# Sets and Groups - Notes for Math3353
# Winter, 2010
# Barry Monson, UNB

## 1 Sets

### 1.1 How does mathematics grow?

In the logical development of any branch of mathematics, each definition of a concept involves other concepts or relations. Thus, the only way to avoid a vicious circle is to accept certain *primitive concepts* or relations as undefined. Likewise, the proof of each proposition (or theorem) uses other propositions. Hence, to again avoid a vicious circle we must accept certain fundamental propositions – called *axioms* or *postulates* – as true but unproved. (Here I have paraphrased a particularly nice description due to H. S. M. Coxeter.)

It is useful to keep these guiding principles in mind when reading or when engaged in a mathematical conversation, as in this course. Although we won't be pursue axiomatics very much, we will attempt to prove things with a seriousness that is appropriate to the context.

Perhaps more importantly, precision in our own thinking is crucial if we are to communicate with a computer on a mathematical level. The high speed electronic moron does not tolerate fuzzy human thinking.

### 1.2 Sets: ideas and notation

Although the starting point for the axiomatic process described above is somewhat arbitrary, most modern mathematicians begin with *set theory* and build up from there[1]. Keeping in mind the vicious circle, we realize that there is no point giving a formal definition of *set*. Instead, at the beginning, we can only discuss informal but sensible ways of thinking about sets[2]. Thus a set $A$ is any collection of objects $x$, called *elements* of the set[3]. We write

$$x \in A$$

(and say '$x$ belongs to $A$' or '$x$ is an element of the set $A$) if indeed $x$ is one of the objects in $A$. Intuitively, if $x \in A$, then $x$ and $A$ have 'different levels of organization'.

[1] In practice, we aren't deterred by 20th century discoveries, due to Gödel and other logicians, that the axiomatic method has inevitable and surprising limitations. For example, any mathematics which accepts the legitimacy of the natural numbers 1, 2, 3, . . . must contain theorems (i.e., true statements) which cannot be proved!

[2] The American mathematician Paul Halmos has written an excellent book 'Naive Set Theory', as an introduction to the foundations of mathematics for working mathematicians.

[3] To repeat: this is a way of thinking, not a precise definition.

If the object $y$ is not in $A$ we write

$$y \notin A.$$

Example.

$$\begin{aligned} A &= \text{one of your classes last term} \\ x &= \text{you (or one of your classmates)} \\ y &= \text{Steven Harper} \end{aligned}$$

so $x \in A$, $y \notin A$. Again, intuitively, the class $A$ has a 'higher level of organization'. Now imagine that students drop the class one by one. The class $A$ changes to $B$, then $C$, etc. as enrolment drops. We say $B$ is a subset of $A$, $C$ is a subset of $B$, indeed $C$ is a subset of $A$:

$$C \subseteq B \subseteq A.$$

The sets $A$, $B$, $C, \ldots$ are different but still have the 'same level of organization'. We can image that everyone drops the class, so it makes sense to allow an empty class $E$, still satisfying

$$E \subseteq A.$$

If we try to make a census of all elements $x$ of $A$, we might ask 'Which of you were born in Fredericton?' or 'Which of you are Math. majors?' or 'Which of you have blond hair?', etc. We wouldn't count twice a person who answered yes to two or more questions. Thus, as is reasonable, we shall agree that:

'order and repetition are irrelevant when
assessing elements in a set'.

(If order and repetition are important, we instead employ a *list*, which is really a *function*, which is really a very special kind of set! See below.)

**Definitions.** Suppose $A$, $B$, etc. are sets.

1. $A$ is a *subset* of $B$, written $A \subseteq B$, if every element of $A$ is an element of $B$:

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

2. $A$ *equals* $B$, written $A = B$, if $A \subseteq B$ and $B \subseteq A$:

   $$\begin{aligned} x \in A &\Rightarrow x \in B \\ x \in B &\Rightarrow x \in A \end{aligned}$$

   In brief, $x \in B$ if and only if $x \in A$. Intuitively, $A$ and $B$ have the same elements.

3. An *empty set* $E$ has no elements.

## 1.3 We don't worry much about the axioms for set theory

This isn't a course in set theory, so we won't say much. Instead, a sensible approach is to learn the material informally through examples and by proving simple theorems, more or less ignoring the axioms.

**However**, we do note that in order to avoid paradoxes we must insist that

$$x \in x$$

**is a meaningless statement** for any mathematical object $x$. Intuitively, $x$ cannot be an element of itself, because it would then simultaneously have two different levels of organization. However,

$$x \in \{x\}$$

is always true; and

$$x \subseteq x$$

is also perfectly okay, so long as $x$ itself is a set.

**Theorem.** The empty set is unique: if $E$ and $E'$ are empty sets, then $E = E'$.

Proof. Convince yourself. □

**Notation.** When an interesting object is shown to be uniquely specified, it often deserves a special notation. The empty set is denoted

$$\emptyset \, .$$

**Exercise**. For any set $A$ whatsoever, prove that

$$\emptyset \subseteq A \, .$$

**More Definitions.**

4. The *union* of two sets $A$, $B$ is the set of all objects in either $A$ or $B$ (or both):

$$A \cup B = \{x : x \in A \ \text{ or } \ x \in B\}.$$

5. The *intersection* of sets $A$, $B$ is the set of all objects in both $A$ and $B$:

$$A \cap B = \{x : x \in A \ \text{ and } \ x \in B\}.$$

6. Sets $A$ and $B$ are *disjoint* if $A \cap B = \emptyset$.

**Remark.** There are similar definitions for any finite family of sets

$$A_1 \cup \ldots \cup A_k \quad \text{or} \quad A_1 \cap \ldots \cap A_k,$$

or even any indexed family of sets $A_t$, where $t \in \mathcal{I}$. The indexing set $\mathcal{I}$ could be infinite. In general then, we write

$$\bigcup_{t \in \mathcal{I}} A_t \quad \text{or} \quad \bigcap_{t \in \mathcal{I}} A_t.$$

***

Sometimes we can explicitly enumerate the elements of a set, as in

$$A = \{5, 6, 7\},$$

or even in

$$Sq = \{1, 4, 9, 16, \ldots\}.$$

(The use of "…" assumes the pattern is clear.) Maybe an explicit description is better as in

$$Sq = \{n \in \mathbb{N} : \; n = a^2, \;\; \text{for some} \;\; a \in \mathbb{N}\}.$$

Thus $Sq$ is a subset of the natural numbers $\mathbb{N}$. Now is a good time to establish some

**Standard Notation for Important Sets**

- $\mathbb{N} = \{\text{natural numbers}\} = \{1, 2, 3, \ldots\}$
- $\mathbb{Z} = \{\text{all integers}\} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$
- $\mathbb{Q} = \{\text{rational numbers}\} = \{\frac{m}{n} \; : \; m, n \in \mathbb{Z}, n \neq 0\}$
- $\mathbb{R} = \{\text{real numbers}\}$
- $\mathbb{C} = \{\text{complex numbers}\} = \{x + \imath y \; : \; x, y \in \mathbb{R}\}$
  The complex numbers thus provide an alternate way to look at the Euclidean plane.

Of course, we also can identify the plane in the usual way with

- $\mathbb{R}^2 = \{\, (x, y) \, : \, x, y \in \mathbb{R}\}$

  A typical element of $\mathbb{R}^2$ is therefore an ordered pair $(x, y)$ (more on that below). By 'usual way', we mean that we set up our coordinates after first chosing an origin, then rectangular axes with unit points.

  Sometimes we want to interpret $\mathbb{R}^2$ as a *vector space*. We commonly use square brackets as a visual reminder that the ordered pair $[x, y]$ is to be treated as a vector.

  Likewise we may describe ordinary Euclidean space by ordered triples:

- $\mathbb{R}^3 = \{\, (x_1, x_2, x_3) \, : \, x_1, x_2, x_3 \in \mathbb{R}\}$

  We may even have a look at the unit sphere:

- $\mathbb{S}^2 = \{\, (x_1, x_2, x_3) \in \mathbb{R}^3 \, : \, x_1^2 + x_2^2 + x_3^2 = 1\}$

  Thus $\mathbb{S}^2$ is a subset of $\mathbb{R}^3$.

Exercises.

1. Let $A = \{2, \{1\}\}, \;\; B = \{\{\emptyset, \{3\}\}\}$.

   (a) What are the elements of $A$?

   (b) What is the cardinality of $A$ (number of distinct elements)?

   (c) What are the distinct elements of $B$? What is its cardinality?

2. (a) What is cardinality of $\emptyset$?

   (b) Of $\{\emptyset\}$?

   (c) Of $\{\emptyset, \{\emptyset\}\}$?

3. We know $\emptyset \subseteq \mathbb{N} \subseteq \mathbb{Z}^{\geq} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$. How many subset relations of the form $A \subseteq B$ are there between these sets?

4. Let

$$\begin{aligned} A &= \{5,6,7\} \\ B &= \{5,7\} \\ C &= \{6,6,5,7,5,7\} \end{aligned}$$

<u>True or False:</u>

$$\begin{aligned} A &= B \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \\ A &= C \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \\ C &\subseteq A \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \\ C &\subseteq B \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \\ B &\subseteq A \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \\ B &\neq A \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \\ B &\in A \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \\ A &\subseteq A \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \\ \{6\} &\subseteq A \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \\ \{6\} &\in A \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \\ 6 &\subseteq A \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \\ 6 &\in A \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \\ \emptyset &\subseteq A \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \\ \emptyset &\in A \text{ \_\_\_\_\_\_\_\_\_\_\_\_} \end{aligned}$$

5. Is it possible that

$$x \in A \quad \text{and} \quad x \subseteq A$$

are both true?

6. Find out what the subset lattice of a set $A$ is and sketch it, when $A = \{1,2,3\}$.

## 1.4 Bulk operations on sets

**Definitions.** If a set $A$ comes with an operation, say "+", then we can 'add' subsets of $A$. Suppose $B \subseteq A$ and $C \subseteq A$ (two subsets of $A$). Then by definition

$$B + C := \{x + y : x \in B \text{ and } y \in C\}.$$

This means *add elements from* $B$ and $C$ in that order *and in all possible ways* .

The sets $B$, $C$ could be finite or infinite. In particular, $B$ could be a singleton (cardinality 1), say

$$B = \{b\}.$$

Then instead of $\{b\} + C$ we write $b + C$ for more pleasant reading.

Similarly, if set $A$ comes equipped with a multiplication "×", we might suppress the operation:

$$\begin{aligned} BC &:= \{xy : x \in B \text{ and } y \in C\} \\ bC &:= \{by : y \in C\}. \end{aligned}$$

Exercises (Continued).

6. Describe, say by a 'clearly understood' listing, these subsets of the integers $\mathbb{Z}$.

   (a) $3\mathbb{Z}$

   (b) $1 + 2\mathbb{Z}$

   (c) $12\mathbb{Z} + 21\mathbb{Z}$

   (d) For specific positive integers $a$, $b$, what is $a\mathbb{Z} + b\mathbb{Z}$ in general?

7. Describe these subsets of the reals $\mathbb{R}$:

   (a) $\sqrt{2}\,\mathbb{Z}$

   (b) $\mathbb{Z} \cap (\sqrt{2}\,\mathbb{Z})$.

8. An example from Euclidean Geometry.

   A typical triangle will be denoted $\triangle ABC$. For simplicity, let $A, B, C$ denote the angles and let $a, b, c$ be the lengths of the opposite edges. Let

$$\begin{aligned} U &= \{\triangle ABC : C = 90^\circ\} \\ V &= \{\triangle ABC : a^2 + b^2 = c^2\} \end{aligned}$$

   In fact $U = V$.

   (a) Rephrase $U \subseteq V$ as a geometrical theorem. What is its conventional name?

   (b) Restate $V \subseteq U$ as such a theorem. (This is the *converse* to the theorem in (a).)

   (c) Restate $U = V$ using 'if and only if' lingo. Using 'necessary and sufficient' lingo.

9. In a vector space, like

$$\mathbb{R}^2 = \{\mathbf{u} = [x, y] : x \in \mathbb{R},\ y \in \mathbb{R}\}$$

   we have two operations:

$$\begin{aligned} \mathbf{u_1} + \mathbf{u_2} &= [x_1, y_1] + [x_2, y_2] := [x_1 + x_2, y_1 + y_2] \\ t\mathbf{u} &= t[x, y] := [tx, ty] \end{aligned}$$

   (component-wise addition and scalar multiplication, for scalars $t \in \mathbb{R}$). We will often use bold type to distinguish vectors, as in $\mathbf{u}$. In class I may use arrows, as in $\vec{u}$.

   Give geometrical descriptions for

   (a) $\mathbb{R}[2, 1]$ (strictly speaking, we here mean $\mathbb{R}\{[2, 1]\}$)

   (b) $[-1, 1] + \mathbb{R}[2, 1]$

   (c) $\mathbb{Z}[1, 0] + \mathbb{Z}[0, 1]$

   (d) $\mathbb{Z}[1, 0] + \mathbb{Z}\left[-\frac{1}{2}, \frac{\sqrt{3}}{2}\right]$

   (e) $\{[x, y] : x \in \mathbb{Z} \text{ and } y \in \mathbb{Z}\}$

   (f) $\{[x, y] : x \in \mathbb{Z} \text{ or } y \in \mathbb{Z}\}$

## 1.5 The Cartesian Product of Sets

**Definition**. For any two sets $A, B$, the (*Cartesian*) *product* $A \times B$ is the set of all ordered pairs $(a, b)$ such that $a \in A$ and $b \in B$:

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

**Example**.

$$\{0, 1\} \times \{x, y, z\} = \{(0, x), (0, y), (0, z), (1, x), (1, y), (1, z)\} .$$

Similarly,

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

is the set of all *ordered* $n$-tuples $(a_1, \ldots, a_n)$ with $a_j \in A_j$, $1 \leq j \leq n$. In the special case that all sets are the same, say $A = A_1 = \ldots = A_n$, we often write $A^n$ instead.

**Example**.

$$\mathbb{R}^2 = \{ [x_1, x_2] \, : \, x_1, x_2 \in \mathbb{R}\} .$$

(Recall once more that the square brackets are commonly used as a visual reminder that the ordered pair is to be treated as a vector.) Of course, $\mathbb{R}^2$ is a very infinite set. But our ideas extend to sets of any cardinality.

## 1.6 Relations

**Definition**. A *relation* $\mathcal{R}$ from a set $A$ to a set $B$ is merely any subset of $A \times B$:

$$\mathcal{R} \subseteq A \times B$$

To indicate that $(a, b) \in \mathcal{R}$ we write

$$a\mathcal{R}b .$$

As we shall see below, a *function* $f : A \to B$ is a very special sort of relation.

Very often we have $A = B$; some extremely useful relations in this case are called *equivalence relations* and *partial orders* on $A$.

**Exercises**

1. If there are $a$ distinct elements in set $A$, $b$ elements in $B$, $c$ elements in set $C$, then how many distinct elements are there in $A \times B \times C$?

2. If $n$ is a positive integer and there are $a$ elements in the set $A$, then how many elements are there in $A^n$?

***

Using the rather abstract point of view from above, analyze the following familiar relations.

