that we will not give an *onthological* definition of a point but rather a *practical one*, in the sense that we will not try to say *what* a point is but *how* it behaves and, as one might expect, its behaviour will depend from the contex. Probably the simplest example comes from the *projective plane*:

The Projective Plane

A *projective plane* Π is a triple $(\mathcal{P}, \mathcal{L}, \iota)$ where $\mathcal{P}$ and $\mathcal{L}$ are disjoint non empty sets (the set of *points* and the set of *lines* respectively) and ι is a correspondence between $\mathcal{P}$ and $\mathcal{L}$ (called *incidence*) such that the following hold:

- $\Pi.1$ Given two distinct points P and Q (i.e. elements of the set $\mathcal{P}$) there exists a unique line l (i.e. a unque element of the set $\mathcal{L}$) such that both P and Q are incident with l (i.e. such that $P \iota l$ and $Q \iota l$).
- $\Pi.2$ Given two distinct lines l and r there exists a unique point P such that both l and r are incident with P (i.e. such that $P \iota l$ and $P \iota r$).
- $\Pi.3$ (nondegeneracy axiom) There are four points such that no line is incident with more than two of them.

Note that, as a difference to the usual *affine*[1], in a projective plane, two distinct lines are never parallel. Indeed we shall see that, in a certain sense, a projective plane can be obtained from a Euclidean plane by adjoining a further line, called the *horizon* whose points are the directions (so that two parallel lines intersect in their directions). If you do not understand this, please have patience, I'll make this clear soon.

[1] If the word *affine* disturbs you just substitute the affine plane with the Euclidean plane you studied at school

Also, note again that this definition just tells how points and lines *relate* to each other, but does not bother about what they *are*: indeed we shall see that objects of completely different nature can behave as points and lines in a projective plane. Just consider the following examples:

Example: constructing a projective plane out of the three dimensional Euclidean space

Projective planes, or, more generally, projective geometry, take their origin in the studies of perspective in painting. Remarkably those studies (the *Flagellation of Christ* of Piero della Francesca being one of the most emblematic examples) date much before its formalisation in mathematics. So perspective could actually be considered as a contribution of art to mathematics.

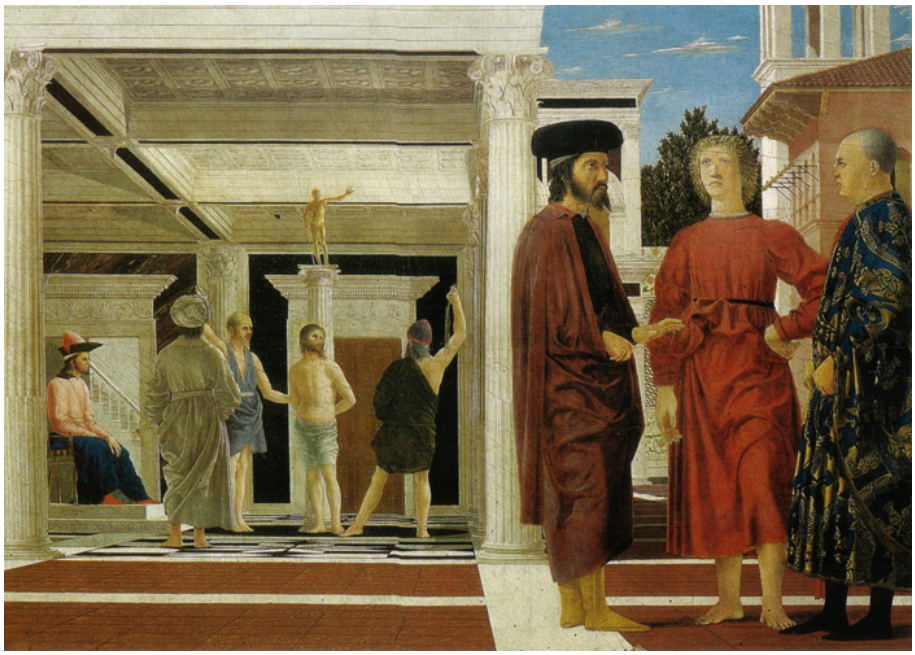
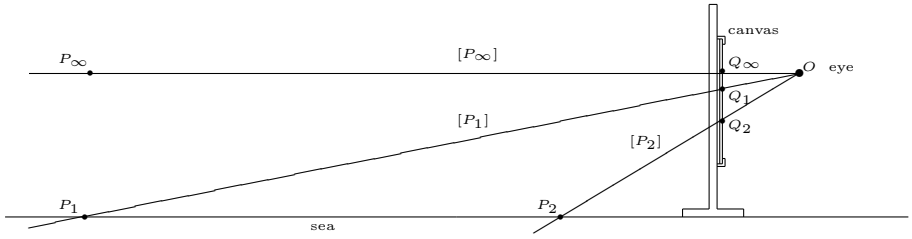


Figure 4.1: Piero della Francesca *The Flagellation of Christ* c. 1455.

Imagine a painter portraying a marine landscape. Let O be the eye of the painter as in the picture below, and π is the plane where the canvas is lying. For safety we also assume that the eye of the painter is above the sea level (otherwise the painter won't have much to live). Then, for each point P in the area portrayed of the sea, with P there is a unique point in the canvas Q which is in the intersection of the line $[P]$ through P and O with the canvas and, for each line l in the sea, there is a unique plane containing l and O and

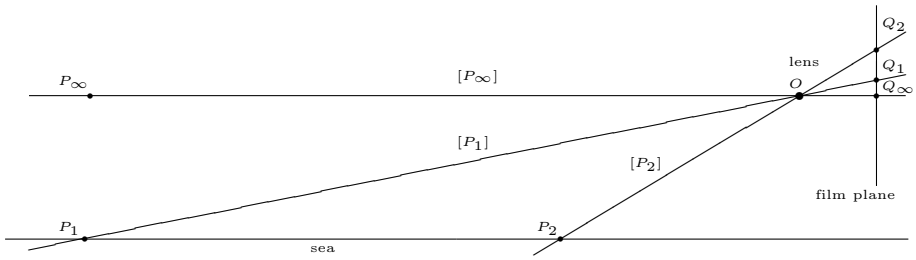
this plane intersects the canvas in a unique line. The horizon (the infinite line) in the painting is exactly the intersection of the canvas with the plane through O which is parallel to the sea. The images in the canvas of any two parallel lines in the sea intersect in the canvas at a point in the horizon, as would be the rails of a railway. The picture below shows how the three points P_1, P_2 and P_∞ (at the horizon) are reproduced respectively to the points Q_1, Q_2 and Q_∞ which are, respectively, the intersections of the lines $[P_1]$, $[P_2]$ and $[P_\infty]$ with the canvas plane.



Note that two distinct parallel lines r and s are mapped to (painted as) parallel lines $p(r)$ and $p(s)$ in the canvas if and only if r and s lie in a plane parallel to the canvas plane. Otherwise $p(r)$ and $p(s)$ will converge at a point: think about one of the thousands photo's of rails (here's one taken by Ansel Adams):

In reality the two rails are parallel, but the two lines that reproduce them in the picture definitely converge at a point.

Incidentally, this fact has to be taken care in architectural photography: indeed the situation is similar to the previous one, the only difference being that, in this case, the film (now sensor) plane is behind the lens O:



Now, if the film plane is not parallel to the façade of a building (for example if you tilt the camera upwards to take a picture of a tall building) parallel

Figure 4.2: Ansel Adams [1, p.161]

vertical lines are reproduced by convergent lines in the picture. This is usually considered disturbing in the reproduction of a building and that's why cameras specialized for architectural purposes (e.g the it Hasselblad SWC) have a spirit level to make it easier for the photographer to keep the film plane vertical.

And this is also why some specialized cameras allow the objective to be *shifted* vertically (usually rail or folding field cameras such as the *Linhof Super Technika* 2×3). Nowadays the problem of convergent lines can be solved by computer aided post-editing.

There is also another situation in photography that is linked with projective geometry (in this case three dimensional): by the Scheimpflug rule, the focus plane, the objective plane, and the film (or sensor) plane should intersect in the same line. Usually in consumer's cameras this line is at infinity, that is the three planes are parallel. But in certain situations it might be convenient to have these three planes not parallel: for example if you stand close to a wall of a room and want to reproduce its floor, which is not parallel to your film plane,

Figure 4.3: The Hasselblad SWC. The spirit level is the white circle visible on the top of the body

Figure 4.4: The Linhof Super Technika 2×3 with shifted objective

so that the whole floor is focused. In this case you need a camera that allows *tilting* the objective. This is common in large format rail cameras, for medium or small format cameras there are special (and expensive) devices or objectives that allow tilting.

The perspective rules give a hint of how a model of a real projective plane can be obtained from the Euclidean three dimensional space:

- take a point O in the three dimensional space;
- let $\mathcal{P}_{\mathbb{R}}$ be the set of lines in the three dimensional space through O, so the"points" are the lines through O and, To avoid confusion, for a point P different from O in the three dimensional space, we will denote by $[P]$ the unique "point" in the real projective plane containing O and P;
- let $\mathcal{L}_{\mathbb{R}}$ be the set of planes through O, so the"lines" are the planes through O and, as above, to avoid confusion, for a line l in the three dimensional

space , we will denote by $[l]$ the unique "line" in the real projective plane containing O and the line l;

- define the incidence relation $\iota_{\mathbb{R}}$ between points and lines as follows: a point $[P]$ is incident with a line $[l]$ if and only if $P \in l$

Then $\Pi_{\mathbb{R}} := (\mathcal{P}_{\mathbb{R}}, \mathcal{L}_R, \iota_{\mathbb{R}})$ is easily seen to be a projective plane: Two lines through O are contained in exactly one plane through O and two planes through O intersect in just one line through O. So Axioms Π.1 and Π.2 are satisfied. Finally take two distinct planes π_1 and π_2 through O and for each one of these two planes take two distinct lines different from the intersection of π_1 andπ_2. A moment's thought will convince you that these lines (i.e. points in the real projective plane) satisfy Axiom Π.3. So $\Pi(\mathbb{R})$ is a projective plane. The concept of (real) projective plane was introduced in geometry in order to avoid the exception of parallel lines (that have empty intersection if they are different) in an affine plane (we will soon define it formally, for the moment think about the Euclidean plane you studied at highschool). To understand the relation betwen the real projective plane a the affine plane, take, in the above example, a plane π in the three dimensional space such that π does not contain O. Then for every point P in π there is exactly one element $[P] \in \mathcal{P}_{\mathbb{R}}$ through P and O, and, for every line l in π there is exactly one element in $\mathcal{L}_{\mathbb{R}}$ through l and O. There is a unique element of $\mathcal{L}_{\mathbb{R}}$ that cannot be obtained this way: this is the plane π_∞ through O which is parallel to π. This element of $\mathcal{L}_{\mathbb{R}}$ is called the *line at infinity* of $\Pi_{\mathbb{R}}$ with respect to π and the points in $\mathcal{P}_{\mathbb{R}}$ incident with π_∞ are called *points at infinity*. Note that if l_1 and l_2 are two parallel lines in π, the intersection of $[l_1]$ and $[l_2]$ is a line through O parallel to π, so a point at infinity.

- Take a point O in the three dimensional Euclidean space;

- let $\mathcal{P}_{\mathbb{R}}$ be the set of lines in the three dimensional space through O, so the"points" are the lines through O and, to avoid confusion, for a point P different from O in the three dimensional space, we will denote by $[P]$ the unique "point" in the real projective plane containing O and P;

- let $\mathcal{L}_{\mathbb{R}}$ be the set of planes through O, so the"lines" are the planes through O and, as above, to avoid confusion, for a line l in the three dimensional space , we will denote by $[l]$ the unique "line" in the real projective plane containing O and the line l;

- define the incidence relation $\iota_{\mathbb{R}}$ between points and lines as follows: a point $[P]$ is incident with a line $[l]$ if and only if $P \in l$

Then $\Pi_{\mathbb{R}} := (\mathcal{P}_{\mathbb{R}}, \mathcal{L}_R, \iota_{\mathbb{R}})$ is easily seen to be a projective plane: Two lines through O are contained in exactly one plane through O and two planes through O intersect in just one line through O. So Axioms Π.1 and Π.2 are satisfied. Finally take two distinct planes π_1 and π_2 through O and for each one of these two planes take two distinct lines different from the intersection of π_1 andπ_2. A moment's thought will convince you that these lines (i.e. points in the real

projective plane) satisfy Axiom Π.3. So $\Pi(\mathbb{R})$ is a projective plane. The concept of (real) projective plane was introduced in geometry in order to avoid the exception of parallel lines (that have empty intersection if they are different) in an affine plane (we will soon define it formally, for the moment think about the Euclidean plane you studied at highschool). To understand the relation betwen the real projective plane a the affine plane, take, in the above example, a plane π in the three dimensional space such that π does not contain O. Then for every point P in π there is exactly one element $[P] \in \mathcal{P}_{\mathbb{R}}$ through P and O, and, for every line l in π there is exactly one element in $\mathcal{L}_{\mathbb{R}}$ through l and O. There is a unique element of $\mathcal{L}_{\mathbb{R}}$ that cannot be obtained this way: this is the plane π_∞ through O which is parallel to π. This element of $\mathcal{L}_{\mathbb{R}}$ is called the *line at infinity* of $\Pi_{\mathbb{R}}$ with respect to π and the points in $\mathcal{P}_{\mathbb{R}}$ incident with π_∞ are called *points at infinity*. Note that if l_1 and l_2 are two parallel lines in π, the intersection of $[l_1]$ and $[l_2]$ is a line through O parallel to π, so a point at infinity.

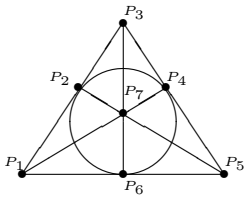
Note that if we had the intuitive notion of a four dimensional real affine space, we could use an analogous construction to get the concept of a three dimensional real projective space. In this case one might guess that the horizon will be a plane and two distinct affine planes are parallel if and only if they intersect in a line at the horizon (this is what happens with the above mentioned Scheimpflug rule for focusing a plane that is not parallel to the film plane). The good news is that, even though we have little or no intuition of a four dimensional real affine space, we can construct it artificially using algebra. We shall see this in the next section.

Figure 4.5: The Linhof Super Technika 2×3 with tilted objective and back

If you want to learn more about this subject, I believe the best book is Ansel Adam's [1].

But let's get to more exotical examples:

The Fano plane: The Fano plane is a projective plane with seven points and seven lines. It is usually represented by the picture below: the dots are the points and the lines are shown as lines or circles.



So

$$\mathcal{P} = \{P_1, P_2, P_3, P_4, P_5, P_6, P_7\}$$

and

$$\begin{aligned}\mathcal{L} \quad = \{ \quad & \{P_1, P_2, P_3\}, \ \{P_1, P_7, P_4\}, \ \{P_1, P_6, P_5\}, \ \{P_3, P_7, P_6\}, \\ & \{P_3, P_4, P_5\}, \ \{P_5, P_7, P_3\}, \ \{P_2, P_4, P_6\} \quad \}\end{aligned}$$

The reader may convince itself by a direct check that the Fano plane satisfies the axioms of a projective plane.

Note that the Fano plane could be a mathematical model for an architectural (better urbanistic) object: you can image the Fano Plane being a village, points being the houses in that village and lines being streets. A strange village, I agree. But here's a more perverse example.

The bloody exercise: Once I was asked to prepare an admittance test in mathematics to a special programme for highly gifted students. I had to find questions that needed no specific preparation in mathematics but, nevertheless needed some brilliant idea to get solved. That was a horrible task for me, anyway one of the question I found was based on a particular example of a projective plane:

Assume in a polygamic community the following hold:

1. *every two males in that community share exactly one wife;*
2. *every two females in that community share exactly one husband;*
3. *there is a set of four males such that, picking any subset of three of them, there is no woman married to all three members of that subset.*

Question 1) Does that community have parity of sexes?
Question 2) How many people live in that community?

Clearly that community satisfies the axioms of a projective plane, males being the points, females the lines and a male and a female are incident if they are married together.

Exercises 4.7.1 and 4.7.2 show how to solve the problem.

We have seen that a "point" can be the mathematical abstraction for at least

- an element of a set,
- a house,
- a male in a community,
- a line (!), see Exercise 4.7.1.

This is the power of mathematical abstraction: whenever you have a real object that satisfies some axioms of an abstract mathematical object you may use the theory you developed for that mathematical object, whatever the real object is. In bare words, when you make the addition 5+4=9 you do not bother if you are adding five apples with four oranges, or five cars with four trucks, you get 9 in any case. So consider a mathematical concept such as a point (a line or a plane) in the same way as you consider a number, without bothering what it represents, but only bothering about how it behaves.

In this sense, Kandinsky's "definition" of a point is somehow closer to mathematics than the description of it given in the officer's text: Kandinsky tries to say what an artist can do with a point. And that's why Kandinsky's book [7] could be a good reading for you.

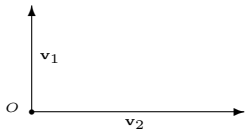
4.2 Vector spaces

In this section we will introduce the formal notion of a vector space beginning with some examples we should know from school. The concept of vector space will be fundamental for defining the concepts of dimension, point, line, plane, and for constructing a mathematical model of the affine and the projective spaces.

4.2.1 From high school physics to linear algebra*

{alpro}

We start with something we should know from school. In physics an intuitive way to visualize and study graphically the effect of forces applied to a (point-like) object O is provided by geometric vectors. These are oriented line segments, with origin in the object (point) O whose direction is the same as the force applied to O and whose length is proportional to the intensity of that force. E.g. In the picture below, the vector $\mathbf{v}_2$ represents a force f_2 applied to the point P whose direction is orthogonal to the force f_1, represented by the vector $\mathbf{v}_1$, and whose intensity is twice that of f_1.



Imagine O is a wagon pulled by a horse $\mathbf{v}$: if we attach to the wagon another horse $-\mathbf{v}$ with the same strength as $\mathbf{v}$, but pulling in the opposite direction of $\mathbf{v}$ then nothing moves. On the other hand if we attach to the wagon two horses, each with the same strength as $\mathbf{v}$, and both pulling in the same direction as $\mathbf{v}$ then the wagon accelerates two times faster as when being pulled by only one horse.

It is convenient to define also the 0-vector, whose origin O coincides with its endpoint, corresponding to no force (or equivalently forces neutralizing each other, such as in the case of the wagon pulled by two horses in opposite directions) applied to the pointlike object O .

Furthermore, given a vector $\mathbf{v}$ and a real number k we can also define the vector $k\mathbf{v}$ in the following way:

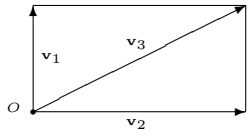
1. if either $\mathbf{v}$ is the 0-vector, or $k = 0$, define $k\mathbf{v}$ to be the 0 vector;

2. if $k > 0$ and $\mathbf{v}$ is not the 0-vector, define $k\mathbf{v}$ to be the vector whose direction is the same as that of $\mathbf{v}$ and whose length equal to k-times the length of $\mathbf{v}$; if $k < 0$ and $\mathbf{v}$ is not the 0-vector, define $k\mathbf{v}$ to be the vector whose direction is the opposite of that of $\mathbf{v}$ and whose length equal to k-times the length of $\mathbf{v}$.

So, in the picture below, given the vector $\mathbf{v}$ with origin in O and endpoint in P_1, $-\mathbf{v}$ is the vector with origin in O and endpoint in P_2, while $2\mathbf{v}$ is the vector with origin in O and endpoint in P_3.

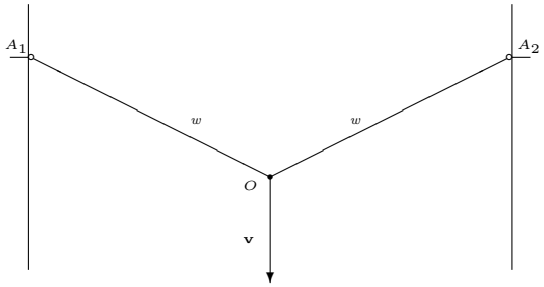
Thus, for each real number k, the multiplication by k induces a sort of *expansion* of the vector applied to O, sending every vector $\mathbf{v}$ with origin at O to its expansion $k\mathbf{v}$.

Further, if we have different forces from different directions applied simultaneously to the object O we can compute graphically the resultant of these forces using the well known *Parallelogram Rule*: indeed we know that applying simultaneously the forces f_1 and f_2 to the object O, is equivalent to applying the force f_3 represented by the vector $\mathbf{v}_3$ in the picture below:



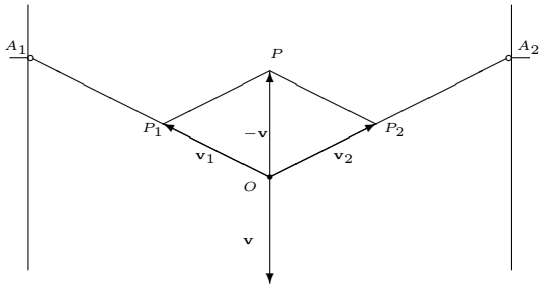
In other words the vector $\mathbf{v}_3$ represents the *sum* f_3 of the forces f_1 and f_2, and we say, therefore, that the vector $\mathbf{v}_3$ is the *sum* $\mathbf{v}_1 + \mathbf{v}_2$ of the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$. Though it is possible to prove graphically, using the axioms of Euclidean geometry, that the set of vectors with origin at a point O endowed with the addition defined by the Parallelogram Rule is an abelian group (and satisfies other properties that we'll formulate in the sequel), we'll choose instead the more comfortable algebraic way. Nevertheless, we want to show a typical application of how problems about forces applied to a point-like object O can be solved graphically using vectors:

Consider a lamp hanged in the middle of a wire which is fixed at the extremities to two anchors. Knowing the weight of the lamp, compute the forces f_1 and f_2 the anchors have to bear. These have directions starting from the anchors and leading to the centre of the wire and their intensity should be such to neutralize the weight. We can represent graphically this situation in the following picture, where O is the middle point of the wire w, A_1 and A_2 represent the anchors and $\mathbf{v}$ is the vector representing the force acting on O (i.e. the weight of the lamp)[2].



We are looking for two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ starting at O and lying on the line segments $\overline{OA_1}$ and $\overline{OA_2}$, respectively, such that their sum is equal to $-\mathbf{v}$ (here $-\mathbf{v}$ is the vector with the same intensity of $\mathbf{v}$ but with opposite direction). This is easily done inverting the procedure of the Parallelogram Rule: namely draw a parallelogram with one edge in O, one in the ending point P of the vector $-\mathbf{v}$, and whose edges are parallel to the line segments $\overline{OA_1}$ and $\overline{OA_2}$:

[2]For the meticolous reader, we are assuming that the wire has no mass and its diameter is 0.



By the Parallelogram Rule

$$-\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$$

So, for each $i \in \{1, 2\}$, the anchor A_i must be strong enough to resist to a force applied to A_i which has intensity equal to f_i and direction opposite to f_i.

Note that once we fix a point O in the plane Π_2, or the three dimensional space Π_3, every point P of Π_2 (resp. Π_3) defines a unique vector $\mathbf{v}_P := \overrightarrow{OP}$ with origin at O and endpoint at P. Conversely every vector $\mathbf{v}$ has a unique endpoint $P_{\mathbf{v}}$. Moreover, for every vector v and every point P, we have, by definition,

$$\mathbf{v} = \mathbf{v}_{P_{\mathbf{v}}} \text{ and } P = P_{\mathbf{v}_P}.$$

In other words the map, that associates to every vector with origin at O its endpoint, is a bijection between the set of points and the set of vectors with origin at O. So, once the point O has been fixed, we can identify each point P with the vector $\mathbf{v}_P$. Apparently this does not seem to make a big difference, but, as a difference to points, we can add vectors with origin in O or multiply them by real numbers obtaining other vectors still with origin in O. It is precisely this algebraic property that vectors have (and points do not) that allows us to define the concepts of lines, planes, dimension and parallelism.

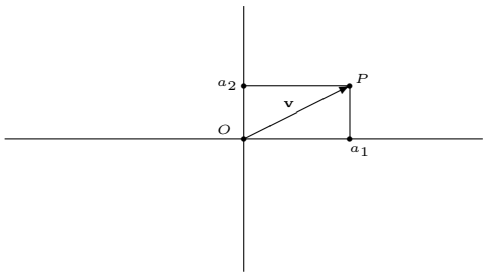
Selling the soul to the devil

> Algebra is the offer made by the devil to the mathematician. The devil says: I will give you this powerful machine, it will answer any question you like. All you need to do is give me your soul: give up geometry and you will have this marvelous machine.
> *Michael Francis Atiyah*

Now let's introduce algebra: fix a system of cartesian coordinates in the plane Π_2 or the space Π_3 with otigin at a point O. Then the set of points (hence the set of vectors with origin at O), is in bijection with the pairs (resp. triples) of real numbers[3].