3. The usual *total order* '<' on the reals $\mathbb{R}$ can be defined as follows:

$$< \; := \; \{(x, y) \in \mathbb{R}^2 : y - x \text{ is positive.}\}$$

   (Presumably in constructing the reals we have somewhere been told which of them are 'positive'). Sketch $<$ as a subset of $\mathbb{R}^2$.

4. On the positive integers $\mathbb{N}$ define the usual *divisibility* relation '$|$' by $a|b$ if $a$ divides $b$ (without remainder).

   Sketch $|$ as a subset of $\mathbb{N}^2$ (itself a subset of $\mathbb{R}^2$).

   How can you tell from your sketch that a number $b$ is prime?

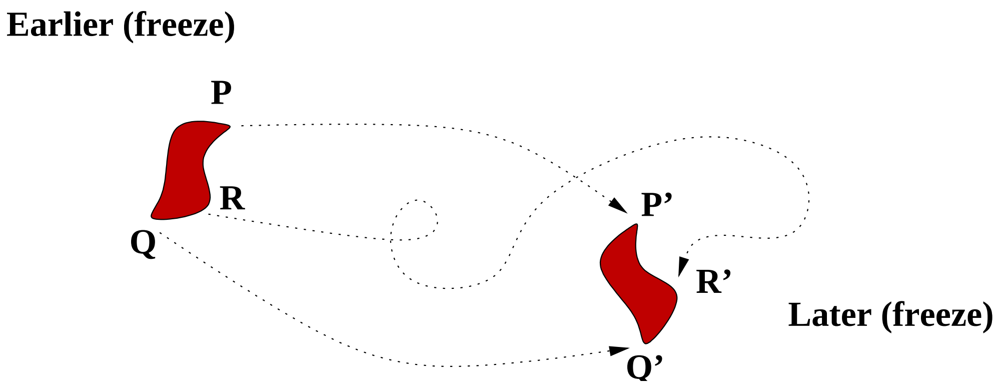
# 2 Functions

## 2.1 Motion and Symmetry - Thinking about Functions

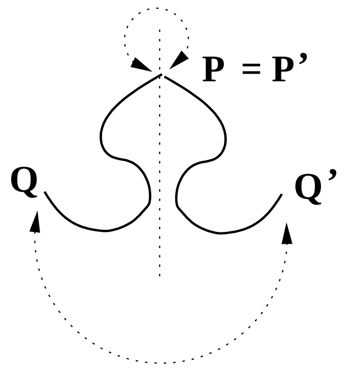
Recall that we let $\mathbb{R}^2$ denote (the set of points in) the *Euclidean plane.* In fact, many of our results will extend to Euclidean spaces of higher dimension.

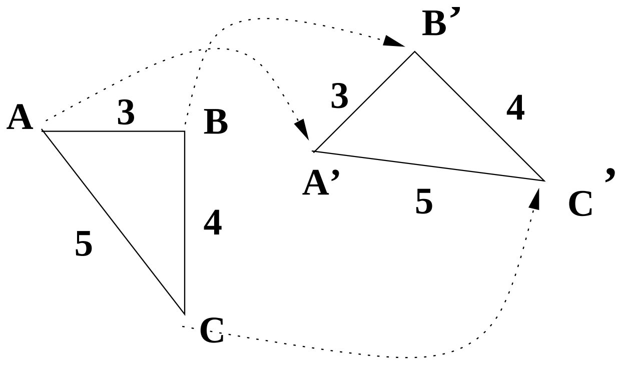
Let us try to make mathematical sense of motions and symmetry.

1. Think about **motion** of a figure in the plane. Freeze a couple of positions.



Or consider **symmetry** or **congruence**:





In all these examples, the geometrical operation gives a function $f$ mapping $P$ to $P'$, $A$ to $A'$, etc. The function preserves shape; specifically, it preserves the distance between pairs of points in the figure:

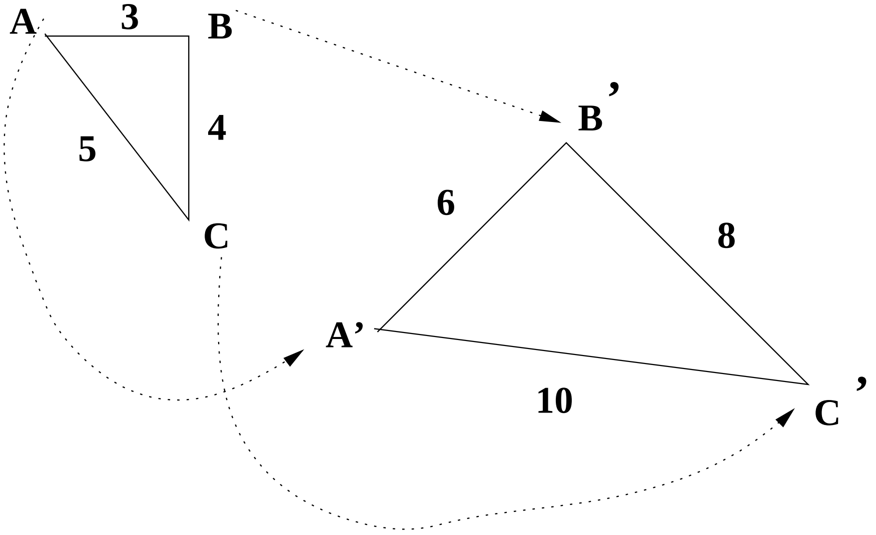
$$PQ = P'Q'$$

for all points $P$, $Q$ in the figure. Since the distance between constituent points is *invariant*, the shape as a whole is unchanged.

A function like this which is respectful of distance is called an *isometry*.

Every motion or symmetry can be reversed: just reverse the arrows and map $P'$ to $P$, $A'$ to $A$, etc. We get the function $f^{-1}$. It too will be respectful of distance; so $f^{-1}$ is also an isometry (closely related to $f$, of course).

2. The idea of **similarity** is much like this, except that every distance now is rescaled by a constant real factor $\lambda \neq 0$. In the picture that follows, $\lambda \approx 1.84$.



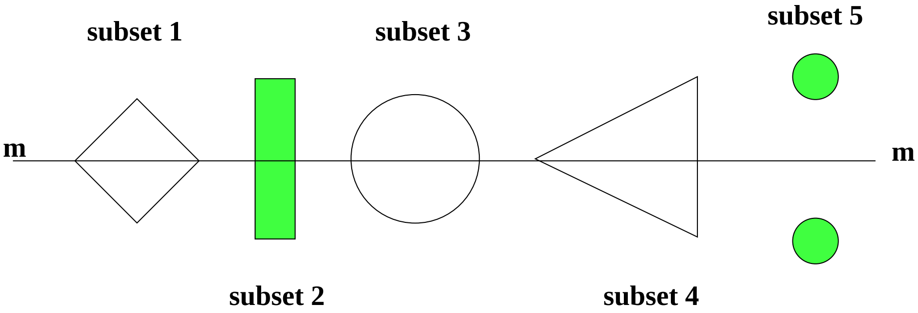
This magnifying function $f$ alters distance, but in a uniform way, so that 'shape' is preserved in some sense. We say $f$ is a *similarity*. It, too, is a nice function. Note that $f^{-1}$ is also a similarity, whose magnifying factor is $\dfrac{1}{\lambda} \approx 0.54$.

3. A general function $f$ might randomly rearrange points of the plane and so utterly destroy structures in the plane. This would be of little use in geometry. Thus we will want to study only 'nice' functions, where 'nice' will depend on our needs and will have to be properly defined.

4. What we have then are certain nice functions mapping a subset of $\mathbb{R}^2$ to another (possibly the same) subset of $\mathbb{R}^2$. It is convenient to start with functions mapping all of $\mathbb{R}^2$ to $\mathbb{R}^2$, since we can then simply restrict our attention to any particular subset of interest.

   For example, a symmetry of the square is really descended from an isometry of the whole plane; but it may suit us to ignore what that isometry does outside the square.

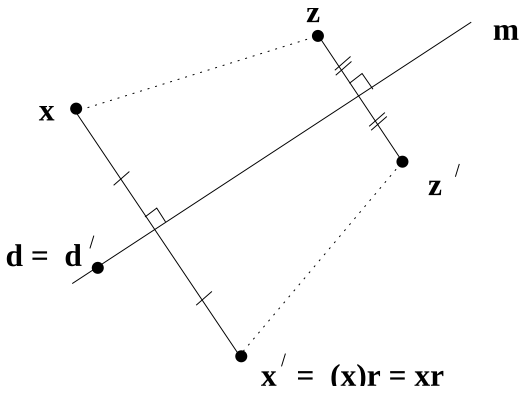
   In the figure below, reflection in the line $m$ preserves not only the whole plane, but also a variety of figures, each of which is said to have *bilateral symmetry.*



5. Reflections are key examples of nice geometrical functions. How, in fact, do we describe a reflection as a function $r : \mathbb{R}^2 \to \mathbb{R}^2$?

   **Definition: the reflection $r$ in a given line $m$**

   Let $m$ be any line in the plane $\mathbb{R}^2$. For each point $P \in \mathbb{R}^2$ let $P' = (P)r$ be the *mirror image* of $P$ by reflection in $m$. More precisely, on the line which passes through $P$ and is perpendicular to $m$ we locate $P'$ an equal distance from $m$ on the opposite side:



   **Remarks**. We check that $r : \mathbb{R}^2 \to \mathbb{R}^2$ is well-defined. This means that our purely geometric definition is unambiguous. For example, there can't be two perpendiculars to $m$ through $P$, so we know exactly where to locate $P'$. (But consider the north pole and equator on the sphere; something subtle is going on!).

   It is easy to believe and not hard to prove that a reflection $r$ is an isometry: reflections preserve distances. For example, the distance from $P$ to $Q$ equals the distance from $P'$ to $Q'$. (In vector language we have $\| \overrightarrow{PQ} \| = \| \overrightarrow{P'Q'} \|$.) Of course, the same is true for <u>any</u> pair of points in the plane. We therefore say that $r$ is an isometry.

   **Exercise**. Prove that $r$ is an isometry.

   We should note some important properties of reflections. First of all, we observe that a point $D$ is *fixed* or *invariant*, in brief, $(D)r = D$, precisely when $D$ lies on $m$. (Prove that this really is what our definition for $r$ says when $D \in m$.)

   We also note that $r$ is an *opposite* (= orientation-reversing = sense-reversing) isometry. The image of any anti-clockwise oriented triangle is a clockwise oriented triangle, and vice-versa.

6. We hinted earlier that 'most' functions $f : \mathbb{R}^2 \to \mathbb{R}^2$ will destroy desirable geometric properties and hence would be of no use to us. Even some simple examples have their faults:

   **Example**. Fix any one point $O$, say, in the plane $\mathbb{R}^2$. Define a *constant* function

$$k : \mathbb{R}^2 \to \mathbb{R}^2$$

   which maps *every* point $P$ to $O$. Note that our definition, naturally enough, is geometric, rather than being produced by an 'algebraic formula'. Note also that the *range* of the function is the one-point set $\{O\}$.

Intuitively, the whole plane is collapsed onto a single point, and we have no way of recovering from this situation. We say that

$$k \text{ has no inverse.}$$

The function is boring from a geometrical point of view. Of course, constant functions are very useful in physics and calculus; for example, the acceleration due to gravity is more or less constant near the the surface of the earth.

7. In class we shall take a closer look at the symmetries of a simple geometric obeject like a square and use that discussion to motivate the more formal descriptions which follow.

## 2.2 So what is a function?

1. **Definitions**. Let $X, Y$ be two sets, finite or infinite. A *function* $f$ is any rule[4] which associates to each *input* element $x$ in $X$ exactly one *output* element, traditionally denoted $f(x)$, in $Y$. In brief, we write

   $$f : X \to Y \ \ .$$

   The set $X$ is the *domain* of $f$.

   Note that the *range*

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

   could be a *proper* subset of $Y$.

   **Definition**. If in fact $f(X) = Y$, then we say $f$ is *onto* or *surjective.*

2. **New Notation** Some arguments are easier to read if we write $x'$ or $y$ or some other letter in place of of $f(x)$. We could write

   $$x \xrightarrow{f} x'$$

   to indicate this. More radically, we will often use the 'algebra-friendly' notation

   $$(x)f \text{ or more simply } xf$$

   instead of $f(P)$.

   Why is this 'natural'?

[4] If you object to the somewhat vague term 'rule', you should note that it is quite possible to give a very precise, but less intuitive definition: the function $f$ is actually a subset of $X \times Y$, with the property that each $x \in X$ is the first entry in exactly one ordered pair $(x, y) \in f$. In other words, $f$ is a special sort of relation from $X$ to $Y$, so

$$f \subseteq X \times Y \ .$$

In fact, the set inclusion has to be proper here, unless $Y$ has what cardinality?

**Example from basic algebra**. Suppose $X = \mathbb{R}^{\geq 0} = \{x \in \mathbb{R} : x \geq 0\}$ is the set of non-negative real numbers. In fact, we will take $Y = X$, too. Now let $f$ be the 'rule' which takes a non-negative real number $x$ as input and outputs the unique non-negative real whose square equals $x$. What function is this?

(Answer: we have described the square root without using a formula.) Thus

$$(4)f = 2$$

or

$$25f = 5\ .$$

Can you rewrite the statement $(ab)f = (af)(bf)$ in traditional notation?

3. We could use the symmetries of geometrical objects to motivate the next idea. Instead, let us think about a more whimsical

   **Example**. Consider the airplane booking function $f$ from the passenger set $X$ on a particular flight to the seat set $Y$ on the airplane: $(x)f = xf$ is the seat occupied by person $x$, for each passenger $x \in X$. Thus, $f$ onto means that every seat is filled. On the other hand, we don't want the flight to be stupidly booked: no two people should be assigned the same seat! A function with this sensible property is said to be $1-1$. Clearly, if $f$ is both $1-1$ and onto, then there is exactly one seat for each passenger; in other words, the number of seats is the same as the number of passengers. Note that you could 'see this' by simply looking at the cabin, without counting or knowing the number of passengers or seats. These considerations motivate the following definitions and observations.

4. **Definitions**. If $f$ maps distinct inputs to distinct outputs, i.e.

$$x_1 \neq x_2 \Rightarrow x_1 f \neq x_2 f\ ,$$

   then we say that $f$ is $1{-}1$ (or *injective*). In contrapositive fashion, we could equivalently say that

$$x_1 f = x_2 f \Rightarrow x_1 = x_2\ .$$

   A function $f : X \to Y$ is *bijective* if it is both $1-1$ and onto.

5. If $f : X \to Y$ is a bijection, then the domain $X$ and range $Y$ must have the same cardinality, even if both are infinite. Put otherwise, $f$ defines a *1–1 correpondence* between the elements of $X$ and the elements of $Y$. In this manner, by construction of suitable functions, we can assess whether certain infinite sets have or do not have the same 'number' of elements.