[3]A word about notation: It is customary, in the theory of vector spaces, to denote the elements of $\mathbb{R}^3$ (or, more generally, $\mathbb{R}^n$) as columns instead of rows, that is we write

Let us call κ the above bijection from the set of vectors with origin in O in the three (or two) dimensional space to the set of triples (resp. pairs) of real numbers, that is the map that associates to each vector $\mathbf{v}$ the triple (resp. pair) of the coordinates of its endpoint $P_{\mathbf{v}}$. So, for example, in dimension 2 let $\mathbf{v}$ be the vector with origin in O and endpoint in P as in the picture below



Let a_1 and a_2 be the coordinates of the point P then κ associates the vector $\mathbf{v}$ with the pair $(a_1, a_2)^t$ (see footnote). We want to study the behaviour of the map κ with respect to the sum and multiple of vectors. Assume we have another vector $\mathbf{w}$ with coordinates b_1 and b_2 and endpoint Q. Let $\mathbf{z}$ be the sum of the vectors $\mathbf{v}$ and $\mathbf{w}$, c_1 and c_2 the coordinates of $\mathbf{z}$ and R its endpoint as in the picture below.

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \text{ instead of } (a, b, c)$$

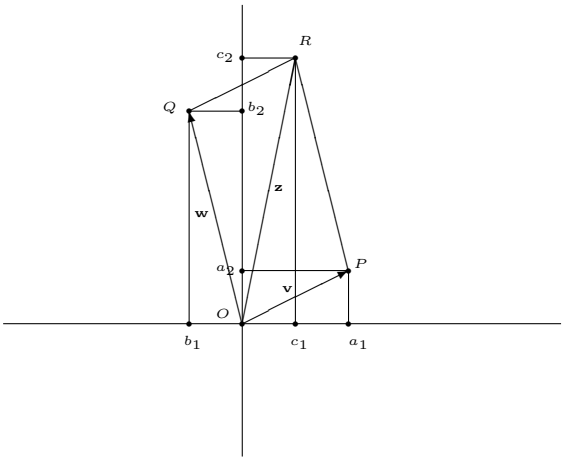
This will turn to be useful when performing matrix calculations. On the other hand the column notation is quite bad for evident typographical reasons. We shall therefore keep the column notation in formulas, but, in the discussions we shall denote by $(a, b, c)^t$ the triple

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

and, more generally, by $(a_1, a_2, \ldots, a_n)^t$ the element

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

in $\mathbb{R}^n$.



Since the line segments $\overline{QR}$ and $\overline{OP}$ are parallel and the same holds for the line segments $\overline{PR}$ and $\overline{OQ}$, it follows that

$$c_1 = a_1 + b_1 \text{ and } c_2 = a_2 + b_2$$

. In other words

$$\kappa(\mathbf{v} + \mathbf{w}) = \kappa(\mathbf{v}) + \kappa(\mathbf{w}). \tag{4.1}$$ {sumve}

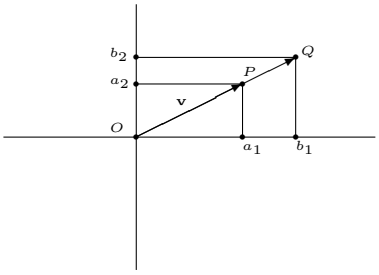
An analogous property holds also for multiples, namely we shall show that for every real number k and every vector $\mathbf{v}$ with coordinates a_1 and a_2, the coordinates of $k\mathbf{v}$ are ka_1 and ka_2, or equivalently

$$\kappa(k\mathbf{v}) = k\kappa(\mathbf{v}). \tag{4.2}$$ {scapro}

If $k = 0$ then $k\mathbf{v}$ is the zero vector, whose coordinates are 0 and 0, and clearly

$$0 = 0a_1 = 0a_2,$$

so the result is trivially true. Assume $k > 0$, let Q be the endpoint of the vector $k\mathbf{v}$ and b_1, b_2 the coordinates of $k\mathbf{v}$, as in the picture below.



Then k is the ratio

$$\frac{|\overline{OQ}|}{|\overline{OP}|}$$

between the lengths of the line segments $\overline{OQ}$ and $\overline{OP}$. Thus, since the triangles

$$\overset{\triangle}{OP}a_1 \text{ and } \overset{\triangle}{OQ}b_1$$

are similar[4], k is also equal to the ratios

$$\frac{b_1}{a_1} \text{ and } \frac{b_2}{a_2},$$

or, equivalently

$$b_1 = ka_1 \text{ and } b_2 = ka_2.$$

The case $k < 0$ is similar and left to the reader as an exercise.

Of course the same holds in dimension 3, it is only more difficult to draw pictures. Equation (4.1) leads us to define an addition in $\mathbb{R}^2$ (or $\mathbb{R}^3$) as follows (we do it for $\mathbb{R}^3$, leaving the reader to define the correspondent addition for $\mathbb{R}^2$):

$$\begin{array}{rccc} +_{\mathbb{R}^3}: & \mathbb{R}^3 \times \mathbb{R}^3 & \longrightarrow & \mathbb{R}^3 \\ & \left(\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \right) & \mapsto & \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ a_3 + b_3 \end{pmatrix} \end{array}$$

It is immediate to check that $+_{\mathbb{R}^3}$ defines a (commutative) group structure on $\mathbb{R}^3$ e.g. the operation $+_{\mathbb{R}^3}$ inherits associativity from $+$, since

$$\begin{aligned} & \left(\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} +_{\mathbb{R}^3} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \right) +_{\mathbb{R}^3} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} \\ = & \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ a_3 + b_3 \end{pmatrix} +_{\mathbb{R}^3} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} \\ = & \begin{pmatrix} (a_1 + b_1) + c_1 \\ (a_2 + b_2) + c_2 \\ (a_3 + b_3) + c_3 \end{pmatrix} = \begin{pmatrix} a_1 + (b_1 + c_1) \\ a_2 + (b_2 + c_2) \\ a_3 + (b_3 + c_3) \end{pmatrix} \\ = & \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} +_{\mathbb{R}^3} \begin{pmatrix} b_1 + c_1 \\ b_2 + c_2 \\ b_3 + c_3 \end{pmatrix} \\ = & \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} +_{\mathbb{R}^3} \left(\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} +_{\mathbb{R}^3} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} \right) \end{aligned}$$

[4]To avoid burdening the picture with extra notations we use the symbols a_1, b_1 (resp. a_2, b_2) to denote both the coordinates and the points with coordinates $(a_1, 0), (b_1, 0)$ (resp. $(0, a_2), (0, b_2)$).

Similarly one can prove that $+_{\mathbb{R}^3}$ is also commutative, that $(0,0,0)^t$ is the neutral element of $\mathbb{R}^3$ and that if $(a_1,a_2,a_3)^t$ is an element of $\mathbb{R}^3$, $(-a_1,-a_2,-a_3)^t$ is its opposite. That is

$$(\mathbb{R}^3, +_{\mathbb{R}^3}) \textit{ is an abelian group.}$$

With a harmless abuse of notation, we shall from now on use the symbol $+$ instead of $+_{\mathbb{R}^3}$. Since the map κ between the set of geometric vectors and $\mathbb{R}^3$ is a bijection that maps the sum of two vectors into the sum in $\mathbb{R}^3$ of their triples of coordinates (cfr. Equation (4.1)) it follows that also the set of geometric vectors endowed with the sum defined by the Parallelogram Rule is an abelian group[5].

In a similar way Equation (4.2) suggests us how one should multiply a triple $(a_1,a_2,a_3)^t$ in $\mathbb{R}^3$ by a real number k: namely define

$$\begin{array}{rccc} \sigma_3: & \mathbb{R}\times\mathbb{R}^3 & \longrightarrow & \mathbb{R}^3 \\ & \left(k, \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}\right) & \mapsto & \begin{pmatrix} ka_1 \\ ka_2 \\ ka_3 \end{pmatrix} \end{array}$$

For $k\in\mathbb{R}$ and $(a_1,a_2,a_3)^t \in R^3$ we denote

$$\sigma_3\left(k, \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}\right) \text{ by } k \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$$

As above (see Exercise 4.7.3), one can easily derive from the associativity and the distributivity properties of the multiplication and the addition in $\mathbb{R}$ that the map σ satisfies the following identities for every h,k in $\mathbb{R}$ and every $\mathbf{a} := (a_1,a_2,a_3)^t$ and every $\mathbf{b} := (b_1,b_2,b_3)^t$ in $\mathbb{R}^3$:

1. $1\mathbf{a} = \mathbf{a}$;
2. $(h+k)\mathbf{a} = h\mathbf{a} +_{\mathbb{R}^3} k\mathbf{a}$;
3. $(hk)\mathbf{a} = h(k\mathbf{a})$;
4. $h(\mathbf{a} +_{\mathbb{R}^3} \mathbf{b}) = h\mathbf{a} +_{\mathbb{R}^3} h\mathbf{b}$.

As above, the fact that κ is a bijection that satisfies Equation (4.2) tells us that the corresponding identities hold also for the set of geometric vectors.

As mentioned before, the fact that $\mathbb{R}^3$ is an abelian group and that the map σ satisfies the identities (1), (2), (3), and (4) is all what we need to define the concepts of line, plane, dimension and parallelism. It will also allow us to construct the affine space and the projective plane. But that's not all: note that if you just put three vertical dots in place of the second coordinate an replace

[5]To see the power of algebra, try to prove directly, using the axioms of Euclidean geometry, that the set of geometric vectors endowed with the sum defined by the Parallelogram Rule is an abelian group

3 with n, you will extend the above results on $\mathbb{R}^3$ to the sets R^n of (column) n-tuples, for every $n \in \mathbb{N} \setminus \{0\}$, so we may generalize to spaces of any finite dimension (there are also spaces of infinite dimension!) and, even if we cannot visualize how they behave, we still can predict their behaviour.

4.2.2 Formal definition and elementary properties of vector spaces

Define a *vector space over* $\mathbb{R}$ (or a *real vector space*) as a triple $(V, +_V, \sigma)$ where $(V, +_V)$ is an abelian group and $\sigma\colon \mathbb{R} \times V \longrightarrow V$ is a map that satisfies the following identities for every h, k in $\mathbb{R}$ and every $\mathbf{v}$, $\mathbf{w}$ in V:

V_1: $1\mathbf{v} = \mathbf{v}$;

V_2: $(h+k)\mathbf{v} = h\mathbf{v} +_V k\mathbf{v}$;

V_3: $(hk)\mathbf{v} = h(k\mathbf{v})$;

V_4: $h(\mathbf{v} +_V \mathbf{w}) = h\mathbf{v} +_V h\mathbf{w}$.

(as in the previous section, for every $h \in \mathbb{R}$ and $\mathbf{v} \in V$, we denote by $h\mathbf{v}$ the element $\sigma(h, \mathbf{v}$ of $V)$. The elements of V are called *vectors* and the elements of $\mathbb{R}$ are called *scalars*. The neutral element of $(V, +)$ is called the *zero vector* and will be denoted by 0_V to distinguish it from the real number 0. With a harmless abuse of notation, when it is not necessary to specify the operation $+_V$ and the function σ, we shall from now on write V instead of $(V, +_V, \sigma)$ and use the symbol $+$ instead of $+_V$.

Examples:

By the previous section and the remark at the end of it, the set of geometric vectors (with sum and multiple as defined in the previous section) and $(\mathbb{R}^3, +_{\mathbb{R}^3}, \sigma_3)$ (or, more generally, $(\mathbb{R}^n, +_{\mathbb{R}^n}, \sigma_n)$ for $n \in \mathbb{N} \setminus \{0\}$), are real vector spaces.

{propel}

Lemma 4.2.1 *Let V be a vector space then, for every $h, k \in \mathbb{R}$ and every vector $\mathbf{v} \in V$, the following identities hold:*

1. $0_V = 0\mathbf{v}$;
2. $-\mathbf{v} = (-1)\mathbf{v}$;
3. $0_V = k0_V$;
4. $k\mathbf{v} = 0_V$ *if and only if either* $k = 0$ *or* $\mathbf{v} = 0_V$;
5. *if* $\mathbf{v} \neq 0_V$ *and* $h\mathbf{v} = k\mathbf{v}$ *then* $h = k$.

Proof. Since $0 + 0 = 0$, by V2 and associativity of the addition in V, we have

$$\begin{aligned} 0_V &= -(0\mathbf{v}) + 0\mathbf{v} = -(0\mathbf{v}) + (0+0)\mathbf{v}) = -(0\mathbf{v}) + (0\mathbf{v} + 0\mathbf{v}) \\ &= (-(0\mathbf{v}) + 0\mathbf{v}) + 0\mathbf{v} = 0_V + 0\mathbf{v} = 0\mathbf{v}, \end{aligned}$$

proving *(1)*.

Similarly by *(1)*, V2, associativity, and V1,

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

giving*(2)*.

Again, since $0_V = 0_V + 0_V$, by V4 and associativity, we get

$$\begin{aligned} 0_V &= k0_V + (-(k0_V)) = k(0_V + 0_V) + (-(k0V)) = (k(0_V + k0_V) + (-(k0_V)) \\ &= k0_V + (k0_V + (-(k0_V))) = k0_V + 0_V = k0_V, \end{aligned}$$

proving *(3)*.

By *(1)* and *(3)*, if $k = 0$ or $\mathbf{v} = 0_V$, we have $k\mathbf{v} = 0_V$. Conversely, assume $k\mathbf{v} = 0_V$ and $k \neq 0$, then k is invertible in $\mathbb{R}$, so, by V1, V3, and *(3)*,

$$\mathbf{v} = 1\mathbf{v} = (k^{-1}k)\mathbf{v} = k^{-1}(k\mathbf{v}) = k^{-1}0_V = 0_V,$$

giving *(4)*.

Finally, if $v \neq 0_V$ and $h\mathbf{v} = k\mathbf{v}$, then, by V2, V3, and *(2)*,

$$(h-k)\mathbf{v} = (h + (-1)k))\mathbf{v} = h\mathbf{v} + ((-1)k)\mathbf{v} = h\mathbf{v} + (-1)(k\mathbf{v}) = k\mathbf{v} + (-(k\mathbf{v})) = 0_V.$$

Since $\mathbf{v} \neq 0_V$, this implies that $h - k = 0$, whence $h = k$. ∎

4.2.3 Subspaces

Given a vector space $V = (V, +_V, \sigma)$, a *subspace* of V is a subgroup W of $(V, +_V)$ that is *closed under multiplication by scalars*, i.e. it satisfies the following condition:

VS $k\mathbf{w} \in W$ for every $k \in \mathbb{R}$ and $\mathbf{w} \in W$.

{subs}

Lemma 4.2.2 *A subset W of a vector space V is a subspace if and only if W is not empty and W is closed under addition and multiplication by scalars.*

Proof. Clearly a subspace is not empty and closed under addition (because it is a subgroup) and by multiplication by scalars by VS. Conversely, assume W is not empty and closed under addition and multiplication by scalars. Then, by Lemma 4.2.1*(2)*, for every $\mathbf{v} \in W$,

$$-\mathbf{v} = (-1)\mathbf{v} \in W,$$

so W is also closed by making of the opposite, that is W is a subgroup, whence a subspace. ∎

Examples:

The set of geometric vectors whose endpoint lies in a line containing the origin O is a subspace of the space of geometric vectors.

The set of geometric vectors whose endpoint lies in a plane containing the origin O is a subspace of the space of geometric vectors.

Let V be a vector space then V itself and $\{0_V\}$ are subspaces of V (we leave to the reader the obvious proofs). These are called the *trivial subspaces*. Any subspace of V different from V is called *proper*.

Let V be a vector space, let $\mathbf{v} \in V$ and let

$$\mathbb{R}\mathbf{v} := \{k\mathbf{v} | k \in \mathbb{R}\}.$$

Then $\mathbb{R}\mathbf{v}$ is a subspace of V. Indeed, by Lemma 4.2.1*(1)*,

$$0_V = 0\mathbf{v} \in \mathbb{R}\mathbf{v},$$

so $\mathbb{R}\mathbf{v}$ is not empty. If $h\mathbf{v}$ and $k\mathbf{v}$ are elements of $\mathbb{R}\mathbf{v}$, with $h, k \in \mathbb{R}$, then, by V2,

$$h\mathbf{v} +_V k\mathbf{v} = (h + k)\mathbf{v} \in \mathbb{R},$$

so $\mathbb{R}\mathbf{v}$ is also closed under addition. Finally, by V3,

$$h(k\mathbf{v}) = (hk)\mathbf{v} \in \mathbb{R}\mathbf{v},$$

so $\mathbb{R}\mathbf{v}$ is also closed under multiplication by scalars, whence $\mathbb{R}\mathbf{v}$ is a subspace of V. $\mathbb{R}\mathbf{v}$ is called the *subspace* of V *spanned* by the vector $\mathbf{v}$.

$\mathbb{Q}^2$ is a subgroup, but not a subspace of $\mathbb{R}^2$: Denote by $\mathbb{Q}^2$ the set $\{(a, b)^t | a, b \in \mathbb{Q}\}$. Then one sees easily that $\mathbb{Q}^2$ is a subgroup of $\mathbb{R}^2$, but it is not a subspace, because it is not closed under multiplication by scalars: e.g. $(1, 1)^t \in \mathbb{Q}^2$ and $\sqrt{2} \in \mathbb{R}$. Since $\sqrt{2} \notin \mathbb{Q}$, $\sqrt{2}(1, 1)^t = (\sqrt{2}, \sqrt{2}) \notin \mathbb{Q}^2$.

Let U and W be subspaces of a vector space V. Define the *sum* $U + W$ of the subspaces U and W as the set of all elements of V that are sums of an element of U with an element of W:

$$U + W := \{u +_V w | u \in U \text{ and } w \in W\}.$$

{sumsps}

Lemma 4.2.3 *Let U and W be subspaces of a vector space V. Then $U + W$ is a subspace of V*

PROOF. Since $0_V \in U \cap W$ and $0_V = 0_V + 0_V$, it follows that

$$0_V \in U + W,$$

whence $U + W \neq \emptyset$. Assume $\mathbf{u}_1 +_V \mathbf{w}_1$ and $\mathbf{u}_2 +_V \mathbf{w}_2$ are elements of $U + W$ with $\mathbf{u}_1, \mathbf{u}_2 \in U$ and $\mathbf{w}_1, \mathbf{w}_2 \in W$. Then, $(\mathbf{u}_1 +_V \mathbf{u}_2) \in U$ and $(\mathbf{w}_1 +_V \mathbf{w}_1) \in W$, since U and W are closed under addition, so by associativity and commutativity of $+_V$

$$\begin{aligned}(\mathbf{u}_1 +_V \mathbf{w}_1) +_V (\mathbf{u}_2 +_V \mathbf{w}_2) &= \mathbf{u}_1 +_V (\mathbf{w}_1 +_V \mathbf{u}_2) +_V \mathbf{w}_2 \\ &= \mathbf{u}_1 +_V (\mathbf{u}_2 +_V \mathbf{w}_1) +_V \mathbf{w}_2 \\ &= (\mathbf{u}_1 +_V \mathbf{u}_2) +_V (\mathbf{w}_1 +_V \mathbf{w}_2) \in U + W,\end{aligned}$$

whence also $U + W$ is closed under addition. Finally assume $h \in \mathbb{R}$ and $\mathbf{u} + \mathbf{w} \in U + W$ with $\mathbf{u} \in U$ and $\mathbf{w} \in W$. Then $h\mathbf{u} \in U$ and $h\mathbf{w} \in W$, since U and W are closed under multiplication by scalars, so, by V4,

$$h(\mathbf{u} +_V \mathbf{w}) = h\mathbf{u} +_V h\mathbf{w} \in U + W,$$

whence $U + W$ is also closed under multiplication by scalars, hence $U + W$ is a subspace of V. ∎

{intsubs}

Lemma 4.2.4 *Let U and W be subspaces of a vector space V. Then $U \cap W$ is a subspace of V*

PROOF. By Lemma 3.2.4, $U \cap W$ is a subgroup of V. It is also closed under multiplication by scalars, for if $h \in \mathbb{R}$ and $\mathbf{z} \in U \cap W$, then $h\mathbf{z} \in U$ and $h\mathbf{z} \in W$, since U and W are closed under multiplication by scalars. So $h\mathbf{z} \in U \cap W$ and the result follows. ∎

4.2.4 Linear combinations

{2.6}

Now assume $\mathbf{w}_1, \ldots, \mathbf{w}_k$ are elements of a subspace W of a vector space V. Since W is closed under multiplication by scalars, it follows that, for every $i \in \{1, \ldots, k\}$ and every $a_i \in \mathbb{R}$,

$$a_i \mathbf{w_i} \in W.$$

Further, since W is also a subgroup of V, it follows that, for every $a_1, a_2, \ldots, a_k \in \mathbb{R}$, the vector

$$\mathbf{w} := a_1\mathbf{w}_1 + a_2\mathbf{w}_2 + \cdots + a_k\mathbf{w}_k \tag{4.3}$$

{lincomb}

is still an element of W. The vector $\mathbf{w}$ in (4.3) is called *a linear combination* of the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_k$ and the scalars $a_1, a_2, \ldots, a_k$ are called the *coefficients* of that linear combination. We have proved that every subspace of V is closed under the making of linear combinations. More precisely

{lincomb1}

Lemma 4.2.5 *If $\mathbf{w}_1, \ldots, \mathbf{w}_k$ are elements of a subspace W of a vector space V, then W contains all the linear combinations with coefficients in $\mathbb{R}$ of $\mathbf{w}_1, \ldots, \mathbf{w}_k$.*

Conversely, if $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are vectors of V denote by $\langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle$ the set of all linear combinations of $\mathbf{v}_1, \ldots, \mathbf{v}_k$ with coefficients in $\mathbb{R}$. Then

{lincomb2}

Lemma 4.2.6 *If $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are vectors of V, then $\langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle$ is a subspace of V containing the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$.*

PROOF. Let $i \in \{1, \ldots, k\}$ then, since $0\mathbf{v} = 0_V$ and $1\mathbf{v} = \mathbf{v}$, for every $\mathbf{v} \in V$, we have

$$\mathbf{v}_i = a_1\mathbf{v_1} + a_2\mathbf{v}_2 + \cdots + a_k\mathbf{v_k}$$

choosing $a_j = 0$ if $j \neq i$ and $a_i = 1$. So $\mathbf{v_i}$ is a linear combination of $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, proving that

$$\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subseteq \langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle,$$

in particular $\langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle$ is not empty. Now assume

$$a_1\mathbf{v_1} + a_2\mathbf{v}_2 + \cdots + a_k\mathbf{v_k} \text{ and } b_1\mathbf{v_1} + b_2\mathbf{v}_2 + \cdots + b_k\mathbf{v_k}$$

are elements of $\langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle$. Then, by associativity and commutativity of the addition in V and by V2,

$$\begin{aligned}&(a_1\mathbf{v_1} + a_2\mathbf{v}_2 + \cdots + a_k\mathbf{v_k}) + (b_1\mathbf{v_1} + b_2\mathbf{v}_2 + \cdots + b_k\mathbf{v_k})\\ &= (a_1 + b_1)\mathbf{v_1} + (a_2 + b_2)\mathbf{v}_2 + \cdots + (a_k + b_k)\mathbf{v_k} \in \langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle,\end{aligned}$$

so $\langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle$ is closed under addition. Finally if $h \in \mathbb{R}$, then, by V4 and V3,

$$\begin{aligned}&h(a_1\mathbf{v_1} + a_2\mathbf{v}_2 + \cdots + a_k\mathbf{v_k}) = h(a_1\mathbf{v_1}) + h(a_2\mathbf{v}_2) + \cdots + h(a_k\mathbf{v_k})\\ &= (ha_1)\mathbf{v_1} + (ha_2)\mathbf{v}_2) + \cdots + (ha_k)\mathbf{v_k} \in \langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle,\end{aligned}$$

so $\langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle$ is also closed under the multiplication by scalars and the result follows ■

The set $\langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle$ is called the subspace of V *spanned* by the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$. By Lemma 4.4 $\langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle$ is the smallest subspace of V containing the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$. Note also that, if $k = 1$,

$$\langle \mathbf{v}_1 \rangle = \mathbb{R}\mathbf{v}_1.$$

Example:

The canonical basis for $\mathbb{R}^3$: Every vector $(a, b, c)^t$ in $\mathbb{R}^3$ is a linear combination of the vectors $(1, 0, 0)^t$, $(0, 1, 0)^t$, and $(0, 0, 1)^t$:

$$\begin{aligned}(a, b, c)^t &= (a, 0, 0)^t + (0, b, 0)^t + (0, 0, c)^t\\ &= a(1, 0, 0)^t + b(0, 1, 0)^t + c(0, 0, 1)^t \in \langle (1, 0, 0)^t, (0, 1, 0)^t, (0, 0, 1)^t \rangle,\end{aligned}$$

so $\mathbb{R}^3 = \langle (1,0,0)^t, (0,1,0)^t, (0,0,1)^t \rangle$. Note that the one above is the unique way to write $(a,b,c)^t$ as a linear combination of the vectors $(1,0,0)^t$, $(0,1,0)^t$, and $(0,0,1)^t$, for, if

$$(a,b,c)^t = a'(1,0,0)^t + b'(0,1,0)^t + c'(0,0,1)^t,$$

then, since

$$a'(1,0,0)^t + b'(0,1,0)^t + c'(0,0,1)^t = (a',b',c')^t,$$

we have

$$(a,b,c)^t = (a',b',c')^t,$$

whence $a = a'$, $b = b'$, and $c = c'$. The set $\{(1,0,0)^t, (0,1,0)^t, (0,0,1)^t\}$ is called the unordered *canonical base* of $\mathbb{R}^3$. It is an unordered basis in the sense that I'll soon define.