**Examples**

- $\mathbb{N}$ and $\mathbb{Z}$ have the same cardinality.
- $\mathbb{N}$ and $\mathbb{Q}$ have the same cardinality. Thus we can *count* the rational numbers even though they seem to fill out an ordinary line.
- $\mathbb{Q}$ and $\mathbb{R}$ do not have the same cardinality: the reals have a 'higher order of infinity'; they are *uncountable*.

  **Remark**. This definitely hampers our ability to represent 'generic' real numbers on the computer. Typically, with the aid of *floating point arithmetic*, we make do with finite (but perhaps rather good!) approximations. This avenue is more appropriate to a course in numerical analysis.

6. **Exercises**

   (a) Exhibit a bijection from the open real interval $(0, 1) = \{x \in \mathbb{R} : 0 < x < 1\}$ to $\mathbb{R}$ itself.

   (b) Exhibit a bijection between the the set $\mathbb{N}$ (of all natural numbers) and the proper subset $Y = \{2, 4, 6, 8, \ldots\}$ of just the even natural numbers.

   (c) If $X$ is finite, say with $n$ elements, how many bijections $f : X \to X$ are there? (Think: given $n$ people in $n$ chairs, how many ways can they rearrange themselves?)

7. Every geometric object has at least one symmetry, namely that in which we leave the object alone. After all, if no point moves, then the object will look the same before as after! This motivates the next definition, which applies to any set $X$, not just to geometrical figures.

   **Definition**. The *identity function*

   $$1 : X \to X$$

   satisfies $(x)1 = x$ for all $x \in X$. We write $1_X$ if we need to emphasize the domain $X$.

   **Remarks**. Identity functions are modest but very useful; they play the same role in symmetry that the number '1' does for the multiplication of real numbers, or that the number '0' does in addition.

   Here we must be careful - the symbol $1 = 1_X$ will often not have a numerical meaning. Instead, it will frequently indicate a trivial symmetry. Thus we must always be aware of our algebraic context.

   **Exercises**.

   (a) Graph the function $1_{\mathbb{R}}$ in the usual way of first-year calculus.

   (b) Prove that $1_X : X \to X$ is a bijection.

   Think about this before you peak at the

   Solution. We must show that $1_X$ is 1–1 and onto. Let $t$ be *any* element of the receiving set $X$. Then $t$ also belongs to the input set (the same $X$), and $(t)1_X = t$, by definition. Thus $1_X$ is onto.

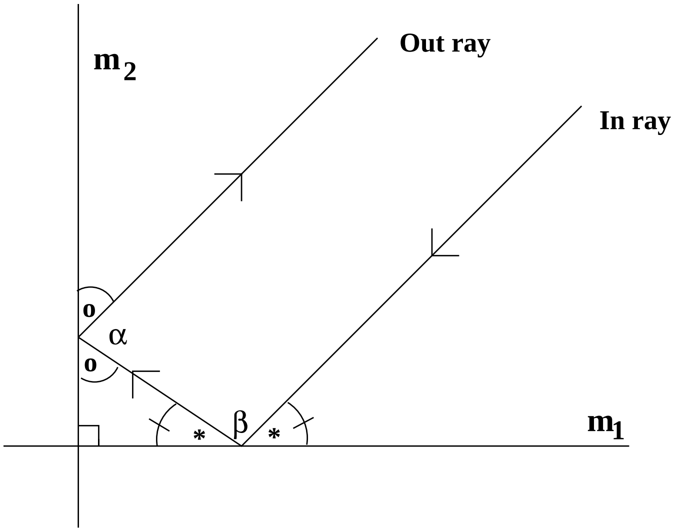
   The proof that $1_X$ is 1–1 is just as easy. □

8. We have observed that a plane reflection can be regarded as a function

$$r : \mathbb{R}^2 \rightarrow \mathbb{R}^2 \ .$$

We use reflections in the next motivational example.

9. **Problem - perpendicular mirrors**. What happens when a light beam bounces successively off two perpendicular mirrors? In other words, what is the net effect of reflections $r_1$ and $r_2$ in two perpendicular mirrors $m_1$ and $m_2$?

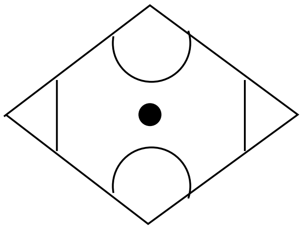


Consider *any* light ray hitting (and bouncing from) $m_1$ at angle $*$. The two equal angles at $m_2$ are $\circ = 90^\circ - *$, so that

$$\begin{aligned} \alpha + \beta &= \alpha + [180^\circ - 2*] \\ &= \alpha + 2[90^\circ - *] \\ &= \alpha + 2\circ \\ &= 180^\circ. \end{aligned}$$

Hence, the in- and out- rays are parallel. In other words, the light beam returns to its source regardless of how that source is positioned relative to the mirrors. A configuration of mirrors similar to this is used in conjunction with lasers to measure distances very accurately. For example, this is how the distance to the moon is computed accurate to a few meters.

10. **Kaleidoscopic symmetry**. Two mirrors like this form a simple *kaleidoscope*. The following plane figure is symmetrical by reflection in precisely two perpendicular mirrors.



Using the orbifold notation of J. H. Conway, et al, we would say that the figure has

$$*2\bullet$$

symmetry, which we read as 'star two point' symmetry. The $*$ is meant to suggest the various mirrors of symmetry; the 2 indicates the number of such mirrors; the $\bullet$ indicates the point at which such mirrors cross.

**Important**: this point common to the mirrors is fixed by reflection in each mirror, indeed is fixed by all symmetries of the figure. We would not use a $\bullet$ for a pattern with no point fixed by all symmetries (example: the bricks of a typical infinite wall).

11. **Products of Functions** The previous discussion suggests that it is worthwhile studying the net effect of two functions.

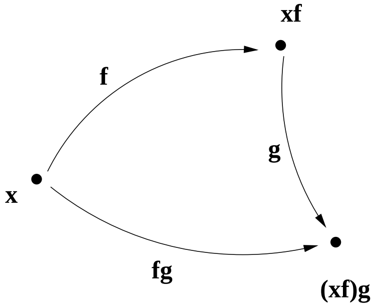
**Definition**. Suppose $f : X \to Y$ and $g : Y \to Z$ are any two functions in which the target set $Y$ for $f$ equals the domain of $g$. Then the *product*

$$fg : X \to Z$$

is the function obtained by first applying $f$ then $g$ (as we read left to right). More precisely, for all points $x \in X$,

$$x(fg) = (xf)g \; .$$

Diagramatically we have



12. **Remarks and examples**

    (a) This is just composition in left-to-right notation. Compare the standard description

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

    Recall '$g \circ f$' traditionally means 'first $f$ then $g$'; we are writing this as $fg$.

    (b) Note that for $fg$ to be defined, the target set for $f$ should be a subset of the domain set for $g$. Here, for convenience, we take both equal $Y$; this is a harmless restriction.

    Typically we shall be concerned with the special case that $X = Y = Z$. Then domain issues subside and the product (i.e. composition) is always defined.

    (c) **Example** Say $r : \mathbb{R}^2 \to \mathbb{R}^2$ is reflection in the line $m$. Note that reflecting twice in succession returns each point to its original position. Put more precisely, this means that

    $$((P)r)r = P = P1$$

    for each point $P \in \mathbb{R}^2$. (Here, of course, 1 denotes the identity $1_{\mathbb{R}^2}$ on $\mathbb{R}^2$.)

    Thus, $rr = r^2$ has the same effect on *every* point in the plane as the identity 1. On these grounds we naturally write

    $$r^2 = 1 \; .$$

    A mapping like $r$ which has period 2 is called an *involution*; we also say that a reflection has *period* 2.

13. **Definition**. An function $f : X \to X$ has *period* $n$ if $f^n = 1$, for some integer $n > 0$, where in fact $n$ is the smallest such positive integer. Here $f^n$ means multiply (i.e. compose) $f$ $n$ times in succession.

14. **Exercises**.

    (a) What is the period of any identity function $1 : X \to X$?

    (b) If $r$ is a reflection, what does $r^{203}$ equal? What about $r^{1268}$?

    (c) Suppose

    $$\begin{aligned} g : \mathbb{R} &\to \mathbb{R} \\ x &\mapsto x^2 \end{aligned}$$

    (a familar function from Calculus). Does $g$ even have a period?

    (d) Suppose $f : X \to X$ satisfies $f^m = 1$, for some integer $m > 0$. Prove that $f$ is a bijection. (Note that by context 1 here indicates the identity function on $X$.)

15. **Properties of Multiplication (Composition) of Functions**. Suppose $f, g, h$ etc. are functions for which the compositions indicated below are defined. Then we have

   (a) *Associativity*: If $f : X \to Y$, $g : Y \to Z$ and $h : Z \to W$, then always

   $$(fg)h = f(gh) \ .$$

   (b) *Possible failure of commutativity*: usually $fg \neq gf$, even in the familiar case when both $f, g : X \to X$.

   (c) *Well-behaved identities*: For any $f : X \to Y$,

   $$f1_Y = f = 1_X f \ .$$

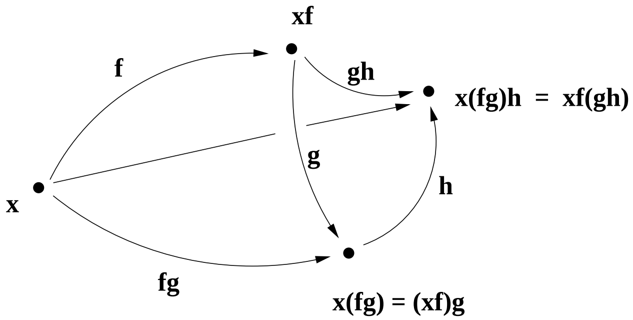
   Remark: we might write $f1 = f = 1f$, but then the symbol 1 means two different things if $X \neq Y$.

   (d) If $f, g$ are both 1–1, so is $fg$.

   (e) If $f, g$ are both onto, so is $fg$.

   (f) If $f, g$ are both bijective, then so is $fg$.

**Proof**. This is routine. The associativity comes from carefully looking at



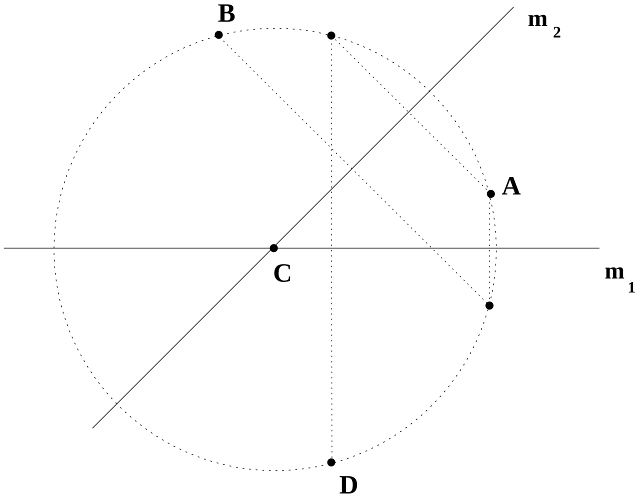
**The straight arrow shows why**
**f(gh) = (fg)h**

Now consider part (e), assuming that both $f : X \to Y$ and $g : Y \to Z$ are onto. Then $fg :\to Z$ is certainly defined. To prove that this function is onto, we must *work backward*. So consider *any* $z \in Z$. Since $g$ is onto, there exists $y \in Y$ with $yg = z$. And since $f$ is onto, there exists $x \in X$ with $xf = y$. Putting this all together, we get $x(fg) = (xf)g = yg = z$. Thus $fg$ is onto.

Part (d) concerning 1–1 is similar; and parts (d) and (e) immediately give part (f).

□

16. **An example: the product of two reflections**. Let $r_1$ and $r_2$ be reflections in lines $m_1$ and $m_2$ which intersect in the point $C$ at a 45° angle.



Since both reflections fix $C$, the two products $r_1r_2$ and $r_2r_1$ also fix $C$; they have that much in common.

However, look at any other point like $A$. We see right away that

$$A \overset{r_1r_2}{\rightarrow} B \text{ and } A \overset{r_2r_1}{\rightarrow} D \ ,$$

where $B$ and $D$ are different points. Hence, we already have that $r_1r_2 \neq r_2r_1$.

**Exercise**. Prove that $C$ is the midpoint of segment $BD$.

17. **Inverses - getting in and out of trouble** Our intuition tells us that ordinary motions are reversible. Similarly, if an object has a symmetry which rearranges the points of the object in a particular way (without altering the appearance or position of the object), then that symmetry should likewise be reversible and should restore each point to its original location.

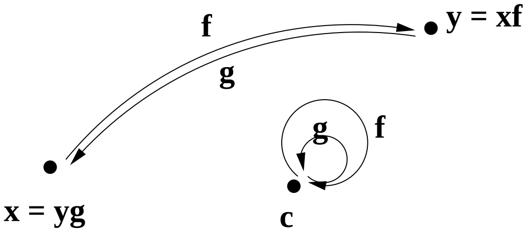
These ideas motivate the idea of an inverse function. Suppose $f : X \rightarrow Y$ is some function. Let us imagine what an inverse function should do for $f$. For any and all elements $x \in X$, if

$$x \xrightarrow{f} y \ ,$$

then the inverse function, which we temporarily call $g$, should restore matters:

$$y \xrightarrow{g} x \ .$$

Likewise, if $a \xrightarrow{f} b$, then $b \xrightarrow{g} a$. Or consider a fixed point: if $c \xrightarrow{f} c$, then $c \xrightarrow{g} c$, as well. In other words, $f$ and $g$ restore one another:



More technically every element is fixed by applying say $f$ then $g$ (or vice versa).

**Definition**. If $f : X \to Y$, then an *inverse* for $f$ is a function $g : Y \to X$ such that

$$fg = 1_X \text{ and } gf = 1_Y \ .$$

We say $f$ is **invertible** if it has such an inverse.

**Remark**. In the special case that $X = Y$, i.e. domain and target set are the same, so that

$$f, g : X \to X \ ,$$

we have $fg = 1_X = gf$. Thus, in this case, $f$ and $g$ do commute, which is unusual.

18. **Theorem**

(a) If $f : X \to Y$ has an inverse, then that inverse is unique.

(b) $f$ is invertible if and only if it is $1 - 1$ and onto (i.e. bijective = injective + surjective).

**Proof** (part (a) only). Suppose $g : Y \to X$ is an inverse and that $h : Y \to X$ also satisfies the inverse requirements:

$$fh = 1_X \text{ and } hf = 1_Y \ .$$

Then $h = h1_X = h(fg) = (hf)g = 1_Y g = g$. So the supposed inverses $g$ and $h$ are really the same.

The equivalent conditions in part (b) may be familiar from Calculus or other classes; the proof is routine. □

19. We have proved that the inverse of a function is unique. When this happens in mathematics, we are justified in assigning special

**Notation**. The inverse of $f$ is denoted $f^{-1}$.

20. **Example. The inverse of a reflection** $r$. Since $r^2 = rr = 1$, we see at once that

$$r^{-1} = r \ .$$

Thus every reflection equals its own inverse.

Think: to undo a reflection, reflect again.

21. **Exercises**

(a) Other kinds of plane isometries $g : \mathbb{R}^2 \to \mathbb{R}^2$ have the property that $g^2 = 1$. These too are self-inverse: $g^{-1} = g$. Can you characterize all such isometries? (Hint: rotations through a special number of degrees.)

(b) Show that the function

$$\begin{array}{rcl} f : \mathbb{R}^{\geq 0} & \to & \mathbb{R}^{\geq 0} \\ x & \mapsto & x^{1/2} \end{array}$$

is bijective. What is its inverse?

(c) Do the same for

$$\begin{array}{rcl} h:\mathbb{R} & \to & \mathbb{R}^{>0} \\ x & \mapsto & e^x \end{array}$$

22. **Properties of Inverse Functions**. Suppose $f : X \to Y$ and $g : Y \to Z$. Then we have

    (a) The identity $1 = 1_X$ has an inverse; indeed $1^{-1} = 1$.

    (b) If $f$ is invertible, the so is $f^{-1}$; indeed,

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

    (c) If $f$ and $g$ are invertible, then so is their product $fg$; indeed,

    $$(fg)^{-1} = g^{-1}f^{-1}\ .$$

    **Proof**. These are routine calculations. □

    **Think**: put on your socks, then your shoes: how do you reverse that? Are the operations commutative?

23. **Where are we?** We have said quite a lot about products and inverses of functions with generally different domain and range sets. But the algebraic applications are much more harmonious when domain and range coincide ($X = Y$). We turn to that special case next and recapitulate what we have already observed above.

## 2.3 The symmetric group on a set $X$

Any bijection from $X$ onto $Y$ defines a $1-1$ correspondence between the elements of $X$ and those of $Y$. These sets therefore have exactly the same *cardinality*, whether infinite or not. The collection of all such bijections does have interesting algebraic properties. However, for a truly rich theory we do need an identity and this forces $X = Y$.

1. **Definition** Let $X$ be any non-empty set, finite or infinite. Let

    $$\mathbb{S}_X = \{\text{all bijections } f : X \to X\},$$

    equipped with composition $fg$ as operation. We say that $\mathbb{S}_X$ is the *symmetric group* on the set $X$.

    We can study this object for any set $X$. Thus we have lots of groups of a particular kind.

2. The term *group* is used in recognition that $\mathbb{S}_X$ has these natural algebraic properties:

   **closure** The product of two bijections $f$, $g$ is another bijection $fg$.

   **associativity** Always $(fg)h = f(gh)$.

   **identity** The identity $1 = 1_X$ is a bijection on $X$.

   **inverses** Each bijection $f$ has an inverse $f^{-1}$.

   **Proof**. We have verified all these properties in more general circumstances. You should reverify that if $f$ and $g$ are bijections from $X$ to $X$, then (i) so is $fg$; (ii) so is $f^{-1}$; (iii) so is $1 = 1_X$. □.

3. For completeness, we restate some important calculations, whose truth actually follows from the above properties:

   (a) Always, $(f^{-1})^{-1} = f$.

   (b) Always, $(fg)^{-1} = g^{-1}f^{-1}$.

4. **Exercises on Symmetric Groups**

   (a) If $X$ is finite, say with $n$ elements, how many bijections are there in $\mathbb{S}_X$? In other words, what is the order of $\mathbb{S}_X$?

   (Once again think of $n$ people in $n$ chairs rearranging themselves.)

   (b) Could $\mathbb{S}_X$ ever be a commutative group?

   (c) Suppose $X$ is finite, say with cardinality $n$. Let $f : X \to X$ be *any* function not necessarily with any special properties. Show that $f$ is 1–1 if and only if it is onto.

   **Remark**. For finite ground sets $X$, we therefore need only check one of the two conditions required for a function to be bijective.

# 3 Symmetric Groups – Permutation Groups

1. A bijection from a set $X$ to itself is often called a *permutation* on $X$; another synonym is *rearrangement*.

   The actual nature of the elements of the *ground set* $X$ is often immaterial. When $X$ has finite cardinality, say

   $$|X| = n ,$$

   it is convenient to take

   $$X = \{1, \ldots, n\} =: [n] ,$$

   in which case we often write $\mathbb{S}_n$ in place of $\mathbb{S}_X$.

   **Notation**. $\mathbb{S}_n$ is the symmetric group of permutations on $\{1, \ldots, n\}$.

   **Remarks**. Such groups are well suited to computer implementation; there are many efficient algorithms for computing with permutations.

   The shorthand $[n] := \{1, \ldots, n\}$ is frequently used in algebra and combinatorics. Thus a permutation is a bijection

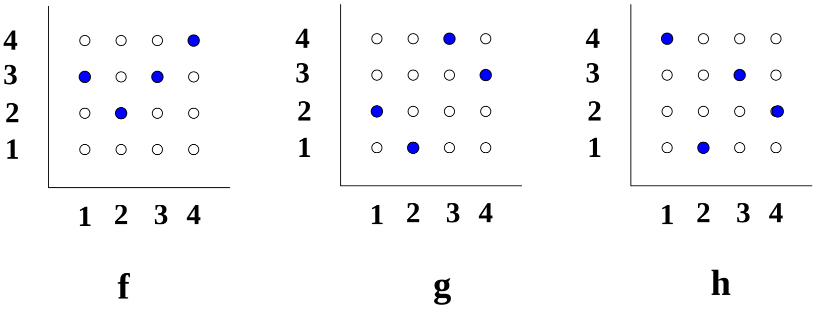
   $$f : [n] \to [n] .$$

2. **Visualizing general functions $f : [n] \to [n]$ (including bijections)**

We can use graphs, arrow digrams or cycle notation. The latter device is efficient and is precisiely how we represent permutations in a computer language like GAP. The case $n = 4$ is typical enough, so we will take

$$X = [4] = \{1, 2, 3, 4\} .$$

(a) Here are some functions represented as conventional graphs in $\mathbb{R}^2$:



Actually, each graph is embedded in the grid

$$[4]^2 = [4] \times [4] = \{1, 2, 3, 4\} \times \{1, 2, 3, 4\} .$$

Note that $f$ is not onto hence also not 1–1. Thus $f \notin \mathbb{S}_4$. However, $g$ and $h$ are typical permutations in $\mathbb{S}_4$.

Since the graph lives in the $4 \times 4$ grid, it makes no sense to join up the dots by line segments! We begin to see that the positions of 1,2,3,4 on the axes are irrelevant. As a matter of fact, the symbols themselves are somewhat arbitrary. We could just as well permute four other symbols, like

$$\alpha, \beta, \gamma, \delta$$

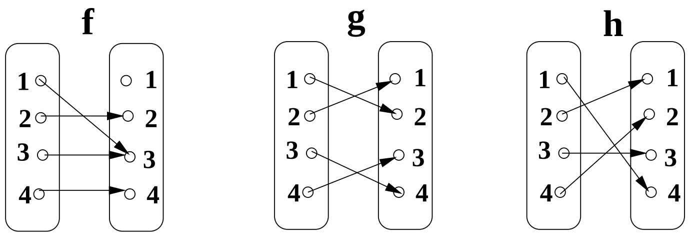
or

$$\heartsuit, \diamondsuit, \clubsuit, \spadesuit ,$$

which have no natural positions in a graph. The mathematical content of the permutations would be unaltered by such a change in the ground set.

However, we stick with 1,2,3,4, since these symbols are familiar and are easy to enter into the computer.

(b) The same three functions are more naturally represented by these arrow diagrams:



Note how easy it is to pick out the 1–1 or onto functions. In fact, we easily see again why 1–1 is equivalent to onto for *finite* ground sets $X$.

(c) Now we investigate cycle notation, which is appropriate only for bijections. Let's look at the function $h$, which could be fully described by the following cumbersome setup:

$$\begin{aligned} 1h &= 4 \\ 2h &= 1 \\ 3h &= 3 \\ 4h &= 2 \end{aligned}$$

The fact that we see a rearrangement of 1,2,3,4 in the right-hand column confirms that $h$ is a bijection. Here is a slightly more compact way of representing the same information:

$$1 \xrightarrow{h} 4 \xrightarrow{h} 2 \xrightarrow{h} 1 \text{ and } 3 \xrightarrow{h} 3 \ .$$

The cycle representation for $H$ is just an abbreviation of this. Here is how it works.

Start with any element of the ground set, say $x = 1$ to be specific, and track where that element is sent by $h$ (here 4), then where that 4 is sent by $h$ (here to 2), and so forth. You will find that your list of elements closes up into a so-called *cycle*. For example, $h$ sends 2 back to the initial element 1. This information can be compressed as follows:

$$(1, 4, 2) \ .$$

Note that we scan a cycle left to right and that each element is mapped by $h$ to the one just after, except that the right-most element is mapped to the front element of the cycle.

Now repeat the process for any unaccounted-for elements in the ground set and manufacture more such cycles.

To finish off our $h$ we note that the remaining element 3 must belong to a trivial cycle (3), indicating that 3 is a *fixed point* for $h$. When done we may express $h$ as a product of *disjoint* cycles, namely

$$h = (1, 4, 2)(3) \ .$$

Note that the same infromation is conveyed if we start any particular cycle at another of its elements. Thus

$$h = (2, 1, 4)(3) \ .$$

If the context is clear, that is to say, if we know we are permuting $\{1, 2, 3, 4\}$, then trivial cycles are often suppressed and understood. In fact, GAP will do just that and print $h$ as

$$(1, 4, 2) \ .$$

Clearly, the above process works for any bijection on a finite set. One encodes the mapping information in a collection of disjoint (i.e. non-overlapping) cycles.

(d) **The identity permutation, inverses, products**

The notation 1 for the identity function would now be a bit confusing; so let us use $e$ to denote the identity permutation on [4]. Thus $e$ fixes every element and so

$$e = (1)(2)(3)(4)\ .$$

After suppressing trivial cycles, GAP would write $e = ()$.

It is easy to write the inverse of a permutation given in cycle form. Since the inverse merely reverses all arrows in an arrow diagram, we must rewrite each cycle in reverse order. Again the starting element for each cycle is a matter of choice. For example,

$$h^{-1} = (1, 2, 4)(3)\ .$$

Trivial cycles, in fact also cycles of length 2, are unaffected by switching to the inverse.

As another example, consider $g$ from above. We can convey the mapping information for $g$ in several ways, so it is quite legal to write

$$g = (1, 2)(3, 4) = (3, 4)(1, 2) = (2, 1)(3, 4)\ .$$

Since every cycle has length 2, $g$ must be self-inverse: $g = g^{-1}$, just like a reflection.

Multiplying permutations in cycle format is is easy – just remember to scan left to right. For example, $g$ maps 1 to 2 and $h$ maps 2 to 1, so $gh$ maps 1 to 1. Now move on to input 2: $g$ maps 2 to 1 and $h$ maps 1 to 4, so $gh$ maps 2 to 4. Next move to input 4 and continue. With practice one can write out products without any effort:

$$gh = (1, 2)(3, 4) \cdot (1, 4, 2)(3) = (1)(2, 4, 3) = (2, 4, 3)$$

and

$$hg = (1, 4, 2)(3) \cdot (1, 2)(3, 4) = (1, 3, 4)(2) = (1, 3, 4)\ .$$

Notice that we end up with disjoint cycles, even though at intermediate steps we might not have disjoint cycles. Also note here that $gh \neq hg$.

(e) **Theorem**. Every permutation $f \in \mathbb{S}_n$ can be factored as a product of disjoint cycles an essentially unique way.

**Proof**. See any text on group theory. The argument just tightens up our informal discussion above. □

(f) Here are the $24 = 4!$ elements of $\mathbb{S}_4$, as produced by GAP:

$$(), (3, 4), (2, 3), (2, 3, 4), (2, 4, 3), (2, 4), (1, 2), (1, 2)(3, 4), (1, 2, 3), (1, 2, 3, 4),$$

$$(1, 2, 4, 3), (1, 2, 4), (1, 3, 2), (1, 3, 4, 2), (1, 3), (1, 3, 4), (1, 3)(2, 4), (1, 3, 2, 4),$$

$$(1, 4, 3, 2), (1, 4, 2), (1, 4, 3), (1, 4), (1, 4, 2, 3), (1, 4)(2, 3)$$

## 3.1 Even and odd permutations

1. An $m$-cycle in $\mathbb{S}_n$ is a permutation which can be written as a single cycle, say

   $$c = (a_1, \ldots, a_m) \ ,$$

   where $a_1, \ldots, a_m$ are distinct elements taken from $\{1, \ldots, n\}$ in any particular order. Of course, this means $m \leq n$; and again we have suppressed fixed points (i.e. 1-cycles).

   A 2-cycle $t = (a, b)$ is often called a *transposition*.

2. We have seen that we can factor a general permutation as a product of disjoint cycles.

   We now observe that any individual cycle can be factored as a product of transpositions, typically in several different ways. These transpositions are unlikely to be disjoint, however. Our proof is by construction:

   $$(a_1, \ldots, a_m) = (a_1, a_2)(a_1, a_3) \cdots (a_1, a_m) \ .$$

   For example,

   - $(1, 3) = (1, 3) = (1, 2)(1, 3)(2, 3)$
   - $(1, 3, 5) = (1, 3)(1, 5)$
   - $(1, 2, 3, 4) = (1, 2)(1, 3)(1, 4)$

   We shall see, however, that if a permutation factors as a product of an odd number of transpositions, then any other factorization of it also involves an odd number of transpositions. Ditto for even numbers of transpositions.

3. Since every permutation can be factored as a product of disjoint cycles, we conclude

   **Proposition 3.1** *Every permutation $f \in \mathbb{S}_n$ can be factored as a product of transpositions, generally in many different ways.*

   For example, in $\mathbb{S}_8$ we have

   $$(1, 3, 5)(2, 4, 6, 8) = (1, 3)(1, 5)(2, 4)(2, 6)(2, 8) \ .$$

   This permutation will soon be called odd; and it does require an odd number of transpositions.

   **Remark**. This factorization essentially says that any shuffle of a deck of cards can be achieved by repeatedly swapping two cards at a time. This is quite believable.

4. **Definition 3.1** *Let $f$ be any permutation in $\mathbb{S}_n$; and suppose when $f$ is written as a product of disjoint cycles that we require $d$ such cycles,* including *all 1-cycles. Then the* sign *of $f$ is*

   $$\operatorname{sgn}(f) = (-1)^{n-d} \ .$$

   *If* $\operatorname{sgn}(f) = +1$*, we say that $f$ is an* even *permutation; if* $\operatorname{sgn}(f) = -1$*, we say that $f$ is* odd.

The Gap lingo is

```
gap> SignPerm(f);
```

5. **Examples**. Remember to count all 1-cycles, which normally we would suppress for ease of reading.