4.2.5 Linear dependence

Let V be a vector space and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be vectors of V. We say that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are *linearly dependent* if one of them is a linear combination of the others. If $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are not linearly dependent we say that they are linearly *independent*.

Examples

- The vectors
$$\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \text{ and } \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$
are linearly dependent because
$$\begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + 2 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

- The vectors
$$\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$
are linearly independent, because for every every scalar a,
$$a \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ a \\ 0 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$
and
$$a \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ a \\ a \end{pmatrix} \neq \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix},$$
hence none of them can be a linear combination (a scalar multiple in this case) of the other.

{rep}

Lemma 4.2.7 *Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be vectors of a vector space V. If there exists $i, j \in \{1, \ldots, n\}$ such that $i < j$ and $\mathbf{v}_i = \mathbf{v}_j$, then $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly dependent.*

PROOF. To simplify notation we can rearrange the indexes such that $i = n-1$ and $j = n$. Then have

$$\mathbf{v}_n = \mathbf{v}_{n-1} = 0a_1\mathbf{v_1} + 0\mathbf{v}_2 + \cdots + 0\mathbf{v}_{n-2} + 1\mathbf{v}_{n-1},$$

so $\mathbf{v}_n$ is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}$. ∎

{null}

Lemma 4.2.8 *Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be vectors of a vector space V. If there exists $i \in \{1, \ldots, n\}$ such that $\mathbf{v}_i = 0_V$, then $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly dependent*

PROOF. As in the proof of Lemma 4.2.7, to simplify notation we can rearrange the indexes such that $i = n$. Then have

$$\mathbf{v}_n = 0\mathbf{v_1} + 0\mathbf{v}_2 + \cdots + 0\mathbf{v}_{n-2} + 0\mathbf{v}_{n-1},$$

so $\mathbf{v}_n$ is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}$. ∎

If $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are vectors of a vector space V, and $\mathbf{v} \in \langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \rangle$ we shall say that $\mathbf{v}$ *can be written in a unique way as a linear combination of* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ if for every $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ in $\mathbb{R}$, the equality

$$a_1\mathbf{v_1} + a_2\mathbf{v_2} + \cdots + a_n\mathbf{v_n} = b_1\mathbf{v_1} + b_2\mathbf{v_2} + \cdots + b_n\mathbf{v_n}$$

implies that

$$a_i = b_i \text{ for every } i \in \{1, \ldots, n\}.$$

Here's a useful characterisation of linear dependence:

{2.9}

Lemma 4.2.9 *Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be vectors of a vector space V and set*

$$W := \langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \rangle.$$

Then the following assertions are equivalent:

1. *$\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly dependent;*
2. *there is a vector $\mathbf{w} \in W$ that can be written in two different ways as a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$;*
3. *the zero vector 0_V can be written as a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ where not all coefficients are equal to 0*

PROOF. *((1)⇒(2))* Assume $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly dependent. Then one of them is a linear combination of the others. As above, rearranging the indexes, we may assume that $\mathbf{v}_n$ is a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}$, say

$$\mathbf{v}_n = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_{n-1}\mathbf{v}_{n-1}.$$

Then also

{lincomb1}
$$\mathbf{v}_n = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_{n-1}\mathbf{v}_{n-1} + 0\mathbf{v}_n. \tag{4.4}$$

On the other hand, we have also

{lincomb2}
$$\mathbf{v}_n = 0\mathbf{v}_1 + 0\mathbf{v}_2 + \cdots + 0\mathbf{v}_{n-1} + 1\mathbf{v}_n, \tag{4.5}$$

so we have found two different ways to write $\mathbf{v}_n$ as a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ and so *(2)* holds. (they are different because the coefficient of $\mathbf{v}_n$ in Equation (4.4) is 0 while in Equation (4.5) it is 1).

((2)⇒(3)) This is obvious: assume

{zero1}
$$0_V = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_{n-1}\mathbf{v}_{n-1} + 0\mathbf{v}_n \tag{4.6}$$

with some of the a_i's different from 0. Since we have also

{zero2}
$$0_V = 0\mathbf{v}_1 + 0\mathbf{v}_2 + \cdots + 0\mathbf{v}_{n-1} + 0\mathbf{v}_n, \tag{4.7}$$

we get that 0_V can be written in two different ways, (4.6) and (4.7), as a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, giving *(3)*, since $0_V \in W$.

((3)⇒(1)) Assume the zero vector 0_V can be written as a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ where not alla coefficients are equal to 0:

{zero3}
$$0_V = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_{n-1}\mathbf{v}_{n-1} + a_n\mathbf{v}_n \tag{4.8}$$

Again, rearranging the indexes, we may assume that in Equation (4.8), the coefficient a_n of $\mathbf{v}_n$ is different from 0. Then we can rewrite Equation (4.8) as

{zero4}
$$+a_n\mathbf{v}_n = (-a_1)\mathbf{v}_1 + (-a_2)\mathbf{v}_2 + \cdots + (-a_{n-1})\mathbf{v}_{n-1}. \tag{4.9}$$

Dividing both members of Equation (4.9), by a_n (which is possible, since $a_n \neq 0$), we get

$$\begin{aligned} \mathbf{v}_n &= a_n^{-1}((-a_1)\mathbf{v}_1 + (-a_2)\mathbf{v}_2 + \cdots + (-a_{n-1})\mathbf{v}_{n-1}) \\ &= a_n^{-1}(-a_1)\mathbf{v}_1 + a_n^{-1}(-a_2)\mathbf{v}_2 + \cdots + a_n^{-1}(-a_{n-1})\mathbf{v}_{n-1}), \end{aligned}$$

so $\mathbf{v}_n$ is a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}$, whence $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly dependent, giving *(1)*. ∎

{a1}

Corollary 4.2.10 *Let V be a vector space, X a finite set of linearly independent vectors of V and Y a nonempty subset of X. Then the vectors in Y are linearly independent.*

PROOF. Let

$$X := \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}.$$

Rearranging the indexes, we may assume that

$$Y := \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\},$$

with $m \leq n$. Assume, by means of contradiction that the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ are linearly dependent. then, by Lemma 4.2.9, there exist $a_1, \ldots, a_k \in \mathbb{R}$ not all equal to 0 such that

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_k\mathbf{v}_m = 0_V.$$

But then

$$0_V = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_k\mathbf{v}_m + 0\mathbf{v}_{m+1} + \cdots + 0\mathbf{v}_n$$

against the asumption that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly independent. ∎

{nongen}

Lemma 4.2.11 *Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be vectors of a vector space V. and set $W := \langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \rangle$. If $\mathbf{v}_n$ is a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}$, then*

$$\langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1} \rangle = \langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \rangle$$

PROOF. Assume

$$v_n = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \ldots + a_{n-1}\mathbf{v}_{n-1}.$$

Clearly

$$\langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1} \rangle \leq \langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \rangle,$$

for if

$$b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + \ldots + b_{n-1}\mathbf{v}_{n-1} \in \langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1} \rangle,$$

then

$$b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + \ldots + b_{n-1}\mathbf{v}_{n-1} = b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + \ldots + b_{n-1}\mathbf{v}_{n-1} + 0\mathbf{v}_n \in \langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \rangle.$$

Conversely assume

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_{n-1}\mathbf{v}_{n-1} + c_n\mathbf{v}_n \in \langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \rangle.$$

Substituting $\mathbf{v}_n$ by $a_1\mathbf{v}_1+a_2\mathbf{v}_2+\ldots+a_{n-1}\mathbf{v}_{n-1}$, we get, by V4, V3, associativity and commutativity of the addition in V and V2,

$$\begin{aligned}
&c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_{n-1}\mathbf{v}_{n-1} + c_n\mathbf{v}_n \\
&= c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_{n-1}\mathbf{v}_{n-1} + c_n(a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \ldots + a_{n-1}\mathbf{v}_{n-1}) \\
&= c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_{n-1}\mathbf{v}_{n-1} + c_n(a_1\mathbf{v}_1) + c_n(a_2\mathbf{v}_2) + \ldots + c_n(a_{n-1}\mathbf{v}_{n-1}) \\
&= c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_{n-1}\mathbf{v}_{n-1} + (c_na_1)\mathbf{v}_1 + (c_na_2)\mathbf{v}_2 + \ldots + (c_na_{n-1})\mathbf{v}_{n-1} \\
&= c_1\mathbf{v}_1 + (c_na_1)\mathbf{v}_1 + c_2\mathbf{v}_2 + \ (c_na_2)\mathbf{v}_2 + \ldots + c_{n-1}\mathbf{v}_{n-1} + (c_na_{n-1})\mathbf{v}_{n-1} \\
&= (c_1 + c_na_1)\mathbf{v}_1 + (c_2 + c_na_2)\mathbf{v}_2 + \ldots + (c_{n-1} + c_na_{n-1})\mathbf{v}_{n-1} \\
&\in \langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}\rangle.
\end{aligned}$$

■

4.2.6 Bases and dimension

We get now to the core of this chapter, where we introduce two fundamental concepts: the concept of basis and the concept of dimension. Bases are essentially minimal spanning sets of a vector space V. The main results of this subsection are that if a vector space V is spanned by a finite set of elements, then V has bases and all bases for V have the same number of elements. This number is the dimension of V. Remarkably this number encodes all the information on V.

An *unordered base* for a vector space V is a subset $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ of V such that

B1 the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ span V,

B2 the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly independent.

Examples

The set

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

is an unordered basis of $\mathbb{R}^3$ since, as we have seen in Subsection 4.2.4, every element of $\mathbb{R}^3$ can be written in a unique way as a linear combination of $1, 0, 0)^t$, $(0, 1, 0)^t$, and $(0, 0, 1)^t$.

The set

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \right\}$$

is not an unordered basis of $\mathbb{R}^3$ since, e.g., $(1, 0, 0)^t$ can be written as a linear combination of $(0, 1, 0)^t$, $(0, 0, 1)^t$, and $(1, 2, 1)^t$ (how?).

The set

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}$$

is not an unordered basis of $\mathbb{R}^3$ since, e. g. the vector $(0,0,1)^t$ cannot be written as a linear combination of $(1,0,0)^t$ and $(0,1,0)^t$.

If $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is an unordered basis for V with all $\mathbf{v}_i$'s distinct, the n-tuple $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ is called a *basis* for V. So the difference between an unordered basis and a basis is that in a basis we take into account the order with which we take the elements: e. g. The sets

The set

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\} \text{ and } \left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}$$

are the same, so they are the same unordered bases of $\mathbb{R}^3$, but the triples

$$\left(\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right) \text{ and } \left(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right)$$

are two different bases of $\mathbb{R}^3$.

A key feature of the bases is that, if $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is a basis of a vector space V, then every element of v can be written in a unique way as a linear combination of the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. More precisely:

{uniqueness}

Theorem 4.2.12 *Let $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ be a basis of a vector space V. Then, for every vector $\mathbf{v}$ of V there is a unique n-tuple of scalars $(a_1, \ldots, a_n)^t$ in $\mathbb{R}^n$ such that*

$$\mathbf{v} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n.$$

Moreover, if also $(b_1, \ldots, b_n)^t$ is another n-tuple in $\mathbb{R}^n$ such that

$$\mathbf{v} = b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + \ldots + b_n\mathbf{v}_n,$$

then

$$a_1 = b_1, a_2 = b_2, \ldots, a_n = b_n.$$

PROOF. Since $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ generate V, every vector of V is a linear combination of them. The uniqueness follows immediately from Lemma 4.2.9. ∎

We now come to the two main results of this section:

{basex}

Theorem 4.2.13 *Let V be a vector space and let $X := \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be a subset of V such that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ span V. Then there is a subset B of X that is an unordered basis for V.*

PROOF. This is done by "purifying" the set X into an unordered basis by iterated use of Lemma 4.2.11. If

$$\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$$

are linearly independent, we are done. Assume $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are not linearly independent, then one of them is a linear combination of the others. As usual, rearranging the indexes, we may assume that $\mathbf{v_n}$ is a linear combination of $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}\}$. By Lemma 4.2.11,

$$V = \langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1} \rangle.$$

Now, if

$$\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}$$

are linearly independent we are done as above, otherwise, again, one of them is a linear combination of the others and we can cancel it. Iterating this process we end up with a spanning set B of V such that $B \subseteq X$ and whose elements are linearly independent, i.e. B is an unordered basis for V and B is contained in X. ∎

{missing}

Lemma 4.2.14 [ROCCO'S LEMMA][6] *Assume $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent elements of a vector space V and let $\mathbf{w} \in V$. If $\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{w}$ are linearly dependent, then*

$$\mathbf{w} \in \langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle.$$

PROOF. Since $\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{w}$ are linearly dependent, by Lemma 4.2.9, there exist scalars $a_1, \ldots, a_{k+1}$ in $\mathbb{R}$ not all equal to 0 such that

$$a_1\mathbf{v}_1 + \cdots + a_k\mathbf{v}_k + a_{k+1}\mathbf{w} = 0_V. \tag{4.10}$$ {Rocco}

Now a_{k+1} cannot be zero, otherwise Equation (4.10) would become

$$a_1\mathbf{v}_1 + \cdots + a_k\mathbf{v}_k = 0_V$$

with one of $a_1, \ldots, a_k$ not zero, against the hypothesis that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent. So, by Equation (4.10)

$$\mathbf{w} = (-a_{k+1}^{-1}a_1)\mathbf{v}_1 + \cdots + (-a_{k+1}^{-1}a_k)\mathbf{v}_k \in \langle \mathbf{v}_1, \ldots, v_k \rangle.$$

∎

[6] Rocco was a student following my 2020/2021 WS course in linear algebra at the Department of Engeneering and Architecture of the University of Udine. He listened very carefully to my lectures, not hesitating to ask for explanations if there was something he did not understand. Such students have always been an immense help for my lectures. That's why he deserves a Lemma

{bodysnatchers}

Theorem 4.2.15 [The Bodysnatcher's Theorem][7]

Let $Y := \{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ be a set of k linearly independent vectors of a vector space V and let $B := \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be an unordered basis of V with n distinct elements. Then

1. *$k \leq n$;*
2. *there is a subset Y' of B such that $|Y'| = k$ and $(B \backslash Y') \cup Y$ is an unordered basis for V.*

Proof. We first show how to replace one of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ with $\mathbf{w}_1$. Since $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ span V and $\mathbf{w}_1 \in V$, there exist $a_1, a_2, \ldots, a_n \in \mathbb{R}$ such that

{b1}
$$\mathbf{w}_1 = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n \tag{4.11}$$

Since $\mathbf{w}_1 \neq 0_V$ some of the coefficients in Equation (4.11) have to be different from 0. Rearranging the indexes, we may assume that $a_1 \neq 0$. Then we can rewrite Equation (4.11), as follows:

{b2}
$$\mathbf{v}_1 = a_1^{-1}\mathbf{w}_1 + (-a_2a_1^{-1})\mathbf{v}_2 + \cdots + (-a_{n-1}a_1^{-1})\mathbf{v}_{n-1}. \tag{4.12}$$

By Lemma 4.2.11 we have

$$\langle \mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \rangle = \langle \mathbf{v}_1, \mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \rangle = \langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \rangle = V.$$

Further $\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are still linearly independent. Indeed, assume by means of contradiction that $\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ were not linearly independent. Since, by Lemma 4.2.10, $\mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly independent, it follows that

$$\mathbf{w}_1 \in \langle \mathbf{v}_2, \ldots, \mathbf{v}_n \rangle.$$

But then, by Equation (4.11) and Lemma 4.2.11,

$$\mathbf{v}_1 \in \langle \mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \rangle \in \langle \mathbf{v}_2, \ldots, \mathbf{v}_n \rangle,$$

which is a contradiction since, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ have been asssumed to be linearly independent. So $\{\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is an unordered basis for V containing n distinct vectors.

Now that we have replaced $\mathbf{v}_1$ with $\mathbf{w}_1$ we proceed with $\mathbf{w}_2$. As above $\mathbf{w}_2$ can be written as a linear combination of $\mathbf{w}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, say

{b2}
$$\mathbf{w}_2 = b_1\mathbf{w}_1 + b_2\mathbf{v}_2 + \cdots + b_n\mathbf{v}_n \tag{4.13}$$

[7] I gave this name to this theorem, because its proof reminda me of the 1952 Don Segals' movie *The invasion of the bodysnatchers*, where humans are replaced by their alien duplicates emerging from giant seed pods, without anyone seeing the difference. In the proof I assume the elements of the set Y as the aliens replacing one by one the "humans", i.e. the elements of the basis B, without anyone seeing the difference, since, after every replacement, we still have a basis.

Note that the coefficients $b_2, \ldots, b_n$ cannot be all equal to 0, for otherwise

$$\mathbf{w}_2 = b_1\mathbf{w}_1$$

which is not possible since, by Lemma 4.2.10, $\mathbf{w}_1$ and $\mathbf{w}_2$ are linearly independent. Again, rearranging the indexes, we may assume that $b_2 \neq 0$ and, proceeding as above, we can replace $\mathbf{v}_2$ with $\mathbf{w_2}$ and get that $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{v}_n\}$ is an unordered basis for V containing n distinct vectors.

Now it should be clear how to proceed in general: Assume that, for $i \in \{1, \ldots, k-1\}$ with $i < n$, we have replaced i vectors of the set B with $\mathbf{w}_1, \ldots, \mathbf{w}_i$. Again, after rearranging the indexes of the vectors in B, we may assume that the vectors in B that have been replaced are $\mathbf{v}_1, \ldots, \mathbf{v}_i$. So $\{\mathbf{w}_1, \ldots, \mathbf{w}_i, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_n\}$ is an unordered basis of V containing n distinct vectors. Again, since

$$\mathbf{w}_{i+1} \in V = \langle \mathbf{w}_1, \ldots, \mathbf{w}_i, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_n \rangle,$$

there are scalars $c_1, \ldots, c_n$, such that

$$\mathbf{w}_{i+1} = c_1\mathbf{w}_1 + \cdots + c_i\mathbf{w}_i + c_{i+1}\mathbf{v}_{i+1} + \cdots + c_n\mathbf{v}_n\rangle. \tag{4.14}$$ {b3}

As above, there exists an index $j \in \{i+1, \ldots, n\}$ such that $c_j \neq 0$, otherwise $\mathbf{w}_1, \ldots, \mathbf{w}_i, \mathbf{w}_{i+1}$ would be linearly dependent, against Lemma 4.2.10. Rearranging the indexes we may assume that $j = i+1$, and, as above, we get that $\{\mathbf{w}_1, \ldots, \mathbf{w}_i, \mathbf{w}_{i+1}, \mathbf{v}_{i+2}, \ldots, \mathbf{v}_n\}$ is an unordered basis of V containing n distinct vectors. Assume $k < n$, then after iterating this procedure k-times we get from B an unordered basis B' of V consisting of n elements, k of which are the elements of Y and the remaining ones are elements of $B \setminus Y$. Assume finally $k \geq n$. Then after n steps we have replaced all elements $\{\mathbf{v}_1, \ldots, \mathbf{v}_n$ of B by the first n elements $\{\mathbf{w}_1, \ldots, \mathbf{w}_n$ of Y obtaining another basis of V. Assume, by means of contradiction that $k > n$, then

$$\mathbf{w}_{n+1} \in V = \langle \mathbf{w}_1, \ldots, \mathbf{w}_n \rangle,$$

against the assumption that $\mathbf{w}_1, \ldots, \mathbf{w}_k$ are linearly independent. ∎

{dim}

Corollary 4.2.16 *Let B be a (finite) unordered basis for a vector space V.*

1. *If Y is a set of linearly independent vectors in V, then $|Y| \leq |B|$.*
2. *If B' is another unordered basis for V, then $|B| = |B'|$.*
3. *If X is a (finite) spanning set for V, then $|X| \geq |B|$.*

PROOF. The first assertion follows immediately from Theorem 4.2.15. Since an unordered basis is also a set of linearly independent elements of V taking Y as B' in *(1)*, we get $|B'| \leq |B|$ and the reverse inequality follows by swapping the rôles of B and B', giving *(2)*. The last assertion follows from *(2)* and the fact

that every finite spanning set of V contains an unordered basis (Theorem 4.2.13). ■

We are now ready to define one of the main concepts of this chapter. First observe that every vector space V over the real numbers, except for the trivial space $\{0\}$ containing only the zero vector, contains infinitely many elements, for if $\mathbf{v} \in V \setminus \{0\}$, then V must contain the set $\{k\mathbf{v}|a \in \mathbb{R}\}$ of all scalar multiples of $\mathbf{v}$ and this set is infinite, since, by Lemma 4.2.1

$$k\mathbf{v} = h\mathbf{v} \text{ if and only if } h = k.$$

This means that the concept of order is nearly of no use for studying vector spaces. On the other hand, As anticipated in Subsection 2.4.3, we still have a parameter that plays for vector spaces a role analogous to the order for finite sets. This is the *dimension*.

If V is a vector space spanned by a finite set of vectors we say that V is *finite dimensional*. If V is finite dimensional, the *dimension* of V is the number of elements in one (any) basis of V. The dimension of V will be denoted by $\dim(V)$.