   - the identity $e = () = (1)(2)\dots(n)$ has $d = n$ one-cycles hence
   $$\mathrm{sgn}(e) = (-1)^{n-n} = +1 \ .$$
   - A transpostion $t = (a, b)$ has a single 2-cycle and $n - 2$ one-cycles, so $d = 1 + (n - 2) = n - 1$, so
   $$\mathrm{sgn}(t) = (-1)^{n-(n-1)} = -1 \ .$$
   - Since the $m$-cycle $c = (a_1, \dots, a_m) = (a_1, a_2)(a_1, a_3) \cdots (a_1, a_m)$, the sign of an $m$-cycle $c$ is
   $$\mathrm{sgn}(c) = (-1)^{n-(1+n-m)} = (-1)^{m-1} \ .$$
   Thus, a little confusingly, a 5-cycle is even, whereas a 6-cycle is odd.

The crucial result is a neat calculation:

**Proposition 3.2** *For any permutation $f$ and transposition $t = (a, b)$ in $\mathbb{S}_n$,*
$$\mathrm{sgn}(ft) = -\mathrm{sgn}(f) = \mathrm{sgn}(f)\mathrm{sgn}(t) \ .$$

**Proof**. Write $f$ as a product of disjoint cycles, taking care to include all 1-cycles:
$$f = (\dots) \cdots (\dots)(\dots) \cdots (\dots) \ .$$

The distinct elements $a$ and $b$ must appear somewhere and exactly once, either in a common cycle or in different cycles.

**Case 1**. A common cycle. We convey the same mapping information by moving $a$ to the front of the cycle. The rest of the cycle must look something like
$$(a, x_1, \dots, x_l, b, y_1, \dots, y_k) \ .$$

(Possibly $l = 0$ so there aren't any $x$'s, etc.; this won't hurt our argument.) Anyway, this cycle for $f$ is disjoint from all other cycles, so we can commute it to the end to get
$$f = (\dots) \cdots (\dots)(\dots) \cdots (a, x_1, \dots, x_l, b, y_1, \dots, y_k) \ ,$$

whence
$$\begin{aligned} ft &= (\dots) \cdots (\dots)(\dots) \cdots (a, x_1, \dots, x_l, b, y_1, \dots, y_k) \cdot (a, b) \\ &= (\dots) \cdots (\dots)(\dots) \cdots (a, x_1, \dots, x_l)(b, y_1, \dots, y_k) \ , \end{aligned}$$

thereby introducing exactly one more cycle! The number of disjoint cycles goes up by 1; since this appears in the exponent in the definition of the sign, the sign must change by the factor -1.

**Case 2**. This is similar and basically reverses the above calculation. Now the number of disjoint cycles goes down by 1; but again this multiplies the sign by $-1$. □

6. **Theorem 3.1** *The function*

$$\begin{array}{rcl} \mathrm{sgn} : \mathbb{S}_n & \to & \{\pm 1\} \\ f & \mapsto & \mathrm{sgn}(f) \end{array}$$

*is a homomorphism from the group* $\mathbb{S}_n$ *(with permutation composition) to the group* $\{\pm 1\}$ *of order* 2 *(now multiplication of integers). In other words, we have*

$$\mathrm{sgn}(fg) = \mathrm{sgn}(f)\mathrm{sgn}(g)$$

*for all* $f, g \in \mathbb{S}_n$.

**Proof**. Any $f$ and $g$ can be written as a product of transpostions, say $f = t_1 t_2 \cdots t_k$ and $g = \tilde{t}_1 \cdots \tilde{t}_l$. By repeatedly applying Proposition 3.2, we get

$$\begin{array}{rcl} \mathrm{sgn}(f) & = & \mathrm{sgn}([t_1 \cdots t_{k-1}]t_k) \\ & = & (-1)\mathrm{sgn}(t_1 \cdots t_{k-1}) \\ & = & (-1)^2\mathrm{sgn}(t_1 \cdots t_{k-2}) \\ & = & (-1)^k\mathrm{sgn}(e) \\ & = & (-1)^k. \end{array}$$

Similarly, $\mathrm{sgn}(g) = (-1)^l$, so that

$$\mathrm{sgn}(fg) = \mathrm{sgn}(t_1 t_2 \cdots t_k \tilde{t}_1 \cdots \tilde{t}_l) = (-1)^{k+l} = (-1)^k(-1)^l = \mathrm{sgn}(f)\mathrm{sgn}(g)\ .$$

□

7. **Corollary 3.1** *Parity of permutations has a new, sensible meaning. An even permutation can be factored only as an even number of transpositions; An odd permutation can be factored only as an odd number of transpositions.*

**Proof**. Note that $(-1)^k = +1$ forces $k$ to be even, never odd. □

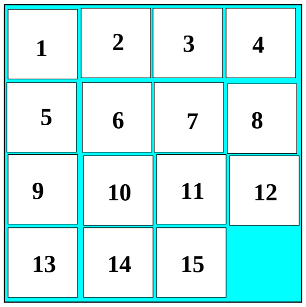
8. **Application 1 - determinants**. Suppose $A = [a_{ij}]$ is an $n \times n$ matrix. As we noted in class, one way to define the determinant is

$$\det(A) := \sum_{f \in \mathbb{S}_n} \mathrm{sgn}(f)\ a_{1,(1)f}a_{2,(2)f} \cdots a_{n,(n)f} \qquad (1)$$

(a sum of $n$! signed terms, each a product of $n$ specially selected entries).

9. **Application 2 - the '15-puzzle'**. This familiar puzzle has 15 sliding blocks labelled $1, \ldots, 15$ and located in a $4 \times 4$ frame. The 16th square is empty or blank, so we label it $b$. By sliding the blank around, we can reconfigure the blocks in various ways.

Here is the starting configuration:



| | | | |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | |

The key problem is this: can we slide the blocks so as to arrive at

| | | | |
|---|---|---|---|
| 2 | 1 | 3 | 4 |
| 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | |

**Solution**. No, we cannot. Each unit move of the blank space (either horizontal or vertical) leaves most of blocks fixed and really amounts to applying a transposition of the form $(i, b)$, where $i \in \{1, 2, \ldots, 15\}$. Any succession of $m$ such moves amounts to a product of $m$ transpositions, whose sign is $(-1)^m$ (Theorem 3.1). However, the blank space is still at the lower right corner, which means that we have moved the blank an even number of times (both horizontally and vertically). In any case, $m$ must be even. Thus we have effected an even permutation on the set $\{1, 2, \ldots, 15, b\}$. The configuration in the second figure amounts to the transposition $(1, 2)$, which is odd.□

10. **Application 3 - the Alternating Group**.

(a) **Definition 3.2** *The* alternating group of degree $n$ *is the set* $\mathbb{A}_n$ *of all even permutations in* $\mathbb{S}_n$*, still with left to right composition.*

(b) For example, $\mathbb{A}_3$ is isomorphic to the group of rotational symmetries of an equilateral triangle (order 3, rotations through $0°, +120°, -120°$).

(c) Likewise, $\mathbb{A}_4$ faithfully represents the group of rotational symmetries of a regular tetrahedron in ordinary space. There are 12 rotations:

- the identity rotation (through $0°$), corresponding to the identity permutation $()$.
- four $120°$ rotations (looking from outside onto a triangular face), corresponding to say $(1, 2, 3), (1, 4, 2), (1, 3, 4), (2, 4, 3)$.
- four $-120°$ rotations (still looking from outside onto a triangular face), corresponding to say $(1, 3, 2), (1, 2, 4), (1, 4, 3), (2, 3, 4)$.
- three $180°$ rotations, corresponding to $(1, 2)(3, 4)$, $(1, 3)(2, 4)$ and $(1, 4)(2, 3)$.

We have twelve symmetries that we can achieve by physically manipulating the tetrahedron; and we have twelve even permutations.

In fact, we have listed the four conjugacy classes of $\mathbb{A}_4$. On the geometrical side, the motions in each class have the 'same geometric effect' on the tetrahedron. In particular, $120°$ rotations and $-120°$ are different: we cannot pass from one to the other without the use of a reflection taking right to left hand. But such reflections must correspond to odd permutations of the four vertices.

In parallel to that, the permutations in each conjugacy class 'look like one another'. In addition, we have seen that we do further have to distinguish say $(1, 2, 3)$ from its inverse $(1, 3, 2)$ when the setting is $\mathbb{A}_4$. However, when the setting is enlarged to $\mathbb{S}_4$ these two classes fuse into one.

(d) Let's see how this plays out in Gap:

```
gap> S4:=SymmetricGroup(4);;
gap> A4:=AlternatingGroup(4);Size(A4);
Alt( [ 1 .. 4 ] )
12
gap> IsSubgroup(S4,A4);
true
gap> conA:=ConjugacyClasses(A4);;Size(conA);
4
gap> for j in [1..4] do Print(j," ",Elements(conA[j]),"\n");od;
1 [ () ]
2 [ (1,2)(3,4), (1,3)(2,4), (1,4)(2,3) ]
3 [ (2,4,3), (1,2,3), (1,3,4), (1,4,2) ]
4 [ (2,3,4), (1,2,4), (1,3,2), (1,4,3) ]
gap> conS:=ConjugacyClasses(S4);;Size(conS);
5
gap> for j in [1..5] do Print(j," ",Elements(conS[j]),"\n");od;
1 [ () ]
2 [ (3,4), (2,3), (2,4), (1,2), (1,3), (1,4) ]
3 [ (1,2)(3,4), (1,3)(2,4), (1,4)(2,3) ]
4 [ (2,3,4), (2,4,3), (1,2,3), (1,2,4), (1,3,2), (1,3,4), (1,4,2),
  (1,4,3) ]
5 [ (1,2,3,4), (1,2,4,3), (1,3,4,2), (1,3,2,4), (1,4,3,2), (1,4,2,3) ]
```

Thus the $4! = 24$ elements of $\mathbb{S}_4$ lie in 5 classes. For example, the second class consists of the six reflection symmetries. The fourth class consists of the rotations of period 3 - now they are all alike since the presence of reflections lets the group waive the distinction between clockwise and anticlockwise. The second class is unchanged (and a clockwise 180° rotation is indeed identical to an anticlockwise 180° rotation).

This leaves the last class, consisting of all six 4-cycles in the symmetric group. (Recall that a 4-cycle is an odd permutation.) These must correspond to a symmetry for the tetrahedron which (like reflections) reverses orientation. In other words, such symmetries would send a left hand sketched on the surface of the solid to a right hand.

So look at a 4-cycle like $(1, 2, 3, 4)$. What does this mean geometrically? It will help to hold a model in your hand, with an edge horizontal and the opposite edge also horizontal but 'perpendicular' to the first. Imagine the vertical *axis* passing through the midpoints of these two opposite edges. The symmetry that we are after here is a composite thing obtained by composing a 90° turn about the vertical axis with a subsequent reflection in the (horizontal) plane perpendicular to the axis and through the centre of the tetrahedron. (Caution: this last reflection on its own is not a symmetry, nor is the 90° turn!) If you track the effect of this symmetry, you will see that it cyclically sends the 4 vertices of the tetrahedron along a ziz-zag path through the edges. (Such a path is called a *Petrie polygon* for the tetrahedron.) Thus we recreate the 4-cycle in a geometrical way.

This sort of combined reflection-rotation is called a *rotatory reflection*. It certainly reverses sense, since the reflection part does but the rotation part does not. Yet it is different from an ordinary relection. An ordinary reflection fixes all points on its planar mirror; a rotatory reflection fixes only one point, in this case the centre of the tetrahedron. This qualitative geometrical distinction translates into distinct conjugacy classes in the group $\mathbb{S}_4$.

11. Our geometrical analysis extends to the regular simplex in $n-1$ dimensions, whose full symmetry group is isomorphic to $\mathbb{S}_n$ and whose rotation subgroup is isomorphic to $\mathbb{A}_n$. Let's consolidate some general results.

**Theorem 3.2** *$\mathbb{A}_n$ is a subgroup of $\mathbb{S}_n$. For $n \geq 2$ its order is $\frac{n!}{2}$.*

**Proof**. Of course the identity $e$ is even, so $e \in \mathbb{A}_n$. If $f, g \in \mathbb{A}_n$, then so is $fg$ by Theorem 3.1. Similarly $f^{-1} \in \mathbb{A}_n$. Thus $\mathbb{A}_n$ is a group (associativity is inherited from $\mathbb{S}_n$).

Suppose $n \geq 2$, and let $\mathbb{O}_n$ be the set of odd permutations in $\mathbb{S}_n$. Clearly $\mathbb{O}_n$ is non-empty; for example, $t = (1,2) \in \mathbb{O}_n$. however the odd permutations definitely do not form a group, since $e \notin \mathbb{O}_n$. We define a function

$$\begin{aligned} \varphi : \mathbb{A}_n &\rightarrow \mathbb{O}_n \\ f &\mapsto ft \end{aligned}$$

This function is well-defined (again by Theorem 3.1). It is 1–1 and onto since we are working in a group. Thus $|\mathbb{A}_n| = |\mathbb{O}_n| = \frac{n!}{2}$, since the two sets here have the same size and exhaust all of $\mathbb{S}_n$ without any overlap. □

**Remark**. We thus have $\mathbb{A}_n = \mathbb{A}_n e$ and $\mathbb{O}_n = \mathbb{A}_n t$. Both sets are said to be *right cosets* of the subgroup $\mathbb{A}_n$ in $\mathbb{S}_n$.

12. **A look ahead to conjugacy- in symmetry groups and elsewhere**.

(a) Suppose $\mathcal{F}$ is some geometrical object, maybe a polygon (such as a square) in the plane, or maybe a polyhedron (such as a tetrahedron) in space, or any other such object. $\mathcal{F}$ need not be particularly symmetric; but, in any case, $\mathcal{F}$ does have a symmetry group $G$. Even if $\mathcal{F}$ is 'totally unsymmetric', we still have $G = \{1\}$, the trivial group of order one.

Now consider two symmetries $f$ and $g$ in $G$. Sometimes these 'look alike', as with two reflection symmetries for the euilateral triangle. Sometimes they are unalike, as with a reflection and rotation. How do we quantify this?

(b) Intuitively, symmetries $f$ and $g$ 'look alike' – are conjugate – if they not only have the same geometric properties but are also 'situated' the same way relative to the object $\mathcal{F}$.

For example, two +120° rotations for the regular tetrahedron are clearly conjugate, since we can turn from one to the other *via a rotational symmetry of the tetrahedron itself.*

On the other hand, a reflection in a diagonal of the square (think $r_2 = (1,3)$) cannot be conjugate to a reflection in a line perpendicular to the edges (think

$r_1 = (1,2)(3,4)$). After all, the first reflection fixes two vertices, the second none. From the point of view of the square, $r_1$ and $r_2$ are different, although both are reflections in the plane.

Switching to Permutationland, the obvious distinction is that $r_1$ and $r_2$ have different cycle structures ($2+2$ versus $2+1+1$).

(c) Another example will will motivate our definition. A $+120°$ rotation $f$ for the tetrahedron is not conjugate to a $-120°$ rotation $g$ so long as we stay inside the rotations only.

But by moving to include reflections the distinction is blurred. Suppose $r$ is some reflection symmetry for the tetrahedron. We observe that $r$ swaps clockwise with anticlockwise. How can we exploit that? Look at the triangular face where $f$ (anticlockwise) is acting; but put $f$ aside for a moment. First reflect that face by $r$ to some other face; in that new face apply say $g$ to turn the face (clockwise); then reflect back. We have done exactly what we need to effect $f$, so that $f = rgr$. Note that $r = r^{-1}$ here, so we really have $f = r^{-1}gr$.