Since a common mistake of the students is confusing the dimension the number of elements of a vector space, I stress again that **the dimension of a vector space V is NOT the number of elements of V, but only the number of elements in a basis of V.**

Here are two elementary, though very useful facts about dimension (compare with the analogous result in Subsection 2.4.3: Lemma 2.6.5 and Lemma 2.4.13);

{bask}

Lemma 4.2.17 *Let V be a finite dimensional vector space, then*

1. *if W is a subspace of V then W is also finite dimensional and* $\dim(W) \leq \dim(V)$*;*
2. *if W is a subspace of V and* $\dim(W) = \dim(V)$*, then $W = V$.*

PROOF. For *(1)*, any set Y of linearly independent vectors of W is also a set of linearly independent vectors of V, so, by Corollary 4.2.16*(1)*, $|Y|$ cannot exceed $\dim(V)$. In particular, if we choose Y so that $|Y|$ is maximal, then the vectors of Y are linearly independent and, by Lemma 4.2.14, for every $w \in W \setminus Y$ w is a linear combination of the vectors in Y, so Y is a basis for W and

$$\dim(W) = |Y| \leq dim(V).$$

For *(2)*, assume $W \leq V$ and $\dim(W) = \dim(V)$. let Y be an unordered basis of W and B be a basis of V. Since $|Y| = |B|$, by Theorem 4.2.15, Y is also a basis for V, whence $W = \langle Y \rangle = V$. ■

{Grassmann}

Theorem 4.2.18 [GRASSMANN'S THEOREM] *If V is a finite dimensional vector space and U and W are subspaces of V, then*

$$\dim(U + W) = (\dim(U) + \dim(W)) - \dim(U \cap W).$$

PROOF. U, W and $U \cap W$ are finite dimensional by Lemma 4.2.17. Set $k := \dim(U)$, $l := \dim(W)$ and $m := \dim(U \cap W)$. Let $\{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$ be an unordered basis of $U \cap W$. Since $\mathbf{z}_1, \ldots, \mathbf{z}_m$ are linearly independent elements contained both in U and W, by Corollary 4.2.16 there are unordered bases

$$\{\mathbf{z}_1, \ldots, \mathbf{z}_m, \mathbf{u}_{m+1}, \ldots, \mathbf{u}_k\} \text{ and } \{\mathbf{z}_1, \ldots, \mathbf{z}_m, \mathbf{w}_{m+1}, \ldots, \mathbf{w}_l\},$$

of U and W respectively, that contain $\{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$. We prove that

$$\{\mathbf{z}_1, \ldots, \mathbf{z}_m, \mathbf{u}_{m+1}, \ldots, \mathbf{u}_k, \mathbf{w}_{m+1}, \ldots, \mathbf{w}_l\}$$

is a basis for $U + W$. Assume $\mathbf{u} + \mathbf{w}$ is an element of $U + W$ with $\mathbf{u} \in U$ and $\mathbf{w} \in W$. Since $\mathbf{u} \in U$ there are scalars $a_1, \ldots, a_m, a_{m+1}, \ldots a_l$ such that

$$u = a_1\mathbf{z}_1 + \cdots + a_m\mathbf{z}_m + a_{m+1}\mathbf{u}_{m+1} + \cdots + a_k\mathbf{u}_k \tag{4.15}$$ {u}

and, since $\mathbf{w} \in W$ there are scalars $b_1, \ldots, b_m, b_{m+1}, \ldots b_k$ such that

$$w = b_1\mathbf{z}_1 + \cdots + b_m\mathbf{z}_m + b_{m+1}\mathbf{w}_{m+1} + \cdots + b_l\mathbf{w}_l. \tag{4.16}$$ {w}

Adding both sides of Equations (4.15) and (4.16), we get

$$\begin{aligned} u + w &= a_1\mathbf{z}_1 + \cdots + a_m\mathbf{z}_m + a_{m+1}\mathbf{u}_{m+1} + \cdots + a_k\mathbf{u}_k \\ &+ b_1\mathbf{z}_1 + \cdots + b_m\mathbf{z}_m + b_{m+1}\mathbf{w}_{m+1} + \cdots + b_l\mathbf{w}_l \\ &= (a_1 + b_1)\mathbf{z}_1 + \cdots + (a_m + b_m)\mathbf{z}_m \\ &+ a_{m+1}\mathbf{u}_{m+1} + \cdots + a_k\mathbf{u}_k \\ &+ b_{m+1}\mathbf{w}_{m+1} + \cdots + b_l\mathbf{w}_l \\ &\in \langle \mathbf{z}_1, \ldots, \mathbf{z}_m, \mathbf{u}_{m+1}, \ldots, \mathbf{u}_k, \mathbf{w}_{m+1}, \ldots, \mathbf{w}_l \rangle. \end{aligned}$$

There remains to prove that the elements

$$\mathbf{z}_1, \ldots, \mathbf{z}_m, \mathbf{u}_{m+1}, \ldots, \mathbf{u}_k, \mathbf{w}_{m+1}, \ldots, \mathbf{w}_l$$

are linearly independent. Assume $c_1, \ldots, c_m, d_{m+1}, \ldots, d_k, e_{m+1} \ldots, e_l$ are scalars such that

$$0_V = c_1\mathbf{z}_1 + \cdots + c_m\mathbf{z}_m + d_{m+1}\mathbf{u}_{m+1} + \cdots + d_k\mathbf{u}_k + e_{m+1}\mathbf{w}_{m+1} + \cdots + e_l\mathbf{w}_l. \tag{4.17}$$ {lini}

Then we can rewrite Equation 4.17 as follows

$$c_1\mathbf{z}_1 + \cdots + c_m\mathbf{z}_m + d_{m+1}\mathbf{u}_{m+1} + \cdots + d_k\mathbf{u}_k = (-e_{m+1})\mathbf{w}_{m+1} + \cdots + (-e_l)\mathbf{w}_l. \tag{4.18}$$ {limi2}

Now note that the member on the left of Equation (4.18) is contained in U, while the member on the right is contained in W, since they are equal they both belong to $U \cap W$ so they are a linear combination of the vectors $\mathbf{z}_1, \ldots, \mathbf{z}_m$. Since $\{\mathbf{z}_1, \ldots, \mathbf{z}_m, \mathbf{u}_{m+1}, \ldots, \mathbf{u}_k\}$ is a basis for U, uniqueness of coefficients (Lemma 4.2.9*(2)*), gives

$$d_{m+1} = \cdots = d_k = 0$$

and Equation (4.18) becomes

$$c_1\mathbf{z}_1+\cdots+c_m\mathbf{z}_m+0\mathbf{u}_{m+1}+\cdots+0\mathbf{u}_k = 0\mathbf{z}_1+\cdots+0\mathbf{z}_m+(-e_{m+1})\mathbf{w}_{m+1}+\cdots+(-e_l)\mathbf{w}_l.$$

Again, since $\{\mathbf{z}_1,\ldots,\mathbf{z}_m,\mathbf{w}_{m+1},\ldots,\mathbf{w}_l\}$ is a basis for W, uniqueness of coefficients forces

$$c_1 = \cdots = c_m = e_{m+1} = \cdots = e_l = 0$$

So the unique linear combination equal to 0_V of the vectors

$$\mathbf{z}_1,\ldots,\mathbf{z}_m,\mathbf{u}_{m+1},\ldots,\mathbf{u}_k,\mathbf{w}_{m+1},\ldots,\mathbf{w}_l$$

is the one with all coefficients equal to 0, whence, by Lemma 4.2.9*(3)*, these vectors are linearly independent. So

$$\{\mathbf{z}_1,\ldots,\mathbf{z}_m,\mathbf{u}_{m+1},\ldots,\mathbf{u}_k,\mathbf{w}_{m+1},\ldots,\mathbf{w}_l\}$$

is an unordered basis of $V+W$ whence

$$\dim(W+W) = k+(l-m) = (k+l)-m = (\dim(U)+\dim(W))-\dim(U\cap W).$$

■

4.3 Linear maps

Linear maps are the "interesting" maps between vector spaces. In this section we fix two vector spaces V and W. A *linear map* between V and W is a

- L_1 homomorphism of groups $\phi\colon V\longrightarrow W$ such that
- L_2 for every scalar a and every vector $\mathbf{v}\in V$, $\phi(a\mathbf{v}) = a\phi(\mathbf{v})$

Linear maps are also called *homomorphisms* (of vector spaces) and bijective linear maps are called *isomorphisms* (of vector spaces) the set of all linear maps between V and W will be denoted by $Hom(V,W)$. A linear map between V and itself is called an *endomorphism* of the vector space V and we'll denote by $End(V)$ the set of all endomorphisms of V and by $GL(V)$ the set of all *automorphisms* (i.e. bijective endomorphisms) of V. In the following lemma the results for linear maps corresponding to those for group homomorphisms proved in Subsection 3.2.5.

{alllin} **Lemma 4.3.1** *Let V and W be two vector spaces and $f\in Hom(V,W)$*

1. *the composition of two linear maps is a linear map;*
2. *the identity map $id_V\colon V\longrightarrow V$ is an automorphism of V;*
3. *the inverse function of an isomorphism of vector spaces is linear, hence an isomorphism of vector spaces;*

4. *$(GL(V), \circ)$ is a group*

5. *$f(0_V) = 0_W$;*

6. *if $\mathbf{v} \in V$, then $f(-\mathbf{v}) = -f(\mathbf{v})$;*

7. *if U is a subspace of V then $f(U)$ is a subspace of W;*

8. *if Z is a subspace of W then $f^{-1}(Z)$ is a subspace of V;*

9. *f is injective if and only if $\ker(f) = \{0_V\}$.*

PROOF. Let $g\colon W \to Y$ be a linear map. By Lemma 3.2.12, $g \circ f$ is a homomorphism of groups. Suppose $a \in \mathbb{R}$ and $\mathbf{v} \in V$, then, since f and g are linear, by L2,

$$(g \circ f)(a\mathbf{v}) = g(f(a\mathbf{v})) = g(af(\mathbf{v})) = ag(f(\mathbf{v})) = a(g \circ f)(\mathbf{v}),$$

proving *(1)*. Similarly, by Lemma 3.2.13, id_V is a homomorphism of groups and, for every $a \in \mathbb{R}$ and $\mathbf{v} \in V$,

$$id_V(av) = av = a\, id_V(v),$$

proving *(2)*. The proof of *(3)* is similar and left as an exercise. Assertion *(4)* follows immediately from *(1)*, *(2)*, and *(3)*. Assertions *(5)* and *(6)* follow from Lemmas 3.2.16 and 3.2.17, which are the correspondent results for groups (note that we use here the additive notation for vector spaces). By Lemma 3.2.18 $f(U)$ is a subgroup of W but it is also closed under scalar multiples, for if $a \in \mathbb{R}$ and $f(\mathbf{u}) \in f(U)$ with $\mathbf{u} \in U$, then $a\mathbf{u} \in U$, since U is a subspace, whence, by L2,

$$af(\mathbf{u}) = f(a\mathbf{u}) \in f(U),$$

so $f(U)$ is a subspace. The proof of *(8)* is similar and is left as an exercise. Finally, *(9)* follows by Lemma 3.2.21. ∎

The group $GL(V)$ is called the *general linear group* of V. It is a very important group (though we won't see why) and it can be considered as the group of symmetries of the vector space V, that is the group of maps that "preserve the structure" of V.

4.3.1 Bases and linear maps

As in the previous subsection, we assume that V and W are vector spaces and $f\colon V \longrightarrow W$ is a linear map. An obvious but central fact about linear maps is that they preserve linear combinations:

{applinco}

Lemma 4.3.2 *Let $a_1, \ldots, a_k$ be scalars and let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be vectors of V. Then*

$$f(a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_k\mathbf{v}_k) = a_1 f(\mathbf{v}_1) + a_2 f(\mathbf{v}_2) + \cdots + a_k f(\mathbf{v}_k).$$

PROOF. Since f is a homomorphism of groups,

$$f(a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_k\mathbf{v}_k) = f(a_1\mathbf{v}_1) + f(a_2\mathbf{v}_2) + \cdots + f(a_k\mathbf{v}_k)$$

and, since $f(a_i\mathbf{v}_i) = a_i f(\mathbf{v}_i)$,

$$f(a_1\mathbf{v}_1) + f(a_2\mathbf{v}_2) + \cdots + f(a_k\mathbf{v}_k) = a_1 f(\mathbf{v}_1) + a_2 f(\mathbf{v}_2) + \cdots + a_k f(\mathbf{v}_k).$$

■

A consequence of Lemma 4.3.2 is that linear maps are "controlled" by the images of the sets of generators of their domains. Precisely

{genapplin}

Lemma 4.3.3 *Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ be a set of generators for V and $g\colon V \longrightarrow W$ a linear map such that, for every $i \in \{1, \ldots, k\}$,*

{unigen}
$$f(\mathbf{v}_i) = g(\mathbf{v}_i). \tag{4.19}$$

Then $f = g$.

PROOF. Since f and g have the same domain, we just have to prove that, for every $\mathbf{v} \in V$,

{unigen1}
$$f(\mathbf{v}) = g(\mathbf{v}) \tag{4.20}$$

So let $\mathbf{v} \in V$. Since V is linearly spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$, there are scalars $a_1, \ldots, a_k$ such that

{linco}
$$\mathbf{v} = a_1\mathbf{v}_1 + \cdots + a_k\mathbf{v}_k. \tag{4.21}$$

Then, by Lemma 4.3.2 and Equation (4.21),

$$\begin{aligned} f(\mathbf{v}) &= f(a_1\mathbf{v}_1 + \cdots + a_k\mathbf{v}_k) \\ &= a_1 f(\mathbf{v}_1) + \cdots + a_k f(\mathbf{v}_k) \\ &= a_1 g(\mathbf{v}_1) + \cdots + a_k g(\mathbf{v}_k) \\ &= g(a_1\mathbf{v}_1 + \cdots + a_k\mathbf{v}_k) = g(\mathbf{v}) \end{aligned}$$

■

We have two important remarks:

1. The above proof shows two golden rules for dealing with linear maps and generators:
   (a) whenever you have a set of generators $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ of a vector space V, you can always write any vector $\mathbf{v}$ as a linear combination of the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$;
   (b) whenever you have a linear map and a linear combination of vectors, you can always use Equation (4.21).

In most proofs of this section we shall constantly use those arguments (and in most exercises, using them will bring you closer to the solution). We'll point out when we use them only at the beginning, leaving to you as an exercise to spot them in the remaining proofs.

2. A consequence of Lemma 4.3.3 is that we cannot choose freely the images $f(\mathbf{v}_1), \dots, f(\mathbf{v}_k)$ of the elements of the generating set $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ of V. More precisely, there is no guarantee that for any set $\{\mathbf{w}_1, \dots, \mathbf{w}_k\}$ of elements in W there is a linear map that sends $\mathbf{v}_i$ to $\mathbf{w}_i$ for every $i \in \{1, \dots, n\}$. Indeed if one of the vectors $\mathbf{v}_1, \dots, \mathbf{v}_k$ is a linear combination of the others, its image has to be the corresponding linear combination of the images of the others: if, e.g.,

$$f\colon \mathbb{R}^3 \longrightarrow \mathbb{R}^2$$

is a linear map such that

$$f\left(\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \text{ and } f\left(\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

then f must necessarily send the vector, say, $(1, 2, 0)^t$ to $(3, 5)^t$, since , by Lemma 4.3.2,

$$\begin{aligned} f\left(\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}\right) &= f\left(1\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 2\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\right) \\ &= 1f\left(\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}\right) + 2f\left(\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\right) \\ &= 1\begin{pmatrix} 1 \\ 1 \end{pmatrix} + 2\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}. \end{aligned}$$

As we see, the obstruction that does not allow us to choose freely the images of $\mathbf{v}_1, \dots, \mathbf{v}_k$, arises from the fact that these vectors can be linearly dependent. So it is natural to ask whether, removing that obstruction, i.e. requiring that $\mathbf{v}_1, \dots, \mathbf{v}_k$ are linearly independent allows us to choose their images freely. Well this is true by the next theorem and has enormous consequences, for it will lead us to a full calssification of the linear maps between two finite dimensional vector spaces.

Theorem 4.3.4 [EXTENSION BY LINEARITY] *Let V and W be vector spaces, $(\mathbf{v}_1, \dots, \mathbf{v}_n)$ a basis of V and $\mathbf{w}_1, \dots, \mathbf{w}_n$ be vectors in W. Then there is a unique linear map $f\colon V \longrightarrow W$ such that $f(\mathbf{v}_i) = \mathbf{w_i}$ for every $i \in \{1, \dots, n\}$.* {extlin}

PROOF. Since $(\mathbf{v}_1, \dots, \mathbf{v}_n)$ is a basis of V, by Theorem 4.2.12, for every $\mathbf{v} \in V$ there is a unique n-tuple of scalars $(a_1, \dots, a_n)$ such that

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

We may therefore define a map $f\colon V \to W$ by setting

$$f(a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n) := a_1\mathbf{w}_1 + \cdots + a_n\mathbf{w}_n.$$

We prove that f is linear. Indeed, if $(b_1, \ldots, b_n)$ is another n-tuple of scalars, then

$$\begin{aligned} f((a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n) + (b_1\mathbf{v}_1 + \cdots + b_n\mathbf{v}_n)) &= f((a_1\mathbf{v}_1 + (b_1\mathbf{v}_1) + \cdots + (a_n\mathbf{v}_n + b_n\mathbf{v}_n)) \\ &= f((a_1 + b_1)\mathbf{v}_1 + \cdots + (a_n + b_n)\mathbf{v}_n) \\ &= (a_1 + b_1)\mathbf{w}_1 + \cdots + (a_n + b_n)\mathbf{w}_n \\ &= (a_1\mathbf{w}_1 + b_1\mathbf{w}_1) + \cdots + (a_n\mathbf{w}_n + b_n\mathbf{w}_n) \\ &= (a_1\mathbf{w}_1 + \cdots + a_n\mathbf{w}_n) + (b_1\mathbf{w}_1 + \cdots + b_n\mathbf{w}_n) \\ &= f(a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n) + f(b_1\mathbf{v}_1 + \cdots + b_n\mathbf{v}_n). \end{aligned}$$

Thus f preserves the sum of vectors. Similarly, if k is a scalar, then

$$\begin{aligned} f(k(a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n)) &= f((ka_1)\mathbf{v}_1 + \cdots + (ka_n)\mathbf{v}_n) \\ &= (ka_1)\mathbf{w}_1 + \cdots + (ka_n)\mathbf{w}_n \\ &= k(a_1\mathbf{w}_1 + \cdots + a_n\mathbf{w}_n) \\ &= kf(a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n)) \end{aligned}$$

So f preserves also the products by scalars, hence f is linear. By definition

$$f(\mathbf{v}_i) = \mathbf{w_i}$$

for every $i \in \{1, \ldots, n\}$. Finally, the uniquenes of f follows from Lemma 4.3.3. ∎

{solitex}

Theorem 4.3.5 *Let $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ be a basis for a vector space V, let W be another vector space, and let $f\colon V \longrightarrow W$ be a linear map. Then*

1. *f is surjective if and only if $f(\mathbf{v}_1), \ldots, f(\mathbf{v}_n)$ span W;*
2. *f is injective if and only if $f(\mathbf{v}_1), \ldots, f(\mathbf{v}_n)$ are linearly independent;*
3. *f is an isomorphism of vector spaces if and only if $f(\mathbf{v}_1), \ldots, f(\mathbf{v}_n)$ is an unordered basis for W.*

Proof. For the first assertion, assume first that f is surjective and let $\mathbf{w} \in W$. Then there is an element $\mathbf{v} \in V$ such that

{v=w}
$$f(\mathbf{v}) = \mathbf{w}. \tag{4.22}$$

Since $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an unordered basis for V and $\mathbf{v} \in \mathbf{V}$, there are scalars $a_1, \ldots, a_n$ such that

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

So, if we apply f to the above equation, we get

$$\begin{aligned}\mathbf{w} &= f(\mathbf{v} = f(a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n)\\ &= a_1 f(\mathbf{v}_1) + \cdots + a_n f(\mathbf{v}_n) \in \langle(\mathbf{v}_1), \ldots, f(\mathbf{v}_n)\rangle.\end{aligned}$$

Conversely, assume $f(\mathbf{v}_1), \ldots, f(\mathbf{v}_n)$ span W. We have to prove that, for every $\mathbf{w} \in \mathbf{W}$ there is an element $\mathbf{v} \in \mathbf{V}$ that satisfies Equation (4.22). Since $\mathbf{w} \in W$ and W is linearly spanned by $f(\mathbf{v}_1), \ldots, f(\mathbf{v}_n)$, there are scalars $b_1, \ldots, b_n$ such that

$$\mathbf{w} = b_1 f(\mathbf{v}_1) + \cdots + b_n f(\mathbf{v}_n) = f(b_1\mathbf{v}_1 + \cdots + b_n\mathbf{v}_n),$$

so the assertion follows taking $\mathbf{v} := b_1\mathbf{v}_1 + \cdots + b_n\mathbf{v}_n$.

For the second assertion, assume f is injective and let $c_1, \ldots, c_n$ be scalars such that

$$0_W = c_1 f(\mathbf{v}_1) + \cdots + c_n f(\mathbf{v}_n). \tag{4.23}$$ {kerlin}

We prove that

$$c_1 = c_2 = \cdots = c_n = 0$$

Indeed, since

$$c_1 f(\mathbf{v}_1) + \cdots + c_n f(\mathbf{v}_n) = f(c_1\mathbf{v}_1) + \cdots + c_n\mathbf{v}_n),$$

Equation (4.23), implies that

$$c_1\mathbf{v}_1) + \cdots + c_n\mathbf{v}_n \in \ker f.$$

Since f is injective, by Lemma 4.3.1*(9)*, $\ker(f) = \{0_V\}$, whence

$$c_1\mathbf{v}_1) + \cdots + c_n\mathbf{v}_n = 0_V.$$

But now $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an unordered basis for V, so $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent, whence $c_1 = c_2 = \cdots = c_n = 0$, and the result follows by Lemma 4.2.9*(3)*. Conversely, assume $f(\mathbf{v}_1), \ldots, f(\mathbf{v}_n)$ are linearly independent and let $\mathbf{v} \in \ker(f)$. Since $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an unordered basis for V, there are scalars $d_1, \ldots, d_n$, such that

$$v = d_1\mathbf{v}_1 + \cdots + d_n\mathbf{v}_n.$$

Applying f to both sides and taking into account that $\mathbf{v} \in \ker(f)$, we get

$$0_W = f(v) = f(d_1\mathbf{v}_1 + \cdots + d_n\mathbf{v}_n) = d_1 f(\mathbf{v}_1) + \cdots + d_n f(\mathbf{v}_n),$$

whence $d_1 = d_2 = \cdots = d_n = 0$, since $f(\mathbf{v}_1), \ldots, f(\mathbf{v}_n)$ are linearly independent. But then $v = 0_V$ and so f is injective by Lemma 4.2.9*(9)*.

The last assertion follows immediately from the first two. ∎

The next two results are quite useful. In both cases the proofs use a strategy similar to that used for proving Grassmann's Theorem (Theorem 4.2.18): namely we start with a basis of a subspace and complete it to a basis of the whole space.

Corollary 4.3.6 [The Witt Property for Vector Spaces] *Let U and W be subspaces of a vector space V and let $f\colon U \to W$ be an isomorphism of vector spaces. Then There are automorphisms of vector spaces $\bar{f}\colon V \to V$ that extend f (i.e. such that $\bar{f}(\mathbf{u}) = f(\mathbf{u})$ for every $\mathbf{u} \in U$.*

{Wittvs}

Proof. Let $(\mathbf{u}_1, \ldots, \mathbf{u}_r)$ be a basis of U and set, for every $i \in \{1, \ldots, r\}$,

$$\mathbf{w}_i := f(\mathbf{v}_i)$$

Since f is an isomorphism of vector spaces, by Theorem 4.3.5 $(\mathbf{w}_1, \ldots, \mathbf{w}_r)$ is a basis of W Since $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are linearly independent elements of V by the Bodysntacther's Theorem (Theorem 4.2.15), we can complete the set $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ to a basis

$$(\{\mathbf{u}_1, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \ldots, \mathbf{u}_n\})$$

of V. Similarly we can complete the set $\{\mathbf{w}_1, \ldots, \mathbf{w}_r\}$ to another basis

$$(\{\mathbf{w}_1, \ldots, \mathbf{w}_r, \mathbf{w}_{r+1}, \ldots, \mathbf{w}_n\})$$

of V. By Theorem 4.3.4, there is a unique linear map $\bar{f}\colon V \to V$ such that

$$\bar{f}(\mathbf{u}_i) = \mathbf{w_i}$$

for every $i \in \{1, \ldots, n\}$ and, again by Theorem 4.3.4, this map extends f. ∎

{nullrang}

Theorem 4.3.7 *Let V and W be vector spaces and $f\colon V \longrightarrow W$ a linear map. Then*

{nullrang}

$$\dim(V) = \dim(\ker(f)) + \dim(f(V)). \tag{4.24}$$

Proof. Let

$$(\mathbf{v}_1, \ldots, \mathbf{v}_k)$$

be a basis of $\ker(f)$. Then, by the Bodysntacther's Theorem (Theorem 4.2.15), we can complete the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ to a basis

$$\{\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\}$$

of V. So that

$$\dim(\ker(f)) = k \text{ and } \dim(V) = n.$$

Hence we just need to prove that

$$\{f(\mathbf{v}_{k+1}), \ldots, f(\mathbf{v}_n)\}$$

is an unordered basis of $f(V)$ with $n-k$ elements. Indeed $\langle f(\mathbf{v}_{k+1}), \ldots, f(\mathbf{v}_n)\rangle = f(V)$, for if $f(\mathbf{v}) \in f(V)$, with $\mathbf{v} \in V$, then, since $\{\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\}$ is an unordered basis of V, there are scalars $a_1, \ldots, a_k, a_{k+1}, \ldots, a_n$ such that

$$\mathbf{v} = a_1\mathbf{v}_1 + \cdots + a_k\mathbf{v}_k + a_{k+1}\mathbf{v}_{k+1} + \cdots + a_n\mathbf{v}_n.$$

Applying f to both sides and taking into account that $f(\mathbf{v}_1) = f(\mathbf{v}_2) = \cdots = f(\mathbf{v}_k) = 0_w$, since these elements are contained in $\ker(f)$, we get

$$\begin{aligned} f(\mathbf{v}) &= f(a_1\mathbf{v}_1 + \cdots + a_k\mathbf{v}_k + a_{k+1}\mathbf{v}_{k+1} + \cdots + a_n\mathbf{v}_n) \\ &= a_1 f(\mathbf{v}_1) + \cdots + a_k f(\mathbf{v}_k) + a_{k+1} f(\mathbf{v}_{k+1}) + \cdots + a_n f(\mathbf{v}_n) \\ &= a_{k+1} f(\mathbf{v}_{k+1}) + \cdots + a_n f(\mathbf{v}_n) \in \langle f(\mathbf{v}_{k+1}), \ldots, f(\mathbf{v}_n)\rangle. \end{aligned}$$

To conclude we are left to show that $f(\mathbf{v}_{k+1}), \ldots, f(\mathbf{v}_n)$ are linearly independent. We use again Lemma 4.2.9*(9)*: assume $b_{k+1}, \ldots, b_n$ are scalars such that

$$b_{k+1} f(\mathbf{v}_{k+1}) + \cdots + b_n f(\mathbf{v}_n) = 0_W.$$

Since

$$b_{k+1} f(\mathbf{v}_{k+1}) + \cdots + b_n f(\mathbf{v}_n) = f(b_{k+1}\mathbf{v}_{k+1} + \cdots + b_n\mathbf{v}_n),$$

this implies that

$$b_{k+1}\mathbf{v}_{k+1} + \cdots + b_n\mathbf{v}_n \in \ker(f),$$

so it is a linear combination of the elements of the unordered basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ of $\ker(f)$, that is there are scalars $b_1, \ldots, b_k$ such that

$$b_{k+1}\mathbf{v}_{k+1} + \cdots + b_n\mathbf{v}_n = b_1\mathbf{v}_1 + \cdots + b_k\mathbf{v}_k.$$

Now we can rewrite the last equation as follows:

$$b_1\mathbf{v}_1 + \cdots + b_k\mathbf{v}_k + (-b_{k+1})\mathbf{v}_{k+1} + \cdots + (-b_n)\mathbf{v}_n.$$

Since $\{\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\}$ is a basis of V, this forces

$$b_1 = b_2 = \cdots = b_k = b_{k+1} = \cdots = b_n = 0,$$

so the unique linear combination of the vectors $f(\mathbf{v}_{k+1}), \ldots, f(\mathbf{v}_n)$ is the one with all coefficients equal to 0, that is these vectors are also linearly independent. ■

The dimension of the kernel of a linear map f is called the *rank* of f and is denoted by $rk(f)$.