(d) Let's generalize. Fix a particular symmetry $f$. For another symmetry $g$ to be 'the same', we really mean that there is a third symmetry $x$ *still in the same group* which moves $g$ to $f$. This really means that we can achieve the effect of $f$ by first moving from $f$ back to $g$, via $x^{-1}$, then performing $g$, then moving back to $f$ via $x$. In other word,

$$f = x^{-1}gx \ .$$

If we can find such an $x$, then $f$ and $g$ will look the same. And it won't hurt if there are several options $x$ for the same pair $f$ and $g$.

But if we cannot find such an $x$ in the group, then $f$ and $g$ won't look the same.

(e) This way of thinking transfers to any group, be it a symmetry group, a permutation group, or any other group, abelian or non-abelian.

**Definition 3.3** *Let $G$ be any group (with operation written multiplicatively). Then two elements $f, g \in G$ are* conjugate *if for some $x \in G$ we have*

$$f = x^{-1}gx \ .$$

*We write $f \sim g$ to indicate this relation.*

*The set of all relatives to $g$ under this relation is called the* conjugacy class *of $g$, written*

$$Cl(g) := \{f \in G : f \sim g\} = \{x^{-1}gx : x \in G\} \ .$$

(f) **Remarks**. In other words, as $x$ runs through the group, we get various $x^{-1}gx$ conjugate to $g$, quite possibly with repeated values for different $x$'s. These constitute the conjugacy class for $g$.

In addiitve notation, multiplication $\cdot$ becomes $+$, $f^{-1}$ becomes $-f$, and conjugacy becomes

$$f = (-x) + g + x \ .$$

Such additive notation is typically used only for abelian groups. What happens to conjugacy then?

(g) In the abelian group $\mathbb{Q}^*$ with multiplication, $f \sim g$ means

$$\begin{aligned} f &= x^{-1}gx \quad \text{(for some } x \neq 0) \\ &= x^{-1}(xg) \quad \text{(since mult. is commutative here)} \\ &= g\ . \end{aligned}$$

Conclusion: if two rational numbers are conjugate, they must in fact be equal. Every conjugacy class contains one element. In this peculiar sense, no rational is 'like' any other.

The same holds in any abelian group.

(h) **Theorem 3.3** *Conjugacy is an equivalence relation. That is, for all $f, g, h \in G$,*

*(a) $f \sim f$*

*(b) $f \sim g$ implies $g \sim f$.*

*(c) $f \sim g$ and $g \sim h$ together imply $f \sim h$.*

This readily checked. The upshot is that the conjugacy classes partion the group $G$; in other words, the classes are non-empty, mutually disjoint sets which together comprise the whole group. Each class consists of all elements which 'look like each other', from the perspective of the group itself.

# 4 Groups

A *group* is a set $G$ equipped with a binary operation satisfying a few key properties. The operation is typically written $+, \times, *, \circ$, etc.

## 4.1 Binary Operations

The idea of a *binary operation* on a set $G$ is that given $a, b \in G$ (allowing $a = b$) we can produce a new element

$$a * b \text{ (also in } G).$$

**Remarks**.

1. Think: apple * apple = apple; do you know any non-binary operations?

2. We emphasize that $a * b$ is *uniquely* given once $a, b$ are known.

3. We typically must define $a * b$. So it is always crucial to check that arbitrary choices, if any, in calculations do not actually affect the final outcome $a * b$. In short, our description of $a * b$ must be *well-defined*.

4. Thus $*$ is really a function

$$* : G \times G \to G$$

5. Example.

$$+ : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}.$$

We have $+(5,3) = 8$, $+(3,-7) = -4$ and $+(-13,0) = -13$. Of course, for brevity we usually write $a+b$ instead of $+(a,b)$.

## 4.2 Some Desirable Properties for the Operation (in a Group)

Let's return to a general group $G$; write $a * b$ for $a, b \in G$. Based on our experience with familiar examples, we want $a * b$ to have the following natural properties:

1. $(a * b) * c = a * (b * c)$ , $\quad \forall a, b, c \in G$. (The operation $*$ is *associative.*)
2. There exists in $G$ some special element $e$ such that
$$e * a = a \qquad \text{for all } a \in G\ .$$
3. For each $a \in G$, there exists an element $b$ (depending on $a$) such that
$$b * a = e\ ,$$
where $e$ is an element mentioned in requirement (2).

**Remarks**:

1. Properties (1), (2), (3) for a binary operation $*$ on a set $G$ actually define a group.
2. We do *not* assume $a * b = b * a$ always holds, though it may occasionally do so. For example, $a * a = a * a$ for all $a \in G$.

   In a *commutative* (or *abelian*) group $G$ we do have
$$a * b = b * a \qquad \forall a, b \in G\ .$$

## 4.3 Some Little but Important Theorems for all Groups

Let $a, b... \in G$ be typical group elements. Let $e$ be an element as defined in requirement (2) above.

1. **Theorem**. If for some $a \in G$ we have $a * a = a$, then $a = e$.
**Proof**. By property (2) there exists a $b$ such that $b * a = e$. Thus
$$\begin{aligned} e &= b * a \\ &= b * (a * a) \\ &= (b * a) * a \\ &= e * a \\ &= a\ . \end{aligned}$$

□

2. **Theorem** Suppose $\tilde{e} \in G$ also satisfies $\tilde{e} * a = a$ for all $a \in G$. Then $\tilde{e} = e$.

   **Proof**. $\tilde{e} * \tilde{e} = \tilde{e}$, so $\tilde{e} = e$ by the preceding theorem. □

   **Meaning**. There is a *unique* element $e \in G$ such that

   $$e * a = a \qquad \forall a \in G \; .$$

   **Definition**.: The unique element $e$ is called the *identity* in $G$.

3. **Theorem**. If $b * a = e$, then $a * b = e$.

   **Remark**. Thus some commuting must happen.

   **Proof**.

   $$\begin{aligned} (a * b) * (a * b) &= a * [b * (a * b)] \\ &= a * [(b * a) * b] \\ &= a * [e * b] \\ &= a * b \end{aligned}$$

   So, by the first theorem, $a * b = e$. □

4. The assumption in defining property (2) for groups is that $e * a = a, \qquad \forall a \in G$. Compare that with the following

   **Theorem** : $\qquad a * e = a \qquad \forall a \in G$.

   **Remark**: again this shows that a little more commuting is forced.

   **Proof**. By property (3), there exists an element $b \in G$ such that $b * a = e$. Thus

   $$\begin{aligned} a * e &= a * (b * a) \\ &= (a * b) * a \\ &= e * a \\ &= a \; . \end{aligned}$$

   □

   **Remark**: The identity $e$ thus commutes with all elements of $G$.

5. **Theorem** Given any $a \in G$, there exists *exactly one* element $b$ such that $b * a = e$.

   **Proof**: Suppose that $b * a = e$ <u>and</u> $c * a = e$. From above we therefore have $a * b = e$ and $a * c = e$. So:

   $$\begin{aligned} c * (a * b) &= c * e \\ (c * a) * b &= c \\ e * b &= c \\ b &= c \end{aligned}$$

□

**Definition**: The unique element guaranteed for each $a$ by this theorem is called is the *inverse* of $a$, and is denoted $a^{-1}$. Thus we have already proved that

$$a * a^{-1} = a^{-1} * a = e \ .$$

6. **Theorem-Exercise** Guess and prove that $\forall a, b \in G$,

$$(a * b)^{-1} = \underline{\hspace{2cm}} \ .$$

Hint: if your guess for the inverse works, it must be <u>the</u> unique inverse.

## 4.4 Examples.

For any particular set $X$, we have seen that $G = \mathbb{S}_X$ is a group; the operation is functional composition $fg$; usually this operation is not commutative.

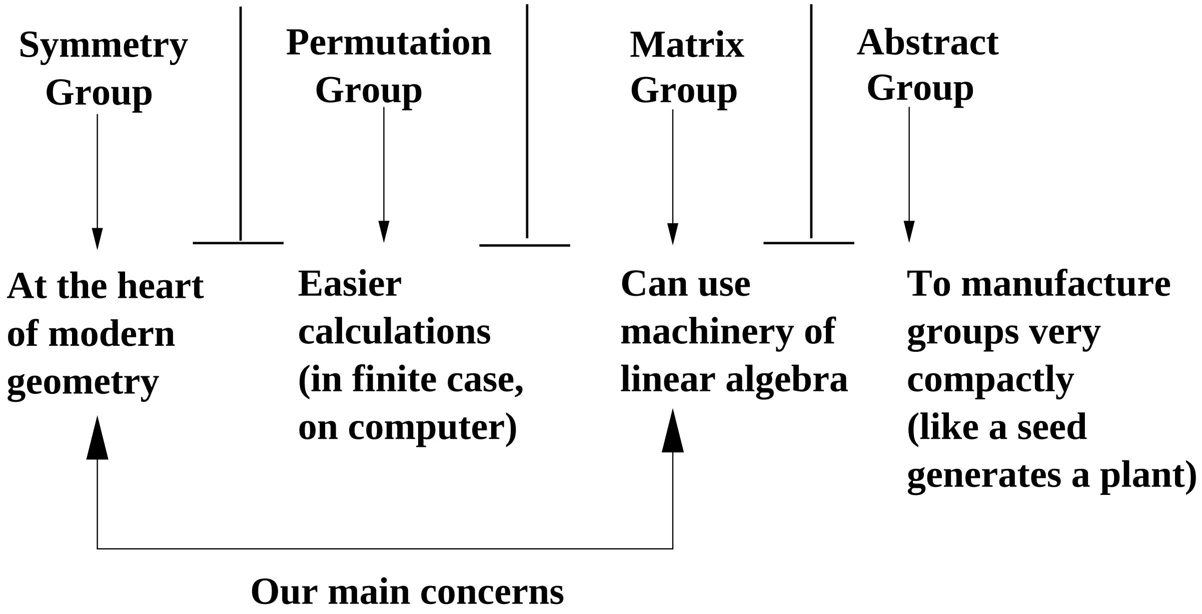
**Exercise**. There are many other more familiar examples of groups. In each case below, indicate whether the given set and operation do yield a group. If not, indicate which of properties (1), (2) or (3) fails. When you do obtain a group, clearly point out the identity and describe the inverse of each element $a$.

| the set $G$ | the operation $*$ |
|---|---|
| $\mathbb{Z}$ | $+$ |
| $\mathbb{Z}$ | $-$ |
| $\mathbb{Z}$ | $\cdot$ |
| $\mathbb{Q}$ | $+$ |
| $\mathbb{R}$ | $+$ |
| $\mathbb{R}$ | $-$ |
| $\mathbb{C}$ | $-$ |
| $\mathbb{Q}^*$ | $\cdot$ |
| $\mathbb{M}_2(\mathbb{R})$ | $+$ |
| $\mathbb{M}_2(\mathbb{R})$ | $\times$ |
| $\{\pm 1\}$ | $\cdot$ |
| $\{1, e^{\frac{2\pi i}{3}}, e^{\frac{4\pi i}{3}}\}$ | $\cdot$ |
| $C_n = \{e^{\frac{(2\pi i)k}{n}}, 0 \le k \le n-1\}$ | $\cdot$ |

Note that $\mathbb{Q}^*$ denotes the non-zero rational numbers. And $\mathbb{M}_2(\mathbb{R})$ is the collection of all $2 \times 2$ real matrices.

## 4.5 Isomorphic Approaches to the Same Group

It is often possible, and fruitful, to look at one and the same group from several points of view:


Symmetry Group
Permutation Group
Matrix Group
Abstract Group
At the heart of modern geometry
Easier calculations (in finite case, on computer)
Can use machinery of linear algebra
To manufacture groups very compactly (like a seed generates a plant)
Our main concerns

In geometry, the idea of *symmetry* is crucial. The parallel idea in algebra is the *matrix group*.

# 5 Subsets and subgroups

Here $G$ is any group, finite or infinite, with the operation written as multiplication. So $ab$ could mean $a \times b$, $a + b$, etc.

## 5.1 Products of Sets

1. Suppose $A, B$ are non-empty subsets of $G$. Then we define

$$AB := \{ab : \ a \in A \ \text{ and } \ b \in B\}.$$

   In short, take all possible products, first an element of $A$, then an element of $B$.

2. $A$ or $B$ could have one element $g \in G$. Then we usually write

$$\begin{array}{l} Ag \text{ (instead of } A\{g\}) \\ gB \text{ (instead of } \{g\}B). \end{array}$$

3. Similarly, we define

$$A^{-1} := \{a^{-1} : \ a \in A\}.$$

4. **Exercises**. Suppose $A, B, C$ are subsets of a group $G$.

   (a) Show $(AB)C = A(BC)$ and $(A^{-1})^{-1} = A$.

   (b) Show that $A, \;\; Ag, \;\; gA$ have the same size.

   (c) Give an example of subsets $A, B$ of $\mathbb{S}_3$ with $|A| = 3, \;\; |B| = 2$ but $|AB| \neq 6$.

   (d) Rewrite the objects in exercises (a) and (b) in *additive notation* .

   (e) Show that $GG = G$.

## 5.2 Subgroups

1. A subset $H$ of $G$ is a *subgroup* if it is a group in its own right, with the operation inherited from $G$.

   Note that associativity is inherited for any subset. Thus for $H$ to be a subgroup we really mean:

   - $1 \in H$.
   - $a, b \in H \Longrightarrow ab \in H$
     (i.e. elements in $H$ also have their product in $H$, in short the group operation is closed on $H$).
   - $b \in H \Longrightarrow b^{-1} \in H$
     (i.e. elements in $H$ have inverses also in $H$ – inverting is also closed on $H$).

2. **Proposition 5.1** (***Subgroup Test***). *A non-empty subset $H$ of $G$ is a subgroup if and only if*

$$a, b \in H \Longrightarrow ab^{-1} \in H.$$

3. **Exercises**.

   (a) $H$ is a subgroup if and only if $HH^{-1} \subseteq H$.

   (b) If $H$ is a subgroup, then $H = H^{-1}$ and $HH = H$.

4. **Example**: Let

$$\begin{aligned} G &= \{2^k : \; k \in Z\} \\ &= \{\ldots \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4 \ldots\} \end{aligned}$$

   with ordinary multiplication.

   **Exercises**.

   (a) Show that $G$ is a group. Thus $G$ is a subgroup of the positive reals, with ordinary multiplication.

   (b) Convince yourself that $G$ is isomorphic to $(Z, +)$. In particular, $G$ is abelian.

(c) Find a subset $H \subset G$ such that $HH = H$ but $H$ is <u>not</u> a subgroup.

(d) Suppose $H$ is a finite subset of any group $G$, and

$$HH = H.$$

Show that $H$ <u>is</u> a subgroup.

5. **Definition** If $a \in G$, then the *cyclic* subgroup generated by $a$ is

$$\langle a \rangle := \{a^k : \ k \in Z\} \ .$$

(Compare $\{2^k : k \in \mathbb{Z}\}$ from above).

6. **More Exercises**. Let $a$ be an element of a group $G$.

(a) Show that $\langle a \rangle$ actually is a subgroup and is, furthermore, abelian.

(b) Give examples of $G$ and $a \in G$ in which $\langle a \rangle$ is infinite. Likewise finite.

7. **Definition**. The *order* of any $a \in G$ can be defined as the smallest positive integer $n$ (if any exists) for which $a^n = 1$. We write

$$|a| = n.$$

If no such $n$ exists, we say $a$ has infinite order: $|a| = \infty$.

**Exercise**. Show that $|a| = |\langle a \rangle|$. (The order of $a$ also equals the number of elements in the cyclic subgroup generated by $a$).

## 5.3 Cosets

When a subset $H$ of $G$ is actually a subgroup, the sets $Hg$ are particularly nicely behaved. For any $g \in G$, we say:

$Hg$ is a *right coset* of $H$
$gH$ is a *left coset* of $H$.

Right and left cosets have analogous properties. So for now let's look only at right cosets.

1. We have already seen that, for any $g \in G$,

$$|gH| = |H| = |Hg|.$$

That is, all cosets have the same size, namely the size of $H$.

Indeed, any subgroup $H$ is itself a coset (of itself). For if 1 is the identity of $G$ we have

$$H = H1.$$

2. **Exercise**. Let $\mathbb{S}_3$ be the group of all permutations on $X = \{1, 2, 3\}$. In other terminology, we say that $\mathbb{S}_3$ is the symmetric group of degree 3. Its order is $3! = 6$, and its elements are

$$S_3 = \{(1), (123), (132), (12), (13), (23)\}.$$

Let $H = \langle(12)\rangle = \{(1), (12)\}$ (a cyclic subgroup of order 2).

(a) Find all right cosets of $H$.

(b) Find a set of right coset representatives (namely, pick an individual element from each coset).

(c) Find all left cosets of $H$.

(d) Are the left and right cosets identical?

(e) Make GAP find all right cosets of $H$. Use GAP to find a set of right coset representatives.

3. **Proposition 5.2** *(a) $Ha = Hb$ if and only if $ab^{-1} \in H$.*
*(b) $aH = bH$ if and only if $b^{-1}a \in H$.*

**Proof**. Part (ii) is similar to part (i). In part (i), there are two things to show.

$$\begin{array}{rll}
\underline{\text{Assume}} & Ha = Hb. & \text{Since } 1 \in H \\
 & 1a = hb, & \text{for some } h \in H. \\
\text{Thus} & ab^{-1} = h \in H. & \\
\underline{\text{Assume}} & ab^{-1} \in H. \text{ We must show } Ha = Hb. & \\
\text{Suppose} & x \in Ha, \text{ say } x = ha. \text{ Then} & \\
 & xb^{-1} = h(ab^{-1}) = hh', \text{ where } h' \in H. & \\
\text{So} & xb^{-1} = \tilde{h} \in H & \\
\text{so} & x = \tilde{h}b \in Hb. & \\
\text{Thus} & Ha \subseteq Hb. & \\
\text{Similarly} & Hb \subseteq Ha\text{: we're done.} &
\end{array}$$

4. **Coset Representatives**

If $Ha$ is any coset, then "a" is called a coset representative for the coset $Ha$. There can be many coset representatives, since $Ha = Hb$ if $ab^{-1} \in H$. Thus, if "a" is one representative, then any $ha = b$ is another ($h \in H$).

In particular, $H = Hb$ whenever $b \in H$.

5. **Proposition 5.3**

Any two right cosets are either identical or disjoint. (The same is true for two left cosets.)

**Proof**. Let $Ha$ and $Hb$ be two cosets of the subgroup $H$.

If they are disjoint (meaning no elements in common) we are done:

$$Ha \cap Hb = \emptyset.$$

So suppose $Ha$ and $Hb$ have at least one element in common, say

$$x \in Ha \cap Hb.$$

The point is that this forces the cosets to be completely the same. Indeed,

$$\begin{aligned} x \in Ha, &\text{ so } x = h_1 a \text{ where } h_1 \in H \\ \text{and } x \in Hb, &\text{ so } x = h_2 b \text{ where } h_2 \in H. \end{aligned}$$

So
$$a = h_1^{-1}x, \quad b = h_2^{-1}x, \quad b^{-1} = x^{-1}h_2,$$

and thus:

$$\begin{aligned} ab^{-1} &= (h_1^{-1}x)(x^{-1}h_2) \\ &= h_1^{-1}h_2 \in H. \end{aligned}$$

By Proposition 5.2 above, $Ha = Hb$. □

6. **Proposition 5.4**

   Every element of $G$ belongs to exactly one coset of $H$.

   **Proof**. Say $g \in G$. Then $g = 1g$, so $g \in Hg$. By Proposition 5.3, $g$ cannot belong to two *different* cosets. □

7. Suppose now that there are finitely many cosets of the subgroup $H$ in $G$. This happens when $G$ itself is finite, but also in other cases.

   We may represent the $k$ cosets by

$$1 = a_0, a_1, a_2, \dots, a_{k-1},$$

so that the different cosets are $H = H1, \;\; Ha_1, \;\; Ha_2, \;\; \dots, \;\; Ha_{k-1}$.

By the previous item, we can represent the situation diagrammatically like this:

| | | | | | |
|---|---|---|---|---|---|
| $G$ | $H$<br>$= H1$ | $Ha_1$ | $Ha_2$ | $\dots$ | $Ha_{k-1}$ |

But all cosets $Ha_j$ have the same size, namely $|H|$.

So

$$\begin{aligned} |G| &= |H| + |H| + \ldots + |H| \\ &= k|H|. \end{aligned}$$

We have therefore proved an unexpected and important result:

**Theorem 5.1 (Lagrange)**

If $H$ is a subgroup of a finite group $G$, then $|H|$ divides $|G|$, and the *index*

$$[G:H] := \frac{|G|}{|H|}$$

is the number of right cosets of $H$. (Also, it is the number of left cosets.)

8. Recall that any element $a \in G$ generates a cyclic group $\langle a \rangle$. We defined the order of $a$ to be the smallest positive integer $n$ such that

$$a^n = 1.$$

(And you should prove that this order equals the size of the corresponding cyclic subgroup.)

**Exercises**

(a) Suppose that $G$ is finite and $a \in G$. Then $|a|$ divides $|G|$.

(b) Suppose that $|G| = p$, a prime. Then $G$ is a cyclic group.
**Conclusion**. Up to isomorphism, there is only one group of prime order $p$. It is cyclic, hence abelian.

(c) Let $G = \mathbb{R}^2$, with vector addition. So

$$[x_1, y_1] + [x_2, y_2] = [x_1 + x_2,\ y_1 + y_2].$$

i. Show that $G$ is an abelian group. What is the (unique) identity element? Of course, $G$ is very infinite.

ii. Let $H = \{[x, 2x] :\ x \in \mathbb{R}\}$. Show that $H$ is a subgroup of $G$. Describe $H$ and its right cosets geometrically. In light of this interpret the above theorems.

(d) Suppose $G$ is a finite group and $H, K$ are subgroups with $K \subseteq H \subseteq G$.
Show: $[G:K] = [G:H] \cdot [H:K]$.

Note:

$$\text{index is } [G:K] \left\{ \begin{array}{cl} G & \\ | & \text{index is the integer}[G:H] \\ H & \\ | & \text{index is the integer}[H:K] \\ K & \end{array} \right.$$

(e) Let $G$ be any group, perhaps infinite.

i. Suppose $H_1$ and $H_2$ are subgroups. Show that $H_1 \cap H_2$ is also a subgroup.

ii. More generally, suppose $\{H_t : t \in \mathcal{I}\}$ is any collection of subgroups. (That is, the index set can be finite or not; the individual groups can be finite or not.) Show that

$$H = \bigcap_{t \in \mathcal{I}} H_t$$

is a subgroup of $G$.

9. Let $X$ be any subset of the group $G$. $X$ need not be a subgroup. By a *word* in $X$ we mean any product.

$$x_1^{\epsilon_1} x_2^{\epsilon_2} \dots x_k^{\epsilon_k}$$

where $\epsilon_j = \pm 1$ and each $x_j \in X$. For example,

$$1 = x_1^1 \cdot x_1^{-1}$$
$$x_1^3 = x_1^1 x_1^1 x_1^1$$
$$x_1 x_2^{-1} x_1 x_3 x_4^{-1}$$

are words in $X = \{x_1, x_2, x_3, x_4\}$.

The subgroup *generated* by $X$ is the set of all words in $X$. We write

$$\langle X \rangle = \{\text{words } \ x_1^{\epsilon_1} \dots x_k^{\epsilon_k} \ \text{ in } \ X\}.$$

(a) Verify that $\langle X \rangle$ is indeed a subgroup. Hint: Use exercise 2.1.

(b) Show that $\langle X \rangle$ is the intersection of all subgroups of $G$ which contain $X$. (There is at least one such subgroup, namely $G$ itself.)

**Remark**: this provides an alternative definition for the subgroup generated by a subset $X$ of the group $G$.

Intuitively, we may therefore say that $\langle X \rangle$ is the *smallest* subgroup which contains the set $X$.

## 5.4 Normal Subgroups

Again $H$ is some subgroup of $G$, perhaps finite or not.

1. (a) We look at right cosets, though the Theorem in part (11) below shows that we could just as well use left cosets.

   (b) Now $H = H1$ is itself a coset, and

$$HH = H.$$

In fact, for any coset $Hb \quad (b \in G)$ we have

$$H(Hb) = (HH)b = Hb.$$

Thus $H$ acts like an identity for multiplication of cosets.

(c) Someone's wonderful idea was to try to make a new group, denoted

$$G/H$$

with:

i. the cosets $Hb \quad (b \in G)$ as individual elements.

ii. coset multiplication as operation.

iii. the coset $H$ itself as identity.

(d) The key difficulty in verifying that this even makes sense is that for certain subgroups $H$, multiplication of cosets need not be a closed operation.

**Example from a previous exercise**.
$G = \mathbb{S}_3$ (order 6)
$H = \langle (12) \rangle$ (order 2)
$a = (23) \qquad b = (13)$
Find $Ha, \ Hb$ then show that $HaHb$ is not even a coset.

(e) Thus we have good reason to study subgroups $H$ for which coset multiplication is closed.

**Definition**. $H$ is a *normal* subgroup of $G$, written

$$H \lhd G,$$

if coset multiplication is closed.

There are many useful and equivalent ways to say the same thing. Some of these equivalent ways are given in the next theorem.

It will be useful now to recall that

$$x \in Hg \text{ if and only if } Hg = Hx.$$

2. **Theorem 5.2** (**Criteria for Normality**)

The following items are equivalent for a subgroup $H$ of $G$.

(a) $H$ is normal in $G$ [meaning "coset multiplication is closed"].

(b) $(Ha)(Hb) = Hab$ for all $a, b \in G$.

(c) $a^{-1}Ha \subseteq H$ for all $a \in G$.

(d) $a^{-1}Ha = H$ for all $a \in G$.

(e) $Ha = aH$ for all $a \in G$.
Caution! This need not mean that $a$ commutes with all *individual* elements of $H$.

(f) Every right coset of $H$ equals some left coset.

**Proof**. We must show that each of the six conditions implies each of the five others, for a tentative total of 30 separate proofs!! However, we get the same result much more economically by showing

$$(a) \Rightarrow (b), (b) \Rightarrow (c), (c) \Rightarrow (d), (d) \Rightarrow (e), (e) \Rightarrow (f) \text{ and } (f) \Rightarrow (a).$$

Here are the details.

$(a) \Rightarrow (b)$ Assume $Ha,\ \ Hb$ are any cosets, so that

$$(Ha)(Hb) = Hg \qquad \text{for some unknown } g \in G.$$

$$\text{But then } (1a)(1b) = ab \in Hg \text{, so}$$

$$Hg = Hab$$

$$\text{Thus } (Ha)(Hb) = Hab.$$

$(b) \Rightarrow (c)$ For any $a \in G$,

$$Ha^{-1}Ha = H(a^{-1}a) = H1 = H.$$

$$\text{Now let } x \in a^{-1}Ha. \text{ Then } x = a^{-1}ha, \text{ for some } h \in H.$$

$$\text{Thus } x = 1a^{-1}ha \in Ha^{-1}Ha = H.$$

$$\text{So } x \in H.$$

Since $x$ was arbitrary in $a^{-1}Ha$, we get

$$a^{-1}Ha \subseteq H.$$

$(c) \Rightarrow (d)$ For any $a \in G$, $\quad a^{-1}Ha \subseteq H$. In particular, this is also true when $a$ is replaced by $a^{-1}$:

$$(a^{-1})^{-1}H(a^{-1}) \subseteq H$$

$$aHa^{-1} \subseteq H$$

$$\text{so } a^{-1}(aHa^{-1})a \subseteq a^{-1}Ha$$

$$\text{so } 1H1 \subseteq a^{-1}Ha$$

$$\text{so } H \subseteq a^{-1}Ha \subseteq H$$

$$\text{Thus } a^{-1}Ha = H.$$

$(d) \Rightarrow (e)$ If $a^{-1}Ha = H$, then

$$a(a^{-1}Ha) = aH$$

$$1Ha = aH$$

$$Ha = aH$$

$(e) \Rightarrow (f)$ Every right coset $Ha = aH$, a left coset.

$(f) \Rightarrow (a)$ Assume every right coset equals some left coset and consider any $a, b \in G$. We want to multiply two right cosets $Ha$ and $Hb$. But the left coset $aH$ is some right coset, say

$$*\ \ aH = Hc, \qquad \text{for some unknown } c.$$

Thus

$$\begin{aligned} (Ha)(Hb) &= H(aH)b \\ &= H(Hc)b \\ &= (HH)(cb) \\ &= H(cb), \end{aligned}$$

so that right coset multiplication is closed.

Since $a \in aH = Hc$, we could have chosen $c = a$, obtaining $aH = Ha$, then

$$(Ha)(Hb) = H(ab).$$

This finishes the proof. □

3. **Theorem 5.3** (**The Factor Theorem**)

   Suppose $H \lhd G$ ($H$ is a normal subgroup of $G$). Then

   $$G/H \quad \text{(the family of right cosets of } H\text{)}$$

   forms a group with coset multiplication. The identity is $H$, and $Ha$ has inverse $H(a^{-1})$.

   **Proof**. We know the operation is closed. It's easy to check associativity, the identity and inverses. □

   **Remark**. (a) $G/H$ is called a *quotient* group, or sometimes a *factor group.*

   (b) By Theorem 5.2(e), we could just as well use left cosets.

4. **Exercises**

   (a) $G \lhd G$ and $\{1\} \lhd G$. Thus, the trivial subgroups of $G$ are each normal.

   (b) If $G$ is abelian, then every subgroup is normal.

   (c) If $[G : H] = 2$, then $H \lhd G$.

   (d) If $H \lhd G$ and $[G : H] = k$, then

   $$|G/H| = \frac{|G|}{|H|}.$$

   (e) Generally a group $G$ has lots of subgroups, say the cyclic subgroups generated by one element $a$, and also subgroups generated by two or more elements.