{insu}

Corollary 4.3.8 [The Pigeonhole Principle for Vector Spaces] *Let V and W be a finitely dimensional vector spaces with $\dim(V) = \dim(W)$ and $\phi\colon V \to W$ a linear map. Then the following assertions are equivalent:*

1. *ϕ is injective;*
2. *ϕ is surjective;*
3. *ϕ is bijective.*

4.3.2 The space $Hom(V,W)$)

Let V and W be vector spaces. Denote by $Hom(V,W)$ the set of all linear maps from V to W. This set can be given a structure of a vector space defining *pointwise* the sum of two linear maps and the product of a linear map by a scalar. Namely let

$$f\colon V \to W \text{ and } g\colon V \to W$$

be two linear maps and $k \in \mathbb{R}$. Define, for every $v \in V$,

$$f+g\colon V \to W$$

to be the map that sends $\mathbf{v}$ to $f(\mathbf{v})+g(\mathbf{v})$ (*pointwise addition of funtions*) and

$$kf\colon V \to W$$

to be the map that sends $\mathbf{v}$ to $k(f(\mathbf{v}))$ (*pointwise multiplication of a function by a scalar*).

{Hom-1}

Lemma 4.3.9 *Let $f,g \in Hom(V,W)$, and $k \in \mathbb{R}$, then $f+g$ and kf are also in $Hom(V,W)$.*

Proof. Since, for every $\mathbf{v} \in V$ $f(\mathbf{v})$ and $g(\mathbf{v})$ are elements of W, also $(f+g)(\mathbf{v}) = f(\mathbf{v})+g(\mathbf{v})$ and $(kf)(\mathbf{v}) = k(f(\mathbf{v}))$ are elements of W. Thus $f+g$ and kf are functions from V to W. We prove that they are linear. Indeed, if $\mathbf{v}_1$ and $\mathbf{v}_2$ are in V and $k \in \mathbb{R}$, then, by the definition and the linearity of f and g,

$$\begin{aligned}(f+g)(\mathbf{v}_1+\mathbf{v}_2) &= f(\mathbf{v}_1+\mathbf{v}_2)+g(\mathbf{v}_1+\mathbf{v}_2)\\ &= f(\mathbf{v}_1)+f(\mathbf{v}_2)+g(\mathbf{v}_1)+g(\mathbf{v}_2)\\ &= f(\mathbf{v}_1)+g(\mathbf{v}_1)+f(\mathbf{v}_2)+g(\mathbf{v}_2)\\ &= (f+g)(\mathbf{v}_1)+(f+g)(\mathbf{v}_2)\end{aligned}$$

and

$$\begin{aligned}(kf)(\mathbf{v}_1+\mathbf{v}_2) &= k(f(\mathbf{v}_1+\mathbf{v}_2)) = k(f(\mathbf{v}_1)+f\mathbf{v}_2))\\ &= k(f(\mathbf{v}_1))+k(f\mathbf{v}_2)) = (kf)(\mathbf{v}_1)+(kf)\mathbf{v}_2)\end{aligned}$$

so $f+g$ and kf are linear. ∎

{Hom}

Theorem 4.3.10 *With the above defined pointwise addition and multiplication by scalars, $Hom(V,W)$ is a vector space.*

Proof. We first prove that $Hom(V,W)$ is an abelian group with respect to the pointwise addition. So we have to prove that

1. the pointwise addition is associative,
2. the pointwise addition is commutative,

3. there exists a zero element in $HomV, W$,

4. for every $f \in Hom(V, W)$ there exists an opposite.

We first prove that the pointwise addition is associative, i.e. that $(f+g)+h = f+(g+h)$ for every f, g, and h in $Hom(V, W)$. Indeed both $(f+g)+h$ and $f+(g+h)$ are linear maps from V to W, so we only need to prove that they coincide pointwise, i.e., that, for every $\mathbf{v} \in V$

$$(f+g)+h)(\mathbf{v}) = (f+(g+h))(\mathbf{v}) \tag{4.25}$$ {Hom1}

Now, by the definition of pointwise addition and the associativity of the addition in W,

$$\begin{aligned}(f+g)+h)(\mathbf{v}) &= (f+g)(\mathbf{v})+h(\mathbf{v}) = (f(\mathbf{v})+g(\mathbf{v}))+h(\mathbf{v}) \\ &= f(\mathbf{v})+(g(\mathbf{v})+h(\mathbf{v})) = f(\mathbf{v})+(g+h)(\mathbf{v}) \\ &= (f+(g+h))(\mathbf{v})\end{aligned}$$

proving (4.25).

Similarly we prove that the pointwise addition is commutative, i.e., that for every f, g in $Hom(V, W)$ and every $\mathbf{v} \in V$,

$$(f+g)(\mathbf{v}) = (g+f)(\mathbf{v}) \tag{4.26}$$ {Hom2}

Indeed, by the definition of pointwise addition and commutativity of addition in W,

$$(f+g)(\mathbf{v}) = f(\mathbf{v})+g(\mathbf{v}) = g(\mathbf{v})+f(\mathbf{v}) = (g+f)(\mathbf{v}),$$

proving (4.26).

Let $\overline{0}$ be the map that sends every vector of V to the zero vector of W. It is immediate to check that $\overline{0} \in Hom(V, W)$. Moreover for every $f \in Hom(V, W)$ and every $\mathbf{v} \in V$,

$$(\overline{0}+f)(\mathbf{v}) = \overline{0}(\mathbf{v})+f(\mathbf{v}) = 0_W + f(\mathbf{v}) = f(\mathbf{v}).$$

Thus $\overline{0}+f = f$ and, by (4.26) also $f+\overline{0} = f$, so $\overline{0}$ is the zero element of $Hom(V, W)$.

We now prove the existence of an opposite for every $f \in Hom(V, W)$. Now let $f \in Hom(V, W)$ and define $-f$ to be the map from V to W that sends a vector $\mathbf{v}$ of V to the vector $-(f(\mathbf{v}))$ of W. We leave the reader the easy task to prove that $-f$ is linear and that $f+(-f) = (-f)+f = \overline{0}$. So $(-f)$ is the opposite of f, for every $f \in Hom(V, W)$. So

$$(Hom(V, W), +) \textit{ is an abelian group}$$

We finally prove that the pointwise multiplication by scalars satisfies the axioms V_1-V_4 of a vector space, i.e., that, for every $f, g \in Hom(V, W)$ and every $h, k \in \mathbb{R}$,

V_1: $1f = f$;

V_2: $(h+k)f = hf + kf$;

V_3: $(hk)f = h(kf)$;

V_4: $h(f+g) = hf + hg$.

By Lemma 4.3.9 all the maps above are linear maps from V to W, so, again, they are equal if and only if they coincide pointwise. So let $\mathbf{v} \in V$. Then, by the definition of pointwise multiplication by scalars and since V_1-V_4 hold for the vector space W,

$$(1f)(\mathbf{v}) = 1(f(\mathbf{v})) = f(v),$$

proving V_1,

$$((h+k)f)(\mathbf{v}) = (h+k)(f(\mathbf{v})) = h(f(\mathbf{v})) + k(f(\mathbf{v})) = (hf+kf)(\mathbf{v}),$$

proving V_2,

$$((hk)f)(\mathbf{v}) = (hk)(f(\mathbf{v})) = h(kf(\mathbf{v})) = (h(kf))(\mathbf{v}),$$

proving V_3, and finally

$$\begin{aligned}(h(f+g))(\mathbf{v}) &= h((f+g)(\mathbf{v})) = h(f(\mathbf{v})) + g(\mathbf{v})) = h(f(\mathbf{v})) + h(g(\mathbf{v})) \\ &= (hf)(\mathbf{v}) + (hg)(\mathbf{v}) = (hf+kf)(\mathbf{v}),\end{aligned}$$

proving V_4. ∎

Let

$$(\mathbf{v}_1, \ldots, \mathbf{v}_n) \text{ be a basis for } V$$

and

$$(\mathbf{w}_1, \ldots, \mathbf{w}_m) \text{ be a basis for } W.$$

By Theorem 4.3.4 for every $i \in \{1, \ldots, n\}$ and every $j \in \{1, \ldots, m\}$, there is a unique linear map $f_{i,j}$ such that

$$f_{i,j}(\mathbf{v}_i) = \mathbf{w}_j \text{ and } f_{i,j}(\mathbf{v}_k) = 0_W, \text{ if } k \neq i.$$

{Hom3}

Theorem 4.3.11 *With the above notations, the linear maps $f_{i,j}$'s where $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$ are linearly independent and generate $Hom(V, W)$. In particular*

$$\dim(Hom(V, W)) = \dim(V) \times \dim(W).$$

Proof. Assume

$$\begin{aligned}\overline{0} &= a_{1,1}f_{1,1} + a_{1,2}f_{1,2} + \cdots + a_{1,m}f_{1,m} \\ &+ a_{2,1}f_{2,1} + a_{2,2}f_{2,2} + \cdots + a_{2,m}f_{2,m} \\ &\vdots \qquad\qquad\qquad \vdots \\ &+ a_{n,1}f_{n,1} + a_{n,2}f_{n,2} + \cdots + a_{n,m}f_{n,m}.\end{aligned}$$

Evaluating the second member on $\mathbf{v}_i$ for every $i \in \{1, \ldots, n\}$, we get the equalities

$$\left\{\begin{array}{rcccl} 0_W & = & a_{1,1}f_{1,1}(\mathbf{v}_1) + \cdots + a_{1,m}f_{1,m}(\mathbf{v}_n) & = & a_{1,1}\mathbf{w}_1 + \cdots, +a_{1,m}\mathbf{w}_n \\ 0_W & = & a_{2,1}f_{2,1}(\mathbf{v}_1) + \cdots + a_{2,m}f_{2,m}(\mathbf{v}_n) & = & a_{2,1}\mathbf{w}_1 + \cdots, +a_{2,m}\mathbf{w}_n \\ & \vdots & \vdots & \vdots & \vdots \\ 0_W & = & a_{n,1}f_{n,1}(\mathbf{v}_1) + \cdots + a_{n,m}f_{n,m}(\mathbf{v}_n) & = & a_{n,1}\mathbf{w}_1 + \cdots, +a_{n,m}\mathbf{w}_n \end{array}\right.$$

which force all the $a_{i,j}$'s to be equal to 0, since $\mathbf{w}_1, \ldots, \mathbf{w}_m$ are linearly independent. Now let $f \in Hom(V, W)$. Then, since, for every $i \in \{1, \ldots, n\}$, $f(\mathbf{v}_i) \in W$ and $(\mathbf{w}_1, \ldots, \mathbf{w}_m)$ is a basis for W there are unique scalars $a_{i,j}$ such that

$$\begin{array}{rcl} f(\mathbf{v}_1) & = & b_{1,1}\mathbf{w}_1 + b_{2,1}\mathbf{w}_2 \cdots + b_{m,1}\mathbf{w}_m \\ f(\mathbf{v}_2) & = & b_{1,2}\mathbf{w}_1 + b_{2,2}\mathbf{w}_2 \cdots + b_{m,2}\mathbf{w}_m \\ \vdots & \vdots & \vdots \\ f(\mathbf{v}_n) & = & b_{1,n}\mathbf{w}_1 + b_{2,n}\mathbf{w}_2 \cdots + b_{m,n}\mathbf{w}_m \end{array}.$$

Let

$$\begin{array}{rcl} g & := & b_{1,1}f_{1,1} + b_{1,2}f_{1,2} + \cdots, +b_{1,m}f_{1,m} \\ & + & b_{2,1}f_{2,1} + b_{2,2}f_{2,2} + \cdots, +b_{2,m}f_{2,m} \\ & \vdots & \vdots \\ & + & b_{n,1}f_{n,1} + b_{n,2}f_{n,2} + \cdots, +b_{n,m}f_{n,m}. \end{array}$$

Then, by Lemma 4.3.9, g is a linear map from V to W and evaluating at $\mathbf{v}_i$ f and g we get $f(\mathbf{v}_i) = g(\mathbf{v}_i)$ for every $i \in \{1, \ldots, n\}$, which implies, by Theorem 4.3.4, $f = g$, since $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ is a basis for V. ∎

4.3.3 The dual space

Let, as before, V and W be vector spaces and $Hom(V, W)$ the space of linear maps between V and W.. The case in which W is $\mathbb{R}^1$ is of particular interest. In this case (1) is a basis since $\mathbb{R}^1$ for $1 \neq 0$, so 1 is linearly independent, and every element k of $\mathbb{R}^1$ is a scalar multiple of 1 ($k = k \cdot 1$), so 1 generate $\mathbb{R}^1$. The vector space $Hom(V, \mathbb{R}^1)$ is called the *dual* space of V and is usually denoted by V^*. By Theorem 4.3.11, V^* has the same dimension as V, and, given a basis

$$(\mathbf{v}_1, \ldots, \mathbf{v}_n)$$

of V, a basis of V^* is given by

$$(\mathbf{v}_1^*, \ldots, \mathbf{v}_n^*)$$

where, for every $i, j \in \{1, \ldots, n\}$,

$$(\mathbf{v}_i^*((\mathbf{v}_i) = 1 \text{ and } (\mathbf{v}_i^*((\mathbf{v}_j) = \text{ if } i \neq j.$$

(Note that, with the above notation, taking $\mathbf{w}_1 = 1$, $\mathbf{v}_i^* = f_{i,1}$). The basis $(\mathbf{v}_1^*, \dots, \mathbf{v}_n^*)$ of V^* is called the *dual basis* relative to the basis $(\mathbf{v}_1, \dots, \mathbf{v}_n)$ of V. The dual basis will turn to be extremely useful for making computations: e.g. if a vector $\mathbf{v}$ of V is written as a linear combinations of the basis vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$:

$$\mathbf{v} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots a_n\mathbf{v}_n$$

then, evaluating $\mathbf{v}_i$ at both sides of the above equation we get

$$a_i = \mathbf{v}_i^*(\mathbf{v})$$

for every $i \in \{1, \dots, n\}$, that is

{coeffici}
$$\mathbf{v} = \mathbf{v}_1^*(\mathbf{v})\mathbf{v}_1 + \mathbf{v}_2^*(\mathbf{v})\mathbf{v}_2 + \cdots \mathbf{v}_n^*(\mathbf{v})\mathbf{v}_n. \tag{4.27}$$

Now the dual space V^* is also a vector space hence, as such, it has all the rights to have a dual V^{**} (the *bidual* space of V. It turns out that V^{**} can be identified with V itself. Namely, for every $\mathbf{v} \in V$, let

$$\begin{array}{rccc} e_{\mathbf{v}}\colon & V^* & \to & \mathbb{R} \\ & f & \mapsto & f(\mathbf{v}) \end{array}$$

that is $e_{\mathbf{v}}$ is the map that sends any linear form $f\colon V \to \mathbb{R}$ to value f takes on the vector $\mathbf{v}$. Then $e_{\mathbf{v}}$ is linear, for, given f and g in V^* and $k \in \mathbb{R}$, we have

$$e_{\mathbf{v}}(f+g) = (f+g)(\mathbf{v}) = f(\mathbf{v}) + g(\mathbf{v}) = e_{\mathbf{v}}(f) + e_{\mathbf{v}}(g)$$

and

$$e_{\mathbf{v}}(kf) = (kf)(\mathbf{v}) = k(f(\mathbf{v})) = k(e_{\mathbf{v}}(f)).$$

So $e_{\mathbf{v}}$ is a linear map from V^* to $\mathbb{R}$, that is an element of the bidual V^{**} of V.

Theorem 4.3.12 *Let V be a vector space and $\mathbf{v} \in V$ then the map*

$$\begin{array}{rccc} e\colon & V & \to & V^{**} \\ & \mathbf{v} & \mapsto & e_{\mathbf{v}} \end{array}$$

is an isomorphism (i.e. a bijective linear map) between V and V^{**}.

Proof. We have to prove that, for every $\mathbf{v}_1$ and $\mathbf{v}_2$ in V and every $k \in \mathbb{R}$,

1. $e_{\mathbf{v}_1+\mathbf{v}_2} = e_{\mathbf{v}_1} + e_{\mathbf{v}_2}$

and

2. $e_{k\mathbf{v}_1} = ke_{\mathbf{v}_1}$.

Both members of the above equations are maps from V^* to $\mathbb{R}$, so we need to prove that they coincide pointwise,
that is, for every $\mathbf{v}^* \in V^*$

(1′) $(e_{\mathbf{v}_1+\mathbf{v}_2})(\mathbf{v}^*) = (e_{\mathbf{v}_1} + e_{\mathbf{v}_2})(\mathbf{v}^*)$

and

(2′) $e_{k\mathbf{v}_1}(\mathbf{v}^*) = (ke_{\mathbf{v}_1})(\mathbf{v}^*)$.

Now, by the definition of e and since $\mathbf{v}^*$ is linear, we have

$$\begin{aligned}(e_{\mathbf{v}_1+\mathbf{v}_2})(\mathbf{v}^*) &= \mathbf{v}^*(\mathbf{v}_1+\mathbf{v}_2) = \mathbf{v}^*(\mathbf{v}_1) + \mathbf{v}^*(\mathbf{v}_2)\\ &= e_{\mathbf{v}_1}(\mathbf{v}^*) + e_{\mathbf{v}_2}(\mathbf{v}^*) = (e_{\mathbf{v}_1} + e_{\mathbf{v}_2})(\mathbf{v}^*),\end{aligned}$$

giving (1′), and

$$(e_{k\mathbf{v}_1})(\mathbf{v}^*) = \mathbf{v}^*(k\mathbf{v}_1) = k\mathbf{v}^*(\mathbf{v}_1) = k((e_{k\mathbf{v}_1})(\mathbf{v}^*)) = (k(e_{k\mathbf{v}_1})(\mathbf{v}^*))$$

proving (2′). ∎

Note that, if

$$(\mathbf{v}_1, \ldots, \mathbf{v}_n)$$

is a basis for V,

$$(\mathbf{v}_1^*, \ldots, \mathbf{v}_n^*)$$

is the relative dual basis of V^*, and

$$(\mathbf{v}_1^{**}, \ldots, \mathbf{v}_n^{**})$$

is the dual basis of V^{**} relative to the basis $(\mathbf{v}_1^*, \ldots, \mathbf{v}_n^*)$, then, for every $i \in \{1, \ldots, n\}$,

$$\mathbf{v}_i^{**} = e_{\mathbf{v}_i} \qquad (4.28) \quad \texttt{\{bidual\}}$$

4.3.4 The annihilator

Let V be a vector space and V^* be its dual space. For every nonempty subset X of V define

$$X^\perp := \{\mathbf{v}_n^* \in V^* | \mathbf{v}_n^*(\mathbf{x}) = 0 \text{ for every } \mathbf{x} \in X\}$$

that is $X^\perp$ is the set of linear forms whose value at the elements of X is 0, that is the elements of V^* that *annihilate* all the elements of X. We'll call therefore $X^\perp$ the *annihilator* of X.

Lemma 4.3.13 *Let X be a non empty subset of a vector space V. Then $X^\perp$ is a subspace of V^**

PROOF. Let $\mathbf{u}^*$ and $\mathbf{v}^*$ be elements in $X^\perp$, so that

$$\mathbf{u}^*(\mathbf{x}) = \mathbf{v}^*(\mathbf{x}) = 0,$$

and let $k \in \mathbb{R}$, then, for every $\mathbf{x} \in X$,

$$(\mathbf{u}^* + \mathbf{v}^*)(\mathbf{x})) = \mathbf{u}^*(\mathbf{x}) + \mathbf{v}^*(\mathbf{x}) = 0 + 0 = 0,$$

so $\mathbf{u}^* + \mathbf{v}^* \in X^\perp$, and

$$(k\mathbf{v}^*)(\mathbf{x}) = k(\mathbf{v}^*(\mathbf{x})) = k0 = 0,$$

thus also $(k\mathbf{v}^*) \in X^\perp$, whence $X^\perp$ is a subspace of V^*. ∎

{dimort}

Theorem 4.3.14 *Let U be a subspace of a vector space V. Then*

{dimot1}
$$\dim U^\perp = n - r = \dim(V) - \dim(U). \tag{4.29}$$

PROOF. Let r be the dimension of U and let

$$(\mathbf{v}_1, \ldots, \mathbf{v}_r)$$

be a basis of U, then, by Theorem 4.2.15, there are elements

$$(\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n)$$

of V such that

$$(\mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{v}_{r+1}, \ldots, \mathbf{v}_n)$$

is a basis for V. Let

$$(\mathbf{v}_1^*, \ldots, \mathbf{v}_r^*, \mathbf{v}_{r+1}^*, \ldots, \mathbf{v}_n^*)$$

its relative dual basis, we prove that

$$(\mathbf{v}_{r+1}^*, \ldots, \mathbf{v}_n^*)$$

is a basis for $U^\perp$. Clearly $\mathbf{v}_{r+1}^*, \ldots, \mathbf{v}_n^*$ are linearly independent since they are distinct elements of a basis. Assume $\mathbf{u}$ is a vector in U, then $\mathbf{u}$ is a linear combination of the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$:

$$\mathbf{u} = a_1\mathbf{v}_1 + \cdots + a_r\mathbf{v}_r,$$

since

{equno}
$$\mathbf{v}_i^*(\mathbf{v}_j) = 0 \text{ if } i \neq j \tag{4.30}$$

for every $i \in \{r+1, \ldots, n\}$, we have

$$\begin{aligned} \mathbf{v}_i^*(\mathbf{u}) &= \mathbf{v}_i^*(a_1\mathbf{v}_1 + \cdots + a_r\mathbf{v}_r) \\ &= \mathbf{v}_i^*(a_1\mathbf{v}_1) + \cdots + \mathbf{v}_i^*(a_r\mathbf{v}_r) \\ &= a_1(\mathbf{v}_i^*(\mathbf{v}_1)) + \cdots + a_r(\mathbf{v}_i^*(\mathbf{v}_r)) \\ &= a_1 0 + \cdots + a_r 0 = 0, \end{aligned}$$

thus $\mathbf{v}_{r+1}^*, \ldots, \mathbf{v}_n^*$ are also contained in $U^\perp$. Finally let $\mathbf{v}^* \in U^\perp$ and write $\mathbf{v}^*$ as a linear combination of the elements $\mathbf{v}_1^*, \ldots, \mathbf{v}_r^*, \mathbf{v}_{r+1}^*, \ldots, \mathbf{v}_n^*$ of the dual basis of V^*:

$$\mathbf{v}^* = b_1\mathbf{v}_1^* + \cdots + b_r\mathbf{v}_r^* + b_{r+1}\mathbf{v}_{r+1}^* + \cdots + b_n\mathbf{v}_n^*$$

then evaluating at $\mathbf{v}_j$ for every $j \in \{1, \ldots, r\}$, while $\mathbf{v}_j \in U$, we get, by Equation (4.30),

$$\begin{aligned} 0 = \mathbf{v}^*(\mathbf{v}_j) &= (b_1\mathbf{v}_1^* + \cdots + b_r\mathbf{v}_r^* + b_{r+1}\mathbf{v}_{r+1}^* + \cdots + b_n\mathbf{v}_n^*)(\mathbf{v}_j) \\ &= b_1(\mathbf{v}_1^*(\mathbf{v}_j)) + \cdots + b_r(\mathbf{v}_r^*(\mathbf{v}_j)) + b_{r+1}(\mathbf{v}_{r+1}^*(\mathbf{v}_j)) + \cdots + b_n(\mathbf{v}_n^*(\mathbf{v}_j)) \\ &= b_j \end{aligned}$$

whence $b_j = 0$ for every $j \in \{1, \ldots r\}$ so $\mathbf{v}^*$ is actually a linear combination of $\mathbf{v}_{r+1}^*, \ldots, \mathbf{v}_n^*$. So $\mathbf{v}_{r+1}^*, \ldots, \mathbf{v}_n^*$ are linearly independent elements of $U^\perp$ that span $U^\perp$, hence $(\mathbf{v}_{r+1}^*, \ldots, \mathbf{v}_n^*)$ is a basis for $U^\perp$. It follows that

$$\dim U^\perp = n - r = \dim(V) - \dim(U).$$

■

4.3.5 The transpose map

Let

$$f\colon V \to W$$

be a linear map between two vector spaces V and W and let

$$V^* \text{ and } W^*$$

be respectively their dual spaces. Note that, for every $\mathbf{v} \in V$ $f(\mathbf{v}) \in W$, so, for every linear form $\mathbf{w}^* \in W^*$, composing $\mathbf{w}^*$ with f, we get a map

$$\begin{array}{rccc} w^* \circ f\colon & V & \to & \mathbb{R} \\ & \mathbf{v} & \mapsto & \mathbf{w}^*(\mathbf{f}(\mathbf{v}) \end{array}$$

which is linear map since it is the composition of the linear map $\mathbf{w}^*$ with the linear map f, hence an element of V^*.

Thus composition with f defines a map

$$\begin{array}{rccc} f^*\colon & W^* & \to & V^* \\ & \mathbf{w}^* & \mapsto & \mathbf{w}^* \circ f \end{array}$$

The map f^* is called the *transpose* map of the map f. The next theorem gives the main properties of the transpose map.

{ortogtrans}

Theorem 4.3.15 *Let $f\colon V \to W$ be a linear map between two vector spaces V and W. Let V^* and W^* be their dual spaces respectively and let $f^*\colon W^* \to V^*$ be the transpose map of f. Then*

1. *f^* is linear;*
2. *$(f(V))^\perp = \ker f^*$*

PROOF. To prove that f^* is linear, we have to prove that, for every $\mathbf{w}^*$, $\mathbf{u}^*$ in W^* and every $k \in \mathbb{R}$

(i) $f^*(\mathbf{w}^* + \mathbf{u}^*) = f^*(\mathbf{w}^*) + f^*(\mathbf{u}^*)$ and

(ii) $f^*(k\mathbf{w}^*) = kf^*(\mathbf{w}^*)$.

To prove (i) observe that the left and the right sides of the equation (i) are functions from V to $\mathbb{R}$, so, in order to prove that they are equal, we need to prove that they coincide elementwise, that is, for every $\mathbf{v} \in V$,

$$(f^*(\mathbf{w}^* + \mathbf{u}^*))(v) = (f^*(\mathbf{w}^*) + f^*(\mathbf{u}^*))(v).$$

Now, by the definition of f^*,

$$\begin{array}{rcl} (f^*(\mathbf{w}^* + \mathbf{u}^*))(v) & = & (\mathbf{w}^* + \mathbf{u}^*) \circ f(\mathbf{v}) \\ & = & (\mathbf{w}^* + \mathbf{u}^*)(f(\mathbf{v})) \\ & = & \mathbf{w}^*(f(\mathbf{v})) + \mathbf{u}^*(f(\mathbf{v})) \\ & = & (\mathbf{w}^* \circ f)(\mathbf{v}) + (\mathbf{u}^* \circ f)(\mathbf{v}) \\ & = & (f^*(\mathbf{w}^*))(\mathbf{v}) + (f^*(\mathbf{u}^*))(v) \\ & = & (f^*(\mathbf{w}^*) + f^*(\mathbf{u}^*))(v), \end{array}$$

proving (i). Similarly, to prove (ii),n we have to prove that, for every $\mathbf{v} \in V$

$$(f^*(k\mathbf{w}^*))(\mathbf{v}) = (kf^*(\mathbf{w}^*))(\mathbf{v}).$$

As above, by the definition of f^*,

$$\begin{array}{rcl} (f^*(k\mathbf{w}^*))(\mathbf{v}) & = & ((k\mathbf{w}^*) \circ f)(\mathbf{v}) \\ & = & (k\mathbf{w}^*)(f(\mathbf{v})) \\ & = & k(\mathbf{w}^*(f(\mathbf{v})) \\ & = & k(\mathbf{w}^*(f(\mathbf{v})) \\ & = & (kf^*(\mathbf{w}^*))(\mathbf{v}). \end{array}$$

proving (ii), hence f^* is linear. Next assume $\mathbf{w}^* \in W^*$. Then $\mathbf{w}^* \in \ker(f^*)$ if and only if $f^*(\mathbf{w}^*) = 0_{V^*}$, that is, if and only if, for every $\mathbf{v} \in V$,

$$0 = 0_{V^*}(\mathbf{v}) = (f^*(\mathbf{w}^*))(\mathbf{v}) = \mathbf{w}^*(f(\mathbf{v})),$$

which is equivalent to say that $w^* \in (f(V))^\perp$. ∎

{ranktras}

Corollary 4.3.16 *Let $f\colon V \to W$ be a linear map and $f^*\colon W \to V$ its transpose map. Then f and f^* have the same rank.*

Proof. Recall that the rank $rk(f)$ of a linear map f is the dimension of $f(V)$. By Theorem 4.3.14

$$\dim(f(V)) = \dim(W) - \dim((f(v))^\perp).$$

On the other hand, by Theorem 4.3.15 $(f(V))^\perp = \ker(f^*)$, and since $\dim(W) = \dim(W^*)$, we get

$$\dim(W) - \dim((f(v))^\perp) = \dim(W^*) - \dim(\ker(f^*))$$

which, on turn is equal to $\dim(f^*)$ by Theorem 4.24. ∎

4.4 Matrices

In this section and in the next section we'll introduce some tools for computing with vector spaces and linear maps.

4.4.1 Matrices, columns, rows, entries and transpose

Theorem 4.3.4 has a consequence of fundamental importance that allows us to describe all linear maps between two vector spaces. In order to do that, we need to introduce a new tool which will be fundamental for the computations: an $m \times n$ *matrix* with entries in $\mathbb{R}$ is just an n-tuple of vectors of $\mathbb{R}^m$. These vectors are called *columns* of the matrix. So, for example, the columns of the 4×3 matrix

$$\begin{pmatrix} 1 & 1 & 0 \\ 2 & 1 & 0 \\ 3 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} \tag{4.31}$$

{massoes}

are the vectors

$$\begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \text{ and } \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Given a matrix

$$M := \begin{pmatrix} a_{1,1}, & a_{1,2}, & \dots & a_{1,n} \\ a_{2,1}, & a_{2,2}, & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1}, & a_{m,2}, & \dots & a_{m,n} \end{pmatrix} \tag{4.32}$$ {masso0}

the n-tuples

$$\begin{gathered} (a_{1,1}, a_{1,2}, \dots a_{1,n}) \\ (a_{2,1}, a_{2,2}, \dots a_{2,n}) \\ \vdots \\ (a_{m,1}, a_{n,2}, \dots a_{m,n}) \end{gathered}$$

are called the *rows* of the matrix M. If $i \in \{1, \dots, n\}$ and $j \in \{1 \dots, m\}$ the (i, j)-*entry* of the matrix M is the scalar $a_{i,j}$ lying in the ith row and jth column.

So, for example, the rows of the matrix in (4.31) are

$$(1, 1, 0), (2, 1, 0), (3, 1, 0), \text{ and } (1, 1, 0),$$

its $(1, 3)$-entry is 3 and its $(2, 2)$-entry is 1.

Finally, the *transpose* of the matrix M is the $m \times n$ matrix M^t whose ith column is the ith row of the matrix (4.32), or, equivalently, the matrix M^t whose (i, j)-entry is the (j, i)- entry of the matrix M for every $i \in \{1, \dots, n\}$ and $j \in \{1 \dots, m\}$, that is

$$M^t := \begin{pmatrix} a_{1,1}, & a_{2,1}, & \dots & a_{m,1} \\ a_{1,2}, & a_{2,2}, & \dots & a_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n}, & a_{2,n}, & \dots & a_{m,n} \end{pmatrix}$$

So, for example, the transpose of the 4×3 matrix

$$\begin{pmatrix} 1 & 1 & 0 \\ 2 & 1 & 0 \\ 3 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

in (4.31) is the 3×4 matrix

$$\begin{pmatrix} 1 & 2 & 3 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

4.4.2 The space of matrices

The set of matrices with n rows and m columns with entries in $\mathbb{R}$ will be denoted by $\mathcal{M}_{n \times m}(\mathbb{R})$. This set can be given an obvious structure of vector space defining entrywise the sum and the product by scalars. That is, given two matrices

A and B with m rows and n columns with entries in $\mathbb{R}$,

$$A := \begin{pmatrix} a_{1,1}, & a_{2,1}, & \dots & a_{m,1} \\ a_{1,2}, & a_{2,2}, & \dots & a_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n}, & a_{2,n}, & \dots & a_{m,n} \end{pmatrix} \text{ and } B := \begin{pmatrix} b_{1,1}, & b_{2,1}, & \dots & b_{m,1} \\ b_{1,2}, & b_{2,2}, & \dots & b_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1,n}, & b_{2,n}, & \dots & b_{m,n} \end{pmatrix},$$

define $A + B$ by

$$A + B := \begin{pmatrix} a_{1,1} + b_{1,1}, & a_{2,1} + b_{2,1}, & \dots & a_{m,1} + b_{m,1} \\ a_{1,2} + b_{1,2}, & a_{2,2} + b_{2,2}, & \dots & a_{m,2} + b_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n} + b_{1,n}, & a_{2,n} + b_{2,n}, & \dots & a_{m,n} + b_{m,n} \end{pmatrix},$$

similarly define the product of the scalar k by the matrix A by:

$$kA := \begin{pmatrix} ka_{1,1}, & ka_{2,1}, & \dots & ka_{m,1} \\ ka_{1,2}, & ka_{2,2}, & \dots & ka_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ ka_{1,n}, & ka_{2,n}, & \dots & ka_{m,n} \end{pmatrix}$$

For every $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$ let $E_{i,j}$ be the matrix whose (i, j)-entry is 1 and whose (h, k)-entry is 0 for every $(h, k) \neq (i, j)$.

We leave the reader the easy check that

1. the set $M_{n\times m}(\mathbb{R})$, with entrywise addition and multiplication by scalars, satisfies the axioms of a vector space;

2. The matrices $E_{i,j}$, for $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$, are linearly independent and generate $M_{m\times n}(\mathbb{R})$.

4.4.3 Matrices associated to linear maps

Now let V and W be vector spaces, fix

$$\text{a basis } \mathcal{V} := (\mathbf{v}_1, \dots, \mathbf{v}_n) \text{ for } V$$

and

$$\text{a basis } \mathcal{W} := (\mathbf{w}_1, \dots, \mathbf{w}_m) \text{ for } W.$$

Suppose

$$f \colon V \to W$$

is a linear map, then, by Theorem 4.3.4 (or by Lemma 4.19), f is fully determined by the images $f(\mathbf{v}_i)$ of the basis vectors $\mathbf{v}_i$ where $i \in \{1, \dots, n\}$. On turn, the images $f(\mathbf{v}_i)$ are vectors of W, so they can be uniquely written as linear

combinations of the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$, that is, there exists a unique $m \times n$ *matrix*

$$M := \begin{pmatrix} a_{1,1}, & a_{2,1}, & \ldots & a_{m,1} \\ a_{1,2}, & a_{2,2}, & \ldots & a_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n}, & a_{2,n}, & \ldots & a_{m,n} \end{pmatrix}$$

such that $a_{j,i}$ is a scalar for every $i \in \{1, \ldots, n\}$ and every $j \in \{1, \ldots, m\}$ and

{masso}

$$\begin{array}{rcl} f(\mathbf{v}_1) & = & a_{1,1}\mathbf{w}_1 + a_{2,1}\mathbf{w}_2 \cdots + a_{m,1}\mathbf{w}_m \\ f(\mathbf{v}_2) & = & a_{1,2}\mathbf{w}_1 + a_{2,2}\mathbf{w}_2 \cdots + a_{m,2}\mathbf{w}_m \\ \vdots \quad & \vdots & \vdots \\ f(\mathbf{v}_n) & = & a_{1,n}\mathbf{w}_1 + a_{2,n}\mathbf{w}_2 \cdots + a_{m,n}\mathbf{w}_m \end{array} \tag{4.33}$$

In other words,

{coefficie}

$$a_{j,i} = \mathbf{w}_j^*(f(\mathbf{v}_i)) \tag{4.34}$$

for every $i \in \{1, \ldots, n\}$ and every $j \in \{1, \ldots, m\}$, where $(\mathbf{w}_1^*, \ldots, \mathbf{w}_m^*)$ is the dual basis of $(\mathbf{w}_1, \ldots, \mathbf{w}_m)$.

Therefore, once the (ordered!) bases $\mathcal{V}$ and $\mathcal{W}$ have been fixed, all we need to know to determine the map f is the matrix M, or equivalently, its transpose matrix

$$M^t \begin{pmatrix} a_{1,1}, & a_{1,2}, & \ldots & a_{1,n} \\ a_{2,1}, & a_{2,2}, & \ldots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1}, & a_{m,2}, & \ldots & a_{m,n} \end{pmatrix}$$

The matrix M^t is called the *matrix associated to* the linear map f *with respect to* the bases $\mathcal{V}$ of V and $\mathcal{W}$ of W.[8] Since M^t depends on f and on the chosen bases $\mathcal{V}$ and $\mathcal{W}$ it will be denoted by

$$M_{\mathcal{W}}^{\mathcal{V}}(f)$$

So, given a linear map $f\colon V \to W$ from a vector space V to a vector space W, and given a basis the procedure to find a matrix associated to f is

1. find an ordered basis $\mathcal{V} := (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ for V;

2. find an ordered basis $\mathcal{W} := (\mathbf{w}_1, \ldots, \mathbf{w}_m)$ for W;

[8] One could reasonably ask why we do not define M to be the matrix associated to f. We will see that the reason is that it depends on the habit of writing $f(x)$ to denote the image of an element via a function f. There is, however an alternative notation for $f(x)$, namely $(x)f$ or x^f, in that case one defines M instead of M^t as the associated matrix

3. write the images via f of the elements of $\mathcal{V}$ as linear combinations of the elements of $\mathcal{W}$:

$$\begin{array}{ccc} f(\mathbf{v}_1) & = & a_{1,1}\mathbf{w}_1 + a_{2,1}\mathbf{w}_2 \cdots + a_{m,1}\mathbf{w}_m \\ f(\mathbf{v}_2) & = & a_{1,2}\mathbf{w}_1 + a_{2,2}\mathbf{w}_2 \cdots + a_{m,2}\mathbf{w}_m \\ \vdots & \vdots & \vdots \\ f(\mathbf{v}_n) & = & a_{1,n}\mathbf{w}_1 + a_{2,n}\mathbf{w}_2 \cdots + a_{m,n}\mathbf{w}_m \end{array}$$

4. erase the $f(\mathbf{v}_i)$'s, the $\mathbf{w}_j$'s, the symbols $=$ and $+$:

$$\begin{pmatrix} a_{1,1}, & a_{2,1}, & \dots & a_{m,1} \\ a_{1,2}, & a_{2,2}, & \dots & a_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n}, & a_{2,n}, & \dots & a_{m,n} \end{pmatrix}$$

5. transpose the above matrix

$$\begin{pmatrix} a_{1,1}, & a_{1,2}, & \dots & a_{1,n} \\ a_{2,1}, & a_{2,2}, & \dots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1}, & a_{m,2}, & \dots & a_{m,n} \end{pmatrix}$$

{matlinmap}

Theorem 4.4.1 *Let V and W be vector spaces, $\mathcal{V} := (\mathbf{v}_1, \dots, \mathbf{v}_n)$ be an ordered basis of V and $\mathcal{W} := (\mathbf{w}_1, \dots, \mathbf{w}_m)$ be an ordered basis of W. Then the map*

$$\begin{array}{rccc} M_{\mathcal{V}}^{\mathcal{W}}: & End(V,W) & \to & M_{m\times n}(\mathbb{R}) \\ & f & \mapsto & M_{\mathcal{W}}^{\mathcal{V}}(f) \end{array}$$

is an isomorphism of vector spaces.

PROOF. Clearly $M_{\mathcal{V}}^{\mathcal{W}}$ is a map from $End(V,W)$ to $M_{m\times n}(\mathbb{R})$ We first prove that $M_{\mathcal{V}}^{\mathcal{W}}$ is linear. Let f and g be in $Hom(V,W)$. For every $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,m\}$, the (i,j)-entry of the matrix $M_{\mathcal{W}}^{\mathcal{V}}(f+g)$ is the sum of the (i,j)-entry of the matrix $M_{\mathcal{W}}^{\mathcal{V}}(f)$ plus the (i,j)-entry of the matrix $M_{\mathcal{W}}^{\mathcal{V}}(g)$. By the Formula (4.34) and linearity of $\mathbf{w}_j^*$, we get that the (i,j)-entry of the matrix $M_{\mathcal{W}}^{\mathcal{V}}(f+g)$ is

$$\mathbf{w}_j^*((f+g)(\mathbf{v}_i)) = \mathbf{w}_j^*((f(\mathbf{v}_i) + g(\mathbf{v}_i)) = \mathbf{w}_j^*(f(\mathbf{v}_i)) + \mathbf{w}_j^*(g(\mathbf{v}_i))$$

which is precisely the sum of the (i,j)-entry of the matrix $M_{\mathcal{W}}^{\mathcal{V}}(f)$ plus the (i,j)-entry of the matrix $M_{\mathcal{W}}^{\mathcal{V}}(g)$.

Similarly, if $k \in \mathbb{R}$, the (i,j)-entry of the matrix $M_{\mathcal{W}}^{\mathcal{V}}(kf)$ is

$$\mathbf{w}_j^*((kf)(\mathbf{v}_i)) = \mathbf{w}_j^*(k(f(\mathbf{v}_i))) = k(\mathbf{w}_j^*(f(\mathbf{v}_i)))$$

which is precisely the (i,j)-entry of the matrix $kM_{\mathcal{W}}^{\mathcal{V}}(f)$. Thus $M_{\mathcal{V}}^{\mathcal{W}}$ is linear. Note also that

$$M_{\mathcal{W}}^{\mathcal{V}}(f_{i,j}) = E_{j,i}$$

so $M_{\mathcal{W}}^{\mathcal{V}}$ sends a basis of $Hom(V,W)$ to a basis of $M_{n\times m}(\mathbb{R})$ hence, by Theorem 4.3.5, $M_{\mathcal{W}}^{\mathcal{V}}$ is bijective. ∎

Now consider the particular case in which $V=\mathbb{R}^n$, $W=\mathbb{R}^m$ and $\mathcal{V}$ and $\mathcal{W}$ are the canonical bases, i.e.

$$\mathcal{V} := (e_1, e_2, \ldots, e_n)$$

where $e_1, e_2, \ldots, e_n$ are the n-tuples

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \ldots, e_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

and

$$\mathcal{W} := (e'_1, e'_2, \ldots, e'_n)$$

where $e'_1, e'_2, \ldots, e'_n$ are the m-tuples

$$e'_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, e'_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \ldots, e'_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

Let $f\colon \mathbb{R}^n \to \mathbb{R}^m$ be a linear map and $M_{\mathcal{W}}^{\mathcal{V}}(f)$ be the matrix associated to f with respect to the bases $\mathcal{V}$ and $\mathcal{W}$. Say

$$M_{\mathcal{W}}^{\mathcal{V}}(f) = \begin{pmatrix} a_{1,1}, & a_{1,2}, & \ldots & a_{1,n} \\ a_{2,1}, & a_{2,2}, & \ldots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1}, & a_{m,2}, & \ldots & a_{m,n} \end{pmatrix}$$

Then

$$\begin{aligned} f(e_1) &= a_{1,1}e'_1 + a_{2,1}e'_2 + \cdots + a_{m,1}e'_m \\ &= a_{1,1}\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + a_{2,1}\begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \cdots + a_{m,1}\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} a_{1,1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ a_{2,1} \\ \vdots \\ 0 \end{pmatrix} + \cdots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ a_{m,1} \end{pmatrix} = \begin{pmatrix} a_{1,1} \\ a_{2,1} \\ \vdots \\ a_{m,1} \end{pmatrix}, \end{aligned}$$

which is precisely the first column of the matrix $M_{\mathcal{W}}^{\mathcal{V}}(f)$. The same computation, shows that $f(e_i)$ is precisely the ith row of the matrix $M_{\mathcal{W}}^{\mathcal{V}}(f)$, for every $i \in \{1,\ldots,n\}$. We have shown:

{linrn}

Lemma 4.4.2 *Let $f\colon \mathbb{R}^n \to \mathbb{R}^m$ be a linear map. Then the columns of the matrix $M(f)$ associated to f with respect to the canonical bases of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, are precisely the images, via f of the vectors of the canonical basis of $\mathbb{R}^n$, i.e. $f(e_i)$ is the i-th column of $M(f)$. In particular $rk(f) = rk(M(f))$.*

Now take a generic element

$$v := \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix} = t_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + t_2 \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \cdots + t_n \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

in $\mathbb{R}^n$ Then

$$\begin{aligned} f(v) &= f\left(t_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + t_2 \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \cdots + t_n \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \right) \\ &= t_1 f \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + t_2 f \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \cdots + t_n f \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \\ &= t_1 \begin{pmatrix} a_{1,1} \\ a_{2,1} \\ \vdots \\ a_{m,1} \end{pmatrix} + t_2 \begin{pmatrix} a_{1,2} \\ a_{2,2} \\ \vdots \\ a_{m,2} \end{pmatrix} + \cdots + t_n \begin{pmatrix} a_{1,n} \\ a_{2,n} \\ \vdots \\ a_{m,n} \end{pmatrix} \\ &= \begin{pmatrix} t_1 a_{1,1} \\ t_1 a_{2,1} \\ \vdots \\ t_1 a_{m,1} \end{pmatrix} + \begin{pmatrix} t_2 a_{1,2} \\ t_2 a_{2,2} \\ \vdots \\ t_2 a_{m,2} \end{pmatrix} + \cdots + \begin{pmatrix} t_n a_{1,n} \\ t_n a_{2,n} \\ \vdots \\ t_n a_{m,n} \end{pmatrix} \\ &= \begin{pmatrix} t_1 a_{1,1} + t_2 a_{1,2} + \cdots + t_n a_{1,n} \\ t_1 a_{2,1} + t_2 a_{2,2}, + \cdots + t_n a_{2,n} \\ \vdots \\ t_1 a_{m,1} + t_2 a_{m,2} + \cdots + t_n a_{m,n} \end{pmatrix} \end{aligned}$$

The latter vector is in $\mathbb{R}^m$ and, for every $j \in \{1, \ldots, m\}$ its jth coordinate is obtained by multiplying pointwise the n entries of jth row of the matrix $M_{\mathcal{W}}^{\mathcal{V}}(f)$ by the n coordinates of the vector V and adding all products. This is called the *rows by columns* product of the $m \times n$ matrix $M_{\mathcal{W}}^{\mathcal{V}}(f)$ by the vector v and is denoted by

$$M_{\mathcal{W}}^{\mathcal{V}}(f) \cdot v.$$

Thus, in our example,

$$\begin{pmatrix} a_{1,1}, & a_{1,2}, & \dots & a_{1,n} \\ a_{2,1}, & a_{2,2}, & \dots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1}, & a_{m,2}, & \dots & a_{m,n} \end{pmatrix} \cdot \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix} = \begin{pmatrix} t_1 a_{1,1} + t_2 a_{1,2} + \cdots + t_n a_{1,n} \\ t_1 a_{2,1} + t_2 a_{2,2}, + \cdots + t_n a_{2,n} \\ \vdots \\ t_1 a_{m,1} + t_2 a_{m,2} + \cdots + t_n a_{m,n} \end{pmatrix}$$

A consequence of the above discussion is

{matasso}

Theorem 4.4.3 *If $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is a linear map and $M(f)$ is the matrix associated to f with respect to the canonical bases of $\mathbb{R}^n$ and $\mathbb{R}^m$, then f maps every vector*

$$T := \begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix}$$

in R^n to $A \cdot T$. Conversely if M is a matrix with m rows and n columns, then the map

$$\begin{array}{rccc} f_M\colon & \mathbb{R}^n & \to \mathbb{R}^m & \\ & T & \mapsto & M \cdot T \end{array}$$

is a linear map and M is the matrix associated to f_M with respect to the canonical bases of $\mathbb{R}^n$ and $\mathbb{R}_m$.

Given a matrix M with m rows and n columns, the map $f_M\colon \mathbb{R}^n \to \mathbb{R}^m$, defined as in the above theorem, is called the linear map *associated* to the matrix M. Note that an immediate consequence of the above theorem is that

{allmat}

Corollary 4.4.4 *For every matrix M there is a linear map f such that M is the matrix associated to f with respect to suitable bases. Moreover the maximum number of linearly independent columns of the matrix M (i.e. the dimension of the subspace generated by these columns) is precisely the rank of the linear map f.*

4.4.4 Rank of a matrix

Let M be a matrix with m rows and n columns as in 4.32, the last result of the previous section suggests to define the *rank* of M as the dimension of the subspace of $\mathbb{R}^m$. This rank will be denoted by $rk(M)$ and, if f is a linear map such that M is associated to f, then $rk(f) = rk(M)$.

Now let, as above,

$$f\colon V \to W$$

be a linear map between two vector spaces V and W, choose a basis

$$\mathcal{V} := (\mathbf{v}_1, \dots, \mathbf{v}_n)$$

of V and a basis

$$\mathcal{W} := (\mathbf{w}_1, \dots, \mathbf{w}_m)$$

of W and let

$$M_{\mathcal{W}}^{\mathcal{V}}(f)$$

be the matrix associated to f with respect to these bases. Let

$$f^*\colon W^* \to V^*$$

be the transpose map. Let

$$\mathcal{W}^* := (\mathbf{w}_1^*, \ldots, \mathbf{w}_m^*)$$

be the dual basis of W relative to the basis $\mathcal{W}$, let

$$\mathcal{V}^* := (\mathbf{v}_1^*, \ldots, \mathbf{v}_n^*)$$

the dual basis of V relative to the basis $\mathcal{W}$ and let

$$M_{\mathcal{V}^*}^{\mathcal{V}^*}(f^*)$$

the matrix associated to f^* with respect to the bases $\mathcal{W}^*$ and $\mathcal{V}^*$.

{mattransp}

Lemma 4.4.5 *With the above notation, $M_{\mathcal{V}^*}^{\mathcal{V}^*}(f^*)$ is the transpose matrix of the matrix $M_{\mathcal{W}}^{\mathcal{V}}(f)$.*

PROOF. Let $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$, let $a_{i,j}$ be the entry in the ith column and jth row of the matrix $M_{\mathcal{W}}^{\mathcal{V}}(f)$. We show that $a_{i,j}$ is equal to the entry $b_{j,i}$ in the jth column and ith row of the matrix $M_{\mathcal{V}^*}^{\mathcal{V}^*}(f^*)$. Indeed by the Formula (4.34), and the Formula (4.28),

$$a_{i,j} = \mathbf{w}_j^*(f(\mathbf{v_i})) = (f^*(\mathbf{w}_j^*))(\mathbf{v}_i) = e_{\mathbf{v}_i}(f^*(\mathbf{w}_j^*)) = \mathbf{v}_i^{**}(f^*(\mathbf{w}_j^*)) = b_{j,i}.$$

■

The above lemma has an important corollary:

{useful}

Corollary 4.4.6 *if M is a matrix, then the maximum number of linearly independent columns of M is equal to the maximum number of linearly independent columns of M^t or, equivalently, to the maximum number of linearly independent rows of M*

PROOF. By Corollary 4.4.4 ther is a linear map f such that M is a matrix associated to f. Then, by Lemma 4.4.5 M^t is the matrix associated to f^* so, by Corollary 4.3.16,

$$rk(M) = rk(f) = rk(f^*) = rk(M^t).$$

■

4.5 Linear systems

4.5.1 Introduction

We have already encountered the problem of writing a vector $\mathbf{v}$ of a vector space V as a linear combination of some vectors $\mathbf{v_1}, \dots \mathbf{v_n}$ spanning V. That is, given $\mathbf{v}$ and knowing $\mathbf{v_1}, \dots \mathbf{v_n}$, find scalars $a-1, \dots, a_n$ such that

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

For example we want to write the vector

$$\begin{pmatrix} 3 \\ 2 \\ 2 \end{pmatrix}$$

as a linear combination of the vectors

$$\begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \text{ and } \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

As mentioned above we need to find scalars a_1, a_2, and a_3 such that

{syst}
$$a_1 \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + a_3 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 2 \end{pmatrix} \tag{4.35}$$

Now the left side of Equation (4.35) is equal to

$$\begin{pmatrix} a_1 + a_3 \\ a_1 + a_2 + a_3 \\ 3a_1 + a_2 \end{pmatrix}$$

so a_1, a_2, and a_3 have to verify simultaneously the three equalities

$$a_1 + a_3 = 3$$

$$a_1 + a_2 + a_3 = 2$$

$$3a_1 + a_2 = 2$$

Or, equivalently, the triple

$$\begin{pmatrix} a_1 \\ a_2 + a_3 \\ a_3 \end{pmatrix}$$

must be a solution of the following *linear system* of three equations in the *indeterminates* x_1, x_2, and x_3

{syst1}
$$\begin{cases} x_1 + & & x_3 = 3 \\ x_1 + x_2 + & x_3 = 2 \\ 3x_1 + x_2 & = 2 \end{cases} \tag{4.36}$$

The most obvious way to solve this system is to get one of the indeterminates as a function of the others in one equation and substitute in the others. For example from the first equation we obtain

$$x_3 = 3 - x_1$$

so, if we replace x_3 with $3 - x_1$ in the second and in the third equations, we get

$$\left\{\begin{array}{rcrcrcl} & & & & x_3 & = & 3 - x_1 \\ x_1 & + & x_2 & + & 3 - x_1 & = & 2 \\ 3x_1 & + & x_2 & & & = & 2 \end{array}\right.$$

that is

$$\left\{\begin{array}{rcrcl} & & & x_3 & = & 3 - x_1 \\ & & x_2 & & = & -1 \\ 3x_1 & + & x_2 & & = & 2 \end{array}\right.$$

obtaining $x_2 = -1$ Finally, replacing x_2 with -1 in the third equation we get

$$\left\{\begin{array}{rcrcl} & & & x_3 & = & 3 - x_1 \\ & & x_2 & & = & -1 \\ 3x_1 & - & 1 & & = & 2 \end{array}\right.$$

that is

$$\left\{\begin{array}{rcrcl} & & & x_3 & = & 3 - x_1 \\ & & x_2 & & = & -1 \\ 3x_1 & & & & = & 3 \end{array}\right.$$

which gives $x_1 = 1$ and, from the first equation, $x_3 = 2$. Thus we have found thet the triple

$$\begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$$

is a solution of the System (4.36). Indeed, if we replace x_1 with 1, x_2 with -1 and x_3 with 2 in (4.36), we get three identities. Therefore the way to write the vector

$$\begin{pmatrix} 3 \\ 2 \\ 2 \end{pmatrix}$$

as a linear combination of the vectors

$$\begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \text{ and } \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

is

$$\begin{pmatrix} 3 \\ 2 \\ 2 \end{pmatrix} = 1 \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix} + (-1) \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + 2 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

In general we shall see that not every linear system has solutions and, also, if there do exixt solution, they do not need to be unique. for example the system

$$\begin{cases} x_1 + x_2 = 0 \\ 2x_1 + 2x_2 = 4 \end{cases}$$

has no solutions (why?). While, for every $a \in \mathbb{R}$, the pair

$$\begin{pmatrix} a \\ -a \end{pmatrix}$$

is a solution of the system

$$\begin{cases} x_1 + x_2 = 0 \\ 2x_1 + 2x_2 = 0 \end{cases}$$

so the latter system has infinitely many solutions. In the sequel we'll find necessary and sufficient conditions for a linear system to have solutions and describe the set of these solutions.

4.5.2 Formal definition of a linear system

let m and n be positive integers. A *linear system of m equations and n indeterminates* is a pair

$$(A, B)$$

where A is a matrix with coefficients in $\mathbb{R}$ with m rows and n columns and B is a vector in $\mathbb{R}^m$, i.e. a matrix with m rows and 1 column. The matrix A is called the *matrix of the coefficients* and the vector B is called the *column of the constant terms* of the system (A, B).

Given a linear system (A, B) of m equations and n indeterminates a *solution* of (A, B) is a vector T in $\mathbb{R}^n$ that verifies the following equality:

{syst3}
$$A \cdot T = B, \tag{4.37}$$

where $A \cdot T$ is the rows by column product of the matrix A by the vector T (note that this makes sense, since the numbers of columns of the matrix A is n which is precisely the number of rows of the vector T).

What has this to do with our usual idea of a linear system of m equations and n indeterminates $x_1, x_2, \ldots, x_n$? Well assume the system is written in the usual form

{syst4}
$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n = b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n = b_2 \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n = b_1 \end{cases} \tag{4.38}$$

with $a_{i,j}$ and b_i in $\mathbb{R}$ for every $i \in \{1,\dots,m\}$ and $j \in \{1,\dots,n\}$. Let A be the $m \times n$ matrix

$$\begin{pmatrix} a_{1,1}, & a_{1,2}, & \dots & a_{1,n} \\ a_{2,1}, & a_{2,2}, & \dots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1}, & a_{m,2}, & \dots & a_{m,n} \end{pmatrix}$$

and B be the vector

$$\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix},$$

Then a vector

$$T := \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix},$$

of R^n is a solution of the System (4.38) if and only if the following identities are simultaneaously verified:

$$\begin{array}{ccccccccccc} a_{1,1}t_1 & + & a_{1,2}t_2 & + & \cdots & + & a_{1,n}t_n & = & b_1 \\ a_{2,1}t_1 & + & a_{2,2}t_2 & + & \cdots & + & a_{2,n}t_n & = & b_2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ a_{1,1}t_1 & + & a_{1,2}t_2 & + & \cdots & + & a_{1,n}t_n & = & b_1 \end{array}$$

or, equivalently, if the following vectorial identity is verified:

$$\begin{pmatrix} a_{1,1}t_1 & + & a_{1,2}t_2 & + & \cdots & + & a_{1,n}t_n \\ a_{2,1}t_1 & + & a_{2,2}t_2 & + & \cdots & + & a_{2,n}t_n \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{1,1}t_1 & + & a_{1,2}t_2 & + & \cdots & + & a_{1,n}t_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}. \qquad (4.39) \quad \texttt{\{eqsyst\}}$$

But now note that the left term of the above identity is precisely the rows by columns product of the matrix A by the vector T, so T is a solution of the system (4.38) if and only if T verifies the identity (4.37), that is T is a solution of the system (A, B). So our definition matches what we usual think a linear system is.

4.5.3 The set of solutions of a linear system

Let (A, B) be a linear system with m equations and n indeterminates. Take a close look at the identity (4.37). Let f_A be the linear map associated to the matrix A, that is the linear map

$$\begin{array}{rccc} f_A\colon & \mathbb{R}^n & \to & \mathbb{R}^m \\ & X & \mapsto & A \cdot X \end{array}$$

Then the identity (4.37) can be rewritten as

$$f_A(T) = B, \tag{4.40}$$

that is a vector T in $\mathbb{R}^n$ is a solution of the system (A, B) if and only if

$$T \in f^{-1}(B).$$

This implies that the system (A, B) has solutions if and only if the preimage $f^{-1}(B)$ of the vector B via f is not empty, or, equivalently, if and only if

{syst7}
$$B \in f_A(\mathbb{R}^n). \tag{4.41}$$

Now, by Lemma 4.4.2 the columns of the matrix A are precisely the images via f_A of the vectors $e_1, e_2, \ldots, e_n$ of the canonical basis of R^n, so, by Lemma 4.2.14 these columns span $f_A(V)$. It follows that $B \in f_A(V)$ if and only if B is a linear combination of the columns of the matrix A. This fact can be seen also in another way: assume Let

$$A_1 := \begin{pmatrix} a_{1,1} \\ a_{2,1} \\ \cdots \\ a_{m,1} \end{pmatrix}, \quad A_2 := \begin{pmatrix} a_{1,2} \\ a_{2,2} \\ \cdots \\ a_{m,2} \end{pmatrix}, \ldots, A_n := \begin{pmatrix} a_{1,n} \\ a_{2,n} \\ \cdots \\ a_{m,n} \end{pmatrix}$$

be the columns of the matrix A. Then there is a solution

$$T := \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix},$$

of the system (A, B) if and only if T satisfies the vectorial identity (4.39) and, since

$$\begin{pmatrix} a_{1,1}t_1 & + & a_{1,2}t_2 & + & \cdots & + & a_{1,n}t_n \\ a_{2,1}t_1 & + & a_{2,2}t_2 & + & \cdots & + & a_{2,n}t_n \\ \vdots & & \vdots & & \ddots & & \vdots \\ a_{1,1}t_1 & + & a_{1,2}t_2 & + & \cdots & + & a_{1,n}t_n \end{pmatrix} = t_1A_1 + t_2A_2 + \cdots + t_nA_n,$$

T satisfies the vectorial identity (4.39) if and only if

$$t_1A_1 + t_2A_2 + \cdots + t_nA_n = B,$$

which is equivalent to say that B is a linear combination of the columns $A_1, A_2, \ldots, A_n$. We have proven in two different ways that:

{rc}
Theorem 4.5.1 *The set of solutions of linear system of m equations and n indeterminates (A, B) is not empty if and only if the column of the constant terms B is a linear combination of the columns $A_1, A_2, \ldots, A_n$ of the matrix A of the coefficients.*

Now let us assume the set of solutions of the linear system (A, B) is not empty. What can we say about this set. We first consider the special case of a *homogeneous system*, that is a system of type $(A, 0_{\mathbb{R}^m})$ where $0_{\mathbb{R}^m}$ is the zero vector of the space $\mathbb{R}^m$, that is all constant terms are equal to 0. By what we have seen above, with B replaced by 0_{R^m}, T is a solution of $(A, 0_{\mathbb{R}^m})$ if and only if

$$T \in f_A^{-1}(0_{\mathbb{R}^m})$$

that is if and only if

$$T \in \ker(f_A).$$

Now we know a lot about the kernel of a linear map: by Lemma 4.3.1 $\ker(f_A)$ is a subspace of $\mathbb{R}^n$ and, by Theorem 4.24,

$$\dim(\ker(f_A)) = \dim \mathbb{R}^n - \dim(f_A(R^n)) = n - rk(f_A) = rk(A).$$

So we have obtained the result about the set of solutions of a homogeneous linear system

{rc1}

Theorem 4.5.2 *A homogeneous linear system $(A, 0)$ of m equations and n indeterminates has always solutions. Moreover the set of its solutions is a subspace of $\mathbb{R}^n$ of dimension $n - rk(A)$.*

Now we turn to the general case. Given a linear system (A, B) of m equations and n indeterminates, the homogeneous linear system $(A, 0_{\mathbb{R}^m})$ is called the *homogeneous linear system associated to* the system (A, B). Now assume T is a solution of (A, B) and T is any solution of $(A, 0)$, so that

$$A \cdot T = B \text{ and } A \cdot Z = 0_{\mathbb{R}^m}.$$

Then, by the distributivity of the rows by columns product withe respect to the addition of matrices,

$$A \cdot (T + Z) = A \cdot T + A \cdot Z = B + 0_{\mathbb{R}^n} = B$$

that is to say the vector $T + Z$ in $\mathbb{R}^n$ is also a solution of the system (A, B). Conversely, assume T' is a solution of (A, B), so $A \cdot T' = 0_{\mathbb{R}^n}$, and let $Z := T' - T$. Then

$$T' = 0_{\mathbb{R}^n} + T' = (T - T) + T' = T + (T' - T) = T + Z$$

and

$$A \cdot Z = A \cdot (T' - T) = A \cdot T' - A \cdot T = B - B = 0_{\mathbb{R}^m}.$$

We have proven

{rc2}

Theorem 4.5.3 *Assume the linear system (A, B) of m equations and n indeterminates has a solution T in $\mathbb{R}^n$. Then the set of solutions of (A, B) is the set*

$$T + \ker(f_A) := \{T + Z | Z \text{ is a solution of } (A, 0_{\mathbb{R}^m})\}.$$

Note that $T + \ker(f_A)$ is precisely the coset of the subspace $\ker(f_A)$ (which is also a subgroup of $\mathbb{R}^n$) of representative T.

4.5.4 Determinants

Introduction

From the previous sections it should be clear that a major problem when dealing with vectors is to determine if, given a finite set of vectors, say $\{\mathbf{w}_1, \dots \mathbf{w}_k\}$, of a vector space V, the vectors $\mathbf{w}_1, \dots \mathbf{w}_k$ are linearly independent. One way was shown In the previous section, namely check that the 0-vector can be uniquely written as a linear combination of $\mathbf{w}_1, \dots \mathbf{w}_k$. To do so, assume $x_1, \dots, x_k$ be scalars such that

{licomb}
$$0_V := x_1\mathbf{w}_1 + \cdots x_k\mathbf{w}_k \tag{4.42}$$

then

1. Choose a basis $(\mathbf{v}_1, \dots \mathbf{v}_n)$ of V and write $\mathbf{w}_1, \dots \mathbf{w}_k$ as linear combinations of the elements of this basis $(\mathbf{v}_1, \dots \mathbf{v}_n)$ of V:

{licomb1}
$$\begin{array}{ccccccccc} \mathbf{w}_1 & = & a_{1,1}\mathbf{v}_1 & + & a_{1,2}\mathbf{v}_2 & + & \cdots + & a_{1,n}\mathbf{v}_n \\ \mathbf{w}_2 & = & a_{2,1}\mathbf{v}_1 & + & a_{2,2}\mathbf{v}_2 & + & \cdots + & a_{2,n}\mathbf{v}_n \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{w}_k & = & a_{k,1}\mathbf{v}_1 & + & a_{k,2}\mathbf{v}_2 & + & \cdots + & a_{k,n}\mathbf{v}_n \end{array} \tag{4.43}$$

2. Substitute in Equation 4.42 the vectors $\mathbf{w}_1, \dots \mathbf{w}_k$ with the corresponding linear combinations of the vectors $\mathbf{v}_1, \dots \mathbf{v}_n$ obtained in Equation 4.43:

$$\begin{array}{rcl} 0_V & = & x_1(a_{1,1}\mathbf{v}_1 + a_{1,2}\mathbf{v}_2 + \cdots + a_{1,n}\mathbf{v}_n) \\ & + & x_2(a_{2,1}\mathbf{v}_1 + a_{2,2}\mathbf{v}_2 + \cdots + a_{2,n}\mathbf{v}_n) \\ & \vdots & \vdots \\ & + & x_k(a_{k,1}\mathbf{v}_1 + a_{k,2}\mathbf{v}_2 + \cdots + a_{k,n}\mathbf{v}_n) \end{array}$$

3. Collecting the coefficients of the $\mathbf{v}_i$'s in the second member of the last equation, we obtain

$$\begin{array}{rcl} 0_V & = & (a_{1,1}x_1 + a_{2,1}x_2 + \cdots + a_{k,1}x_k)\mathbf{v_1} \\ & + & (a_{1,2}x_1 + a_{2,2}x_2 + \cdots + a_{k,2}x_k)\mathbf{v_2} \\ & \vdots & \vdots \\ & + & (a_{1,n}x_1 + a_{2,n}x_2 + \cdots + a_{k,n}x_k)\mathbf{v_n} \end{array}$$

4. Since $(\mathbf{v}_1, \dots \mathbf{v}_n)$ is a basis of V, the last equation implies that all the coefficients of $\mathbf{v}_1$ up to $\mathbf{v}_n$ have to be zero:

$$\left\{ \begin{array}{ccccccccc} a_{1,1}x_1 & + & a_{2,1}x_2 & + & \cdots & + & a_{k,1}x_k & = & 0 \\ a_{1,2}x_1 & + & a_{2,2}x_2 & + & \cdots & + & a_{k,2}x_k & = & 0 \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{1,n}x_1 & + & a_{2,n}x_2 & + & \cdots & + & a_{k,n}x_k & = & 0 \end{array} \right.$$

In other words, we need to check if there are no nonzero solutions $(x_1, \ldots, x_k)$ of the homogeneous linear system whose matrix of coefficients is

$$\begin{pmatrix} a_{1,1} & a_{2,1} & \cdots & a_{k,1} \\ a_{1,2} & a_{2,2} & \cdots & a_{k,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n} & a_{2,n} & \cdots & a_{k,n} \end{pmatrix}$$

This can be effective in many situations, however there are also more efficient ways to check the linear independence of a set of vectors. One of these is by using the determinants.

Multilinear maps

Let k and n be positive integers and V be a vector space of dimension n. Set

$$V^k := \{(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n) | \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n \in V\}$$

(i.e. V^n is the set of n-tuples with entries in V). Given another vector space X, a map

$$\mu \colon V^n \to X$$

is called *k-multilinear* if the following conditions are satisfied:
For every $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k) \in V^k$, every $i \in \{1, \ldots, k\}$,

M1 if $\mathbf{w}_i = \mathbf{u}_i + \mathbf{z}_i$, then

$$\begin{aligned} & f(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{u}_i + \mathbf{z}_i, \ldots, \mathbf{w}_k) \\ = \; & f(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{u}_i, \ldots, \mathbf{w}_k) + f(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{z}_i, \ldots, \mathbf{w}_k) \end{aligned}$$

M2 for every $a \in \mathbb{R}$,

$$f(\mathbf{w}_1, \mathbf{w}_2, \ldots, a\mathbf{w}_i, \ldots, \mathbf{w}_k) = af(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_i, \ldots, \mathbf{w}_k).$$

When $X = \mathbb{R}$ a k-multilinear map is also called a (k-multilinear) *form*.

Let $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ be a basis of V. We have seen that every linear map from V to another space X is completely determined by its action on the elements of the basis $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$. Similarly k-multilinear maps are completely determined by their actions on the finite number k^n of k-tuples whose entries are in the set

$$\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$$

[9] A formal proof of this requires heavy notation and is not of great use. Instead we'll give an example which should make clear enough how a formal proof should be (and leave it to the reader). Suppose $k = 2$, $n = 3$, $(\mathbf{v}_1, \mathbf{v}_2), \mathbf{v}_3)$ is a basis

[9] Note that linear maps are precisely the 1-multilinear maps.

of V and $X = \mathbb{R}$. Then there are 9 possible pairs whose entries are in the set $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, namely

$$\begin{array}{lll} (\mathbf{v}_1, \mathbf{v}_1), & (\mathbf{v}_1, \mathbf{v}_2), & (\mathbf{v}_1, \mathbf{v}_3), \\ (\mathbf{v}_2, \mathbf{v}_1), & (\mathbf{v}_2, \mathbf{v}_2), & (\mathbf{v}_2, \mathbf{v}_3), \\ (\mathbf{v}_3, \mathbf{v}_1), & (\mathbf{v}_3, \mathbf{v}_2), & (\mathbf{v}_3, \mathbf{v}_3), \end{array}$$

Assume f is a *bilinear* (i.e. 2-multilinear) from V to $\mathbb{R}$ that maps the nine pairs above as follows:

$$\begin{array}{lll} f(\mathbf{v}_1, \mathbf{v}_1) = 1, & f(\mathbf{v}_1, \mathbf{v}_2) = 0, & f(\mathbf{v}_1, \mathbf{v}_3) = -1, \\ f(\mathbf{v}_2, \mathbf{v}_1) = -1, & f(\mathbf{v}_2, \mathbf{v}_2) = 3, & f(\mathbf{v}_2, \mathbf{v}_3) = -2, \\ f(\mathbf{v}_3, \mathbf{v}_1) = 4, & f(\mathbf{v}_3, \mathbf{v}_2) - 4, & f(\mathbf{v}_3, \mathbf{v}_3) = 2. \end{array} \tag{4.44}$$ {values}

Now, given any pair $(\mathbf{w}_1, \mathbf{w}_2)$ of vectors in V^3, its image $f(\mathbf{w}_1, \mathbf{w}_2)$ is uniquely determined by the values given in Equation (4.44): for example, if

$$\mathbf{w}_1 = \mathbf{v}_1 - \mathbf{v}_2 + \mathbf{v}_3 \text{ and } \mathbf{w}_2 = 3\mathbf{v}_1 + 2\mathbf{v}_2),$$

then, using M1, M2, and the values given in Equation (4.44), we obtain

$$\begin{aligned} f(\mathbf{w}_1, \mathbf{w}_2) &= f(\mathbf{v}_1 - \mathbf{v}_2 + \mathbf{v}_3, 3\mathbf{v}_1 + 2\mathbf{v}_2) \\ &= f(\mathbf{v}_1, 3\mathbf{v}_1 + 2\mathbf{v}_2) + f(-\mathbf{v}_2, 3\mathbf{v}_1 + 2\mathbf{v}_2) + f(\mathbf{v}_3, 3\mathbf{v}_1 + 2\mathbf{v}_2) \\ &= f(\mathbf{v}_1, 3\mathbf{v}_1) + f(\mathbf{v}_1, 2\mathbf{v}_2) \\ &+ f(-\mathbf{v}_2, 3\mathbf{v}_1) + f(-\mathbf{v}_2, 2\mathbf{v}_2) \\ &+ f(\mathbf{v}_3, 3\mathbf{v}_1) + f(\mathbf{v}_3, 2\mathbf{v}_2) \\ &= 3f(\mathbf{v}_1, \mathbf{v}_1) + 2f(\mathbf{v}_1, \mathbf{v}_2) \\ &+ (-3)f(\mathbf{v}_2, \mathbf{v}_1) + (-2)f(\mathbf{v}_2, \mathbf{v}_2) \\ &+ 3f(\mathbf{v}_3, \mathbf{v}_1) + 2f(\mathbf{v}_3, \mathbf{v}_2) \\ &= 3 \cdot 1 + 2 \cdot 0 + (-3) \cdot (-1) + (-2) \cdot 3 + 3 \cdot 4 + 2 \cdot (-4) \\ &= 3 + 3 + (-6) + 12 + (-8) \\ &= 4. \end{aligned}$$

Alternating multilinear forms

Let k, n, V and X be as in the previous subsection. A k-multilinear map $f\colon V \to X$ is called *alternating*, if f satisfies also the following condition.

MA Assume that, for $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k) \in V^n$, there exist $i, j \in \{1, \ldots, k\}$ with $i \neq j$, such that $\mathbf{w}_i = \mathbf{w}_j$, then

$$f(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k) = 0_X.$$

In other words, whenever the same vector appears twice in a k-tuple, the value of f in that k-tuple is the 0 vector of X.

Condition MA has strong consequences:

{altform}

Lemma 4.5.4 *Let $f: V \to X$ be an alternating form, $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k) \in V^n$, $i, j \in \{1, \ldots, k\}$ with $i < j$ and σ a permutation of the set $\{1, \ldots, k\}$. Then*

$$f(\mathbf{w}_{\sigma(1)}, \mathbf{w}_{\sigma(2)}, \ldots, \ldots, \mathbf{w}_{\sigma(k)}) = sgn(\sigma) f(\mathbf{w}_1, \mathbf{w}_2, \ldots, \ldots, \mathbf{w}_k)$$

PROOF.

■

4.6 Affine spaces

4.6.1 Affine planes

An affine plane is a pair

$$(\Pi, \Lambda)$$

where Π is a set (the set of *points*) and Λ (the set of *lines*) is a set of subsets of Π such that the following axioms hold:

AF1 Any two distinct points lie in a unique line

AF2 Every line contains at least two points

AF3 Given any line r and a point P not lying in r, there is a unique line r' containing P and having empty intersection with r

AF4 There exist three points P, Q, and R that are not contained in a unique line.

Define two lines r and r' to be *parallel* if their intersection is empty. Then axiom AF3 says that given a line r and a point P not in that line, there exists a unique line r' containing P and parallel to r (Euclid's well known Fifth Axiom). Finally AF4 (the nondegeneracy axiom) is needed to avoid the plane to collapse into a single line. Readers confident with Euclid's axioms will recognize that the above axioms are a part of Euclid's axioms[10] , the ones not included are those which deal with circles, i.e. those involving the concept of distance. This will be discussed in one of the next sections. Still, in an affine plane we can speak of dimension (pionts have dimension 0 and lines have dimension 1) and parallelism.

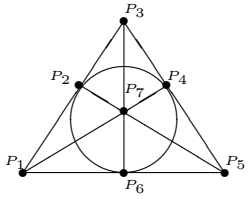
Examples:

- The standard example of a projective plane (Π, Λ) is taking Π to be the set of vectors of $\mathbb{R}^2$ (or more generally of any two dimensional vector space V) and Λ to be the set of cosets of one dimensional subspaces of $\mathbb{R}^2$ (resp.

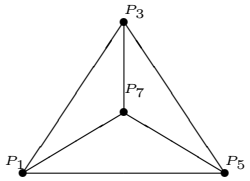
[10]Actually instead of axiom AF2 Euclid requires a line to have infinite points, but this would exclude many natural and interesting examples of finite planes without making the theory much different.

of the vector space V. We leave the reader the easy task of showing that this is indeed an affine plane. This will be discussed in more generality in the next subsection.

- Take the Fano Plane



and erase one line and all the points in that line (e.g. the circle line):



Let, in this case,

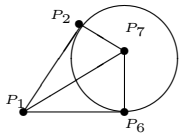
$$\Pi := \{P_1, P_3, P_5, P_7\}$$

be the set of points and

$$\Lambda := \{\{P_1, P_3\}, \{P_1, P_5\}, \{P_1, P_7\}, \{P_3, P_5\}, \{P_3, P_7\}, \{P_5, P_7\}\}$$

be the set of lines. Again one can easily check that the pair (Π, Λ) satisfies all axioms AF1-AF4. Note that, for example, the point P_7 does not belong to the line $\{P_1, P_3\}$ and the line $\{P_5, P_7\}$ is the unique line containing P_7 and parallel (i.e with empty intersection with) the line $\{P_1, P_3\}$. This is a general feature of affine spaces: we may consider the circle line as the *horizon*, and the points we have erased, i.e. the points at the horizon, as *directions*. This is because the Fano plane is the projective plane associated to the affine plane (Π, Λ), so the lines $\{P_1, P_3\}$ and $\{P_5, P_7\}$ are parallel, because their corresponding lines in the Fano plane, i.e the lines $\{P_1, P_2, P_3\}$ and $\{P_2, P_5, P_7\}$ meet at the point P_2 in the horizon, that is, they have the same direction.

Note that we would have achieved the same geometry erasing any line and all the points in that line. So, for example, taking as the horizon the line $\{P_3, P_4, P_5\}$ in the Fano plane, we would get the affine space



which is essentially the same as the one we got before (don't get confused by the curved line in the diagram, we are only interested to it as the set of points it contains, i.e. the set $\{P_2, P_6\}$. Enjoy yourself finding for each line a parallel one (actually there is a unique parallel line to a given one, for, given a line, there is only one point not in that line) and for any two parallel lines their directions.

4.6.2 Affine spaces

Let V be a vector space of dimension n. The *affine space* $A(V)$ associated to V is the set of all cosets of subspaces of V, i.e. the set of all subsets of the type

$$W + \mathbf{v} := \{\mathbf{w} + \mathbf{v} | \mathbf{w} \in W\}$$

where W varies among the subspaces of V and $\mathbf{v}$ varies in V.

1. a *point* in $A(V)$ is a coset of the subspace $\{0_V\}$, that is a set containing one vector of V;
2. a *line* in $A(V)$ is a coset of a subspace of dimension 1;
3. a *plane* in $A(V)$ is a coset of a subspace of dimension 2;
4. a *hyperplane* in $A(V)$ is a coset of a subspace of dimension $\dim(V) - 1$.

Two elements X and Y of the affine space are *incident* if either $X \subset Y$ or $Y \subset X$ (so, by this definition no element of $A(V)$ is incident with itself). Note that the subspaces of V are precisely the elements of $A(V)$ that contain $\{0_V\}$.

If $W + \mathbf{v}$ is an element of $A(V)$, with $\mathbf{v} \in V$, the *dimension* of $W + \mathbf{v}$ is the dimension of W. So points have dimension 0, lines have dimension 2, planes have dimension 3, hyperplanes have dimension $n - 1$ where n is the dimension of the vector space V (which is the dimension of the affine space $A(V)$).

Further, we say thet two elements $W_1 + \mathbf{v}_1$ and $W_2 + \mathbf{v}_2$ of $A(V)$ are *parallel* if either $W_1 \leq W_2$ or $W_2 \leq W_1$ (or, equivalently, if W_1 and W_2 are incident).

Theorem 4.6.1 *An affine space of dimension* 2 *is an affine plane*

PROOF. Let V be a vector space of dimension 2 and let $A(V)$ be the associated affine space. Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be any two distinct points. Then $\mathbf{w} := \mathbf{v}_1 - \mathbf{v}_2$ is a nonzero vector of V, so

$$W := \langle \mathbf{w} \rangle$$

is a subspace of V of dimension 1. Consider the affine subspace $W + \mathbf{v}_1$. Then $W + \mathbf{v}_1$ is a line and it contains $\mathbf{v}_2$, since

$$\mathbf{v_2} = -\mathbf{w} + \mathbf{v_1}$$

which is an element of $W + \mathbf{v}_1$. In order to prove that $W + \mathbf{v}_1$ is the unique line containing $\mathbf{v}_1$ and $\mathbf{v}_2$, assume $W' + \mathbf{v}_1$ is another line containing $\mathbf{v}_1$ and $\mathbf{v}_2$. Since $\mathbf{v}_2 \in W' + \mathbf{v}_1$, we have

$$\mathbf{v}_2 = \mathbf{w}' + \mathbf{v}_1$$

for a non zero vector $\mathbf{w}'$ in W'. In particular

$$W' = \langle \mathbf{w}' \rangle$$

and

$$\mathbf{w}' = \mathbf{v}_2 - \mathbf{v}_1 = -\mathbf{w},$$

so that

$$W = \langle \mathbf{w} \rangle = \langle \mathbf{w}' \rangle = W',$$

that is

$$W + \mathbf{v}_1 = W' + \mathbf{v}_1.$$

Thus $A(V)$ satisfies axiom AF1.

Let $W + \mathbf{v}$ be a line with W a subspace of dimension 1 of V and $v \in V$. Let $\mathbf{w}$ be any non zero vector of W. Then $W + \mathbf{v}$ contains, at least, the two distinct vectors

$$\mathbf{v} = \mathbf{0}_V + \mathbf{v} \text{ and } \mathbf{w} + \mathbf{v}.$$

Thus also axiom AF2 is satisfied.

Let $W + \mathbf{v}$ be a line in $A(V)$, with W a subspace of dimension 1 of V and $v \in V$ and let $\mathbf{v}'$ be a point not in $W + \mathbf{v}$. We prove that

$$W + \mathbf{v}'$$

is the unique line parallel to $W + \mathbf{v}$ and containing $\mathbf{v}'$. Assume U a 1-dimensional subspace of V such that $U + \mathbf{v}'$ is a line parallel to $W + \mathbf{v}$ and containing $\mathbf{v}'$. Then either

$$U \leq W \text{ or } W \leq U$$

But since both of them are 1-dimensional subspaces of V, we must have $U = W$, whence

$$U + \mathbf{v}' = W + \mathbf{v}',$$

giving axiom AF3.

Finally, since V has dimension 2, V contains two linearly independent vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ Let $U := \langle \mathbf{u_1} \rangle$ Then U is the unique line containing $\mathbf{u}_1$ and $\mathbf{0}_V$, and it cannot contain $\mathbf{u}_2$, since $\mathbf{u}_1$ and $\mathbf{u}_2$ are linearly independent. So the vectors $\mathbf{0}_V$, $\mathbf{u}_1$, and $\mathbf{u}_2$ are not contained in a unique line, giving axiom AF4. ∎

{intaff}

Theorem 4.6.2 *Let V be a vector space and let $W + \mathbf{x}$ and $U + \mathbf{y}$ be two affine subspaces of $A(V)$, with W and U subspaces of V and $\mathbf{w}$ and $\mathbf{v}$ vectors of V. Assume*

$$(W + \mathbf{x}) \cap (U + \mathbf{y}) \neq \emptyset.$$

Then

(i) for every element $\mathbf{v} \in W + \mathbf{v} \cap U + \mathbf{v}$, we have

$$W + \mathbf{x} = W + \mathbf{v} \text{ and } U + \mathbf{y} = U + \mathbf{v};$$

(ii)

$$(W + \mathbf{v}) \cap (U + \mathbf{v}) = (W \cap U) + \mathbf{v}.$$

(iii) the affine dimension of $(W + \mathbf{x}) \cap (U + \mathbf{y})$ is $\dim(W \cap U)$.

Proof. Let

$$\mathbf{v} \in W + \mathbf{x} \text{ and } \mathbf{v} \in U + \mathbf{y},$$

so by Lemma **??**

$$W + \mathbf{x} = W + \mathbf{v} \text{ and } U + \mathbf{y} = U + \mathbf{v}.$$

We prove that

$$(W + \mathbf{v}) \cap (U + \mathbf{v}) = (W \cap U) + \mathbf{v}.$$

Let $\mathbf{z} \in (W + \mathbf{v}) \cap (U + \mathbf{v})$. Then there are elements $\mathbf{w}' \in W$ and $\mathbf{u}' \in U$, such that

$$\mathbf{w}' + \mathbf{v} = \mathbf{z} = \mathbf{u}' + \mathbf{v}$$

Which forces

$$\mathbf{w}' = \mathbf{u}' \in W \cap U.$$

So

$$\mathbf{z} \in (W \cap U) + \mathbf{v}.$$

Conversely, if $\mathbf{z} \in (W \cap U) + \mathbf{v}$, then there are elements $\mathbf{w}' \in W$ and $\mathbf{u}' \in U$, such that

$$\mathbf{z} = \mathbf{w}' + \mathbf{v},$$

so $\mathbf{z} \in W + \mathbf{v}$, and

$$\mathbf{z} = \mathbf{u}' + \mathbf{v},$$

so $\mathbf{z} \in U + \mathbf{v}$, whence

$$\mathbf{z} \in (W + \mathbf{v}) \cap (U + \mathbf{v}).$$

The last assertion now follows from (ii) and the definition of affine dimension.
■

{intaff2}

Lemma 4.6.3 *Let V be a vector space. Let $W + \mathbf{x}$ be an affine hyperplane of $A(V)$ and $U + \mathbf{y}$ be an affine subspace of $A(V)$ such that $U + \mathbf{y}$ is not parallel to $W + \mathbf{x}$ (i.e. $U \not\leq W$). Then $(W + \mathbf{x}) \cap (U + \mathbf{y}) \neq \emptyset$*

PROOF. Since U is not contained in W, $W+U$ must be a subspace of V strictly containing W, and since $W + \mathbf{x}$ is an affine hyperplane, W has codimension 1 in V. It follows that

$$V = W + U$$

In particular there are vectors $\mathbf{w} \in W$ and $\mathbf{u} \in U$ such that

$$x - y = w + u$$

or, equivalently,

{inhyp}
$$(-w) + x = u + y. \tag{4.45}$$

Now, the left side of Equation (4.45) is an element of $W + \mathbf{x}$, whilst the right side is an element of $U + \mathbf{y}$, whence

$$(-w) + x = u + y \in (W + \mathbf{x}) \cap (U + \mathbf{y}) \neq \emptyset.$$

■

4.6.3 Affine transformations

4.6.4 Projective spaces

A projective space can be seen as an affine space joined with the horizon. We first construct the projective plane associated to the 2-dimensional affine space. In order to do that we use the ideas from perspective mentioned in the introduction. Recall the painter painting a sea landscape. Now choose the coordinates in $\mathbb{R}^3$ so that

1. the (pointform) eye of the painter is at the origin of the cartesian plane, i.e. is the point
$$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$
2. the sea is the "horizontal" plane Π at level -1, that is the affine plane of all vectors
$$\begin{pmatrix} a \\ b \\ -1 \end{pmatrix} \in \mathbb{R}^3$$
with a and b in $\mathbb{R}$.

In other words, if $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e_3})$ is the canonical basis of $\mathbb{R}^3$ and W is the subspace of $\mathbb{R}^3$ spanned by $\mathbf{e}_1$ and $\mathbf{e}_2$, then Π is the affine plane

$$W - \mathbf{e}_3.$$

Now let

$$P := (a, b, -1)^t$$

be a point in Π, then P is the intersection of Π with the 1-dimensional subspace U_P of $\mathbb{R}^3$ spanned by the vector $(a, b, -1)^t$. Note that U_P is not contained in W, since $P \in U_P$ while the vectors of W have the third coordinate equal to 0. Conversely, let U be a 1-dimensional subspace of $\mathbb{R}^3$ such that U is not contained in W. Then, by Lemma 4.6.2 and Lemma 4.6.3, $U \cap P = U' \cap (W - \mathbf{e}_3)$ has dimension 0, i.e. U' intersects P in a point , which we'll denote by P_U. Clearly

$$P_{U_P} = P \text{ and } U_{P_U} = U,$$

thus the maps

$$P \mapsto U_P \text{ and } U \mapsto P_U$$

are bijections, one inverse to the other, between the set of points of Π and the set of subspaces of dimension 1 of $\mathbb{R}^3$ (i.e. affine lines through $(0, 0, 0)$) that are not contained in W. Similarly, if Γ is a line in Π, denote by U_γ the 2-dimensional subspace of $\mathbb{R}^3$ generated by any two linear independent vectors of Γ. Then, as above, U_Γ is not contained in W, hence, by Lemma 4.6.2, Γ is the intersection of U_γ with Π. Conversely, if U is a two dimensional subspace of $\mathbb{R}^3$ which is not contained in W, then, as above, by Lemma 4.6.2 and Lemma 4.6.3, $U \cap \Pi$ is an affine line, which we'll denote by Γ_U. Again we have

$$\Gamma_{U_\Gamma} = \Gamma \text{ and } U_{\Gamma_U} = U,$$

thus the maps

$$\Gamma \mapsto U_\Gamma \text{ and } U \mapsto \Gamma_U$$

are bijections, one inverse to the other, between the set of lines of Π and the set of subspaces of dimension 2 of $\mathbb{R}^3$.

Let us now consider the canvas. Assume this is set vertically, that is we may think of it as the affine plane $\Delta := Z + \mathbf{e}_1$ where Z is the 2-dimensional subspace of $\mathbb{R}^3$ spanned by the vectors $\mathbf{e}_2$ and $\mathbf{e}_3$ (an infinitely large canvas!). Then, as above, the map

$$U \mapsto U \cap \Delta$$

is a bijection between the set of 1-dimensional subspaces of $\mathbb{R}^3$ not contained in Z and the set of points of Δ and the map

$$R \mapsto R \cap \Delta$$

is and between is a bijection between the set of 2-dimensional subspaces of $\mathbb{R}^3$ not contained in Z and the set of lines of Δ

In the affine space $A(\mathbb{R}^3)$ let

$$\mathbf{e}_1 := \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \ \mathbf{e}_2 := \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \text{ and } \mathbf{e}_3 := \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

so that $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ is the canonical basis for $\mathbb{R}^3$. Let W be the subspace of $\mathbb{R}^3$ spanned by the vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ and consider the affine 2 dimensional subspace $W - \mathbf{e}_3$, so that Π is the set of vectors of the form

$$\begin{pmatrix} t_1 \\ t_2 \\ -1 \end{pmatrix}.$$

Let $k \in \{1, 2\}$ and let U be any subspace of dimension k of $\mathbb{R}^3$ with $U \not\leq W$ (i.e. U is not parallel to $W - \mathbf{e}_3$). Then $W - \mathbf{e}_3 \cap U$ is an affine space of dimension $k - 1$, i.e. a point (if U has dimension 1) or a line (if U has dimension 2).

4.7 Exercises

{finitegeom1}

Exercise 4.7.1 *Prove that if $(\mathcal{P}, \mathcal{L})$ is a projective plane, then also $(\mathcal{L}, \mathcal{P})$ is a projective plane.*

{finitegeom2}

Exercise 4.7.2 *Let $(\mathcal{P}, \mathcal{L})$ be a projective plane. For a point $P \in \mathcal{P}$, let $\mathcal{L}_P$ be the set of lines incident with P and, for a line r let $\mathcal{P}_r$ be the set of points incident with r.*

1. *Prove that there exist $P \in \mathcal{P}$ and $l \in \mathcal{L}$ such that P is not incident with l.*
2. *Prove that in any finite projective plane the number of lines incident with a point P is the same as the number of points incident with a line l not incident with P.*
3. *Prove that if $|\mathcal{P}|$ is finite, the number of lines incident with a point is constant (i.e. does not depend on the chosen point).*
4. *Prove that if $|\mathcal{L})|$ is finite, the number of points incident with a line is constant.*
5. *Let $P \in \mathcal{P}$. Prove that $\mathcal{P}$ is the union of the sets $\mathcal{P}_r$ where $r \in \mathcal{L}_P$.*
6. *Prove that if either $|\mathcal{P}|$ or $|\mathcal{L}|$ is finite, then $|\mathcal{P}| = |\mathcal{L}|$ and the number of points incident with a given line is equal to the number of lines incident with a given point.*
7. *Prove that if there are exactly n distinct lines through a point, then $|\mathcal{P}| = n(n-1)+1$.*

Remark It is still an open problem to determine for which natural numbers n there exists a projective plane such that each line is incident with precisely n points. It is known (and I'll sketch it later how to prove this) that if $n-1$ is the power of a prime then such a projective plane exists, but only partial results are known for the other natural numbers (e.g. with hard machine computing it has been proved that no such plane exists for $n = 11$). If you solve this problem you will be most probably a good architect and surely an excellent mathematician.

{identities}

Exercise 4.7.3 *Prove the identities (1), (2), (3), and (4) in the subsection 4.2.1*

{b2}

Exercise 4.7.4 *Let V be a vector space, Y a finite spanning subsset of V and X a finite subset of V containing X. Then X also spans V.*

Exercise 4.7.5 *1. Prove that the vectors*

$$\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \text{ and } \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

span $\mathbb{R}^3$

2. Compute all possible unordered bases of $\mathbb{R}^3$ that are contained in the set

$$\left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right\}$$

Exercise 4.7.6 *1. Prove that the vectors*

$$\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

are linearly independent.

2. Prove that the set

$$\left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}$$

is an unordered basis of $\mathbb{R}^3$.

3. Compute all possible ways to replace two vectors in the set

$$\left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}$$

with the vectors

$$\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

so that the resulting set is still an unordered basis of $\mathbb{R}^3$.

Chapter 5

Distance

Figure 5.1: Caspar David Friedrich *Sunset*

Chapter 6

Continuity

Figure 6.1: Frank Lloyd Wright *The Solomon R. Guggenheim Museum - New York*

Bibliography

[1] Ansel Adams, *The Camera*, New York Graphic Society, 1980

[2] Michael G. Aschbacher, *Finite Group Theory*, Cambridge

[3] Euclid, *The Elements of Euclid*,Desilver Thomas & co.,Philadelphia, 1838.

[4] Leonhard Euler, *Solutio problematis ad geometriam situs pertinentis*, Comment. Acad. Sci. U. Petrop 8, 128?40 (1736).

[5] Carl Hierholzer, "Ueber die Mglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechung zu umfahren", Mathematische Annalen, **6** (1), 30?32 (1873).

[6] Nathan Jacobson *Basic Algebra 1* W. H. Freeman and Company, San Francisco 1974.

[7] Wassily Kandinsky *Punkt und Linie zu Fläche* Bauhaus Books **9** edited by Walter Gropius and Lazlo Moholy Nagy, 1926. The translation by Howard Dearstyne and Hilla Rebay is taken from the English edition published by the Solomon R. Guggenheim Foundation, 1947.

[8] John L. Kelley *General Topology* Springer Graduate texts in mathematics, **27**, New York-Heidelberg-Berlin 1955.

[9] Hans Kurzweil, Bernd Stellmacher, Finite Group Theory, and introduction Springer

[10] Mark Ronan, Symmetry and the Monster, Oxford University Press, Oxford 2006.

[11] Michio Suzuki, Finite Groups I and II, Springer

[12] Herman Weyl, *Symmetry*, Princeton University Press, Princeton N.J. 1951. (you can download the book at the following website:

http://people.math.harvard.edu/ knill teaching/mathe320_2017/blog17/ Hermann_Weyl_Symmetry.pdf)

[13] Italo Zannier, *Architettura e fotografia*, Laterza, Bari 1991