   Also, by exercise (b) above, every subgroup of an abelian group is normal. However, for non-abelian groups $G$ it sometimes happens that normal subgroups are scarce. Such groups $G$ are interesting and important: we call them simple.

   **Definition**. A group $G$ is *simple* if it has *no* non-trivial normal subgroups. (Thus, $\{1\}$ and $G$ are the only normal subgroups.)

   **Remark**: in a sense, simple groups play the same role in group theory as prime numbers play in number theory. The actual details in either case are very deep and complicated.

   (f) Characterize all cyclic groups which are simple. (Hint: $\langle a \rangle$ of order $n$ or order $\infty$ is abelian: use exercise (b).)

(g) If $\{H_t : t \in \mathcal{I}\}$ is any family of normal subgroups of $G$, then

$$H = \bigcap_{t \in \mathcal{I}} H_t$$

is also a normal subgroup.

**Remark**. In particular, if $H_1$ and $H_2$ are normal, so is $H_1 \cap H_2$.

(h) **Definition**. If $x, b \in G$, then $x^{-1}bx$ is called a *conjugate* of $b$. If $S$ is any subset of $G$, let

$$\tilde{S} = \{x^{-1}sx \ : \ s \in S, \ \ x \in G\}$$

be the set of all conjugates of elements of $S$.

(i) Show that $H = \langle \tilde{S} \rangle$ is a normal subgroup of $G$ and that $H \subseteq S$.

Show that $H$ is the intersection of all normal subgroups of $G$ which contain $S$. (In some sense, $H$ is the *smallest* normal subgroup containing $S$.)

**Definition**. $H = \langle \tilde{S} \rangle$ is the *normal closure* of $G$.

# 6 More useful mappings on the plane $\mathbb{R}^2$.

1. **The reflection $r$ in a given line $m$**

   We have already defined reflections and have noted that each reflection $r : \mathbb{R}^2 \to \mathbb{R}^2$ is a bijection.

   In fact $r$ is a nice bijection, in that it preserves distances. One can prove this from the geometric definition of $r$, coupled with a few congruent triangles. Thus $r$ is an isometry, according to the following:

2. **Definition:** An *isometry* (of the plane $\mathbb{R}^2$) is a distance preserving bijection

   $$f : \mathbb{R}^2 \to \mathbb{R}^2 \ .$$

   Intuitively isometries preserve shape because of this. In this regard, it is useful to note the following

3. **Proposition 6.1**

   The collection of all points $P$ equidistant from two distinct points $P, Q$ is a line (the *perpendicular bisector* of segment $PQ$).

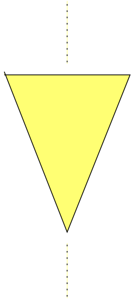
   **Proof**. Uses only a few applications of basic congruence theorems, such as SAS. □

4. Thus any isometry (in particular any reflection)

   - maps a circle to a (congruent) circle of the same radius
   - maps a straight line to a straight line
   - maps a triangle $\triangle ABC$ to a congruent $\triangle A'B'C'$

5. We have seen that any reflection $r$ is an *opposite isometry*: it maps any clockwise oriented triangle to a congruent but anticlockwise oriented triangle.

6. $r$ is defined on all of $\mathbb{R}^2$; in typical applications we restrict to a subset of interest

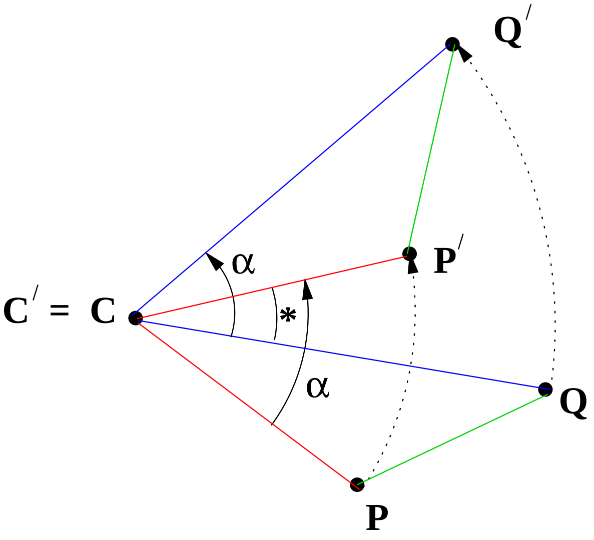


**Here the subset is an isosceles triangle, which has bilateral symmetry.**

7. **The rotation $s$ with centre $C$ and angle $\alpha$**

**Definition**. For each point $P \in \mathbb{R}^2$ let $P' = Ps$ be the same distance from $C$ as $P$ and located so that

$$\angle PCP' = \alpha.$$



(a) Again it is easy to check that this description is well-defined and that $s$ is invertible. Once more an application of SAS shows that $s$ is an isometry. You can see from the figure that $\triangle PQC$ and $\triangle P'Q'C'$ have the same *orientation*. (Note that the centre $C = C'$ is invariant.) Thus a rotation $s$ is a *direct* isometry.

(b) Note that we must distinguish clockwise from anticlockwise rotations. We do this in the usual manner, taking

$$\alpha : \begin{array}{l} \oplus \text{ if anti-clockwise} \\ \ominus \text{ if clockwise} \end{array} .$$

(c) If $\hat{s}$ is the rotation with the same centre $C$, but with the opposite angle $-\alpha$, then $Ps = P'$ implies $P'\hat{s} = P$. Thus

$$s\hat{s} = 1 = \hat{s}s.$$

We have shown that $\hat{s}$ is the inverse of $s$. Therefore the inverse of a rotation $s$ is also a rotation; and $s^{-1}$ has the same centre as $s$, but the opposite angle.

(d) If $\alpha = 0°,\ \pm 360°,\ n(360°)$ and $C$ is any centre, then

$$s = 1\ .$$

Thus the identity is actually a rotation with ambiguous centre.

Otherwise, if $\alpha \neq n(360°)$, where $n \in \mathbb{Z}$, then $s$ fixes only one specific central point $C$.

(e) In any rotation, the angles $\alpha+n(360°)$ all define the same rotation, here considered to be a particular mapping on the plane.

(f) **Definition:** The *half-turn* $h = h_C$ with centre $C$ is the rotation at $C$ with angle $180°$.

Observe that every half-turn is an involution.

(g) **Proposition 6.2**

The product of rotations $s_1$, $s_2$ with the *same* centre $C$ but possibly different angles $\alpha_1$, $\alpha_2$, is also a rotation, namely that with centre $C$ and angle $\alpha_1 + \alpha_2$. Hence, such rotations commute:

$$s_1 s_2 = s_2 s_1 \ .$$

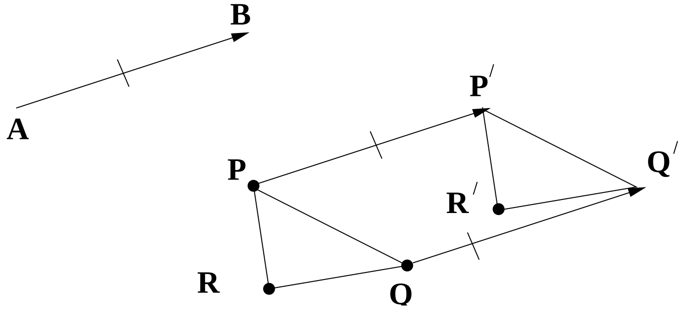
(This is unusual isometry behaviour.)

8. There could be other kinds of isometry. So far, we have two distinct species. In fact, it will follow from the three reflections theorem and the nature of Euclidean parallelism, that there are just two more species of Euclidean isometries: *translations* and *glides.*

9. **The translation $t$ with vector $\overrightarrow{AB}$**

Recall that vector $\overrightarrow{AB}$ is the directed line segment from point $A$ (the tail) to point $B$ (the head).

**Definition** The **translation** $t$ with vector $\overrightarrow{AB}$ maps point $P \in \mathbb{R}^2$ to the point $P' = Pt$ located so that $\overrightarrow{AB}$ and $\overrightarrow{PP'}$ are equal and parallel, in the same sense. (We say that these two vectors are equal, of course.)



(a) You may check that $t$ is an isometry. The verification uses properties of parallel lines and hence is heavily dependent on Euclidean parallelism.

(b) $t$ is direct: $\triangle PQR$ and $\triangle P'Q'R'$ have the same orientation.

(c) The inverse of translation $t$ is another translation; and $t^{-1}$ has vector $\overrightarrow{BA}= - \overrightarrow{AB}$.

(d) The identity 1 can be considered anew as a translation with the vector $\mathbf{0} =\overrightarrow{AA}$. The zero vector has length 0 and ambiguous direction.

(e) **Theorem**. The product of translations $t_1$, $t_2$ with vectors $\overrightarrow{A_1B_1}$, $\overrightarrow{A_2B_2}$, respectively, is also a translation, namely that with vector $\overrightarrow{A_1B_1} + \overrightarrow{A_2B_2}$ (obtained by 'head-to-tail' vector addition). Hence, all translations commute:

$$t_1t_2 = t_2t_1 \ .$$

(This, too, is unusual isometry behaviour.)

10. **The glide (or glide reflection) $g$ with non-zero 'hook' $\overrightarrow{AB}$**

**Definition** The **glide** $g$ with *hook* $\overrightarrow{AB}$ is the product of the translation $t$ with vector $\overrightarrow{AB}$ and the reflection $r$ in the line through $A$ and $B$. (We insist that $\overrightarrow{AB}\neq \mathbf{0}$ merely to guarantee that $A$, $B$ are distinct points.) Thus

$$g = rt \ .$$

(a) I use the non-standard word 'hook' in the hope of avoiding confusion. The hook $\overrightarrow{AB}$ is a directed line segment, so secretly it is a vector! But here it is issuing reflection as well as translation instructions.

(b) In this set up we actually have $rt = tr$. Prove this! Thus the definition is not as touchy as one might think.

(c) The line through $A$ and $B$ is callled the *axis* of the glide.

(d) Since we have defined a glide as a product of two isometries, it must itself be an isometry. In fact, $g^{-1}$ is a also a glide, with the negative hook $\overrightarrow{BA} = -\overrightarrow{AB}$. The actual axis is the same.

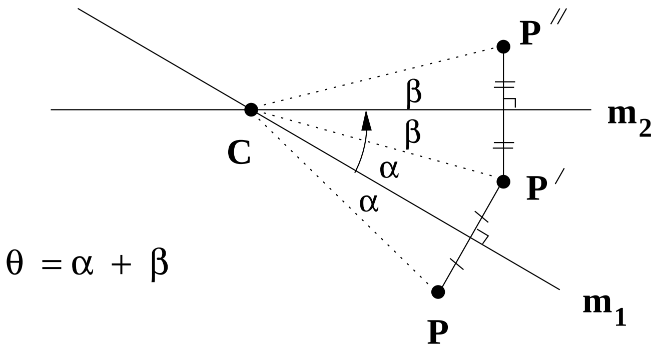
(e) Being a product of a direct and opposite isometry, a glide must itself be opposite.

11. **Summary: the four species of Euclidean plane isometries and core data**

| Type | Core Data | Sense | Fixed points | Remarks |
|---|---|---|---|---|
| reflection $r$ | mirror $m$ | opposite | all points on mirror $m$ | $r^{-1} = r$ is the same reflection |
| rotation $s$ | centre $C$ and angle $\alpha$ (signed) | direct | just $C$ if $s \neq 1$ | $s^{-1}$ is a rotation with same $C$, angle $-\alpha$ (identity 1 is a trivial rotation with angle 0°) |
| translation $t$ | vector $\overrightarrow{AB}$ | direct | none when $\overrightarrow{AB}\neq \mathbf{0}$ | $t^{-1}$ is translation with negative vector $-\overrightarrow{AB}=\overrightarrow{BA}$ (identity 1 is a trivial translation with vector $\mathbf{0}$) |
| glide $g$ | hook $\overrightarrow{AB}$ | opposite | none, if hook $\overrightarrow{AB}\neq \mathbf{0}$ | $g^{-1}$ is the glide with the negative hook $\overrightarrow{BA}$ |

# 7 Practical calculations with isometries.

1. **Theorem** Suppose lines $m_1, m_2$ meet at $C$ and the angle *from* $m_1$ *to* $m_2$ is $\theta$:



Let $r_j$ be the reflection with mirror $m_j$. Then $r_1 r_2$ is the rotation with centre $C$ and angle $2\theta$.

**Proof**. See the diagram. Note that this result does not depend on explicit properties of parallelism. It holds in non-Euclidean geometry, too. □

(a) Note the interaction of species here. A analogous result holds when $m_1$ and $m_2$ are parallel. In that case $r_1 r_2$ is the translation through *twice* the vector running orthogonally from $m_1$ to $m_2$. (See below.)

(b) How is $\theta$ ambiguous? Why doesn't it matter?

(c) Order does matter. Usually $r_1 r_2 \neq r_2 r_1$.

(d) If $m_1 = m_2$, then $r_1 = r_2$ and $r_1 r_2 = r_1^2 = 1$. Suppose $m_1 \neq m_2$. When does $r_1 r_2 = r_2 r_1$?

Answer: When $m_1 \perp m_2$ and then $r_1 r_2 = r_2 r_1 = h_C$.

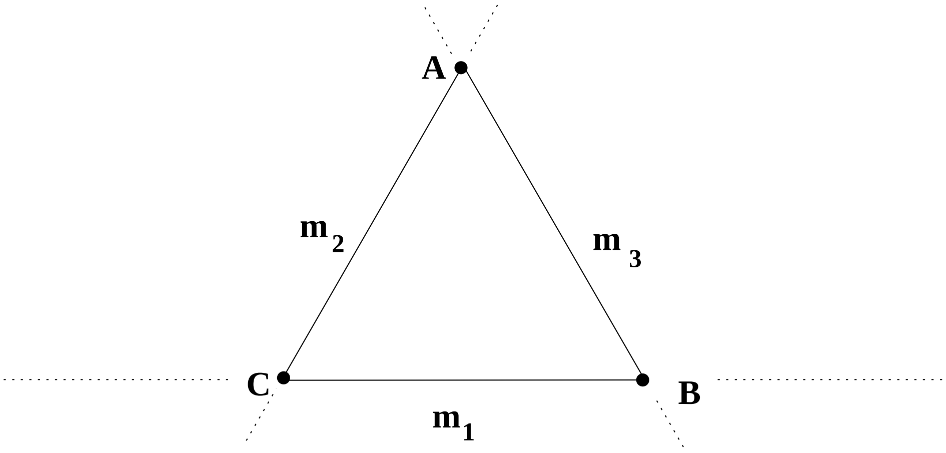
(e) $r_1 r_2 = s =$ rotation with centre $C$, and angle $2\theta$.
↑↑ ↑
mirrors no mirrors any more!
$m_1, m_2$ through
$C$

The devious insight here comes from turning this around. Given a rotation $s$ we can factor

$$s = r_2 r_1$$

choosing $m_1$ (or $m_2$) through $C$ arbitrarily, adjusting $m_2$ (or $m_1$) appropriately.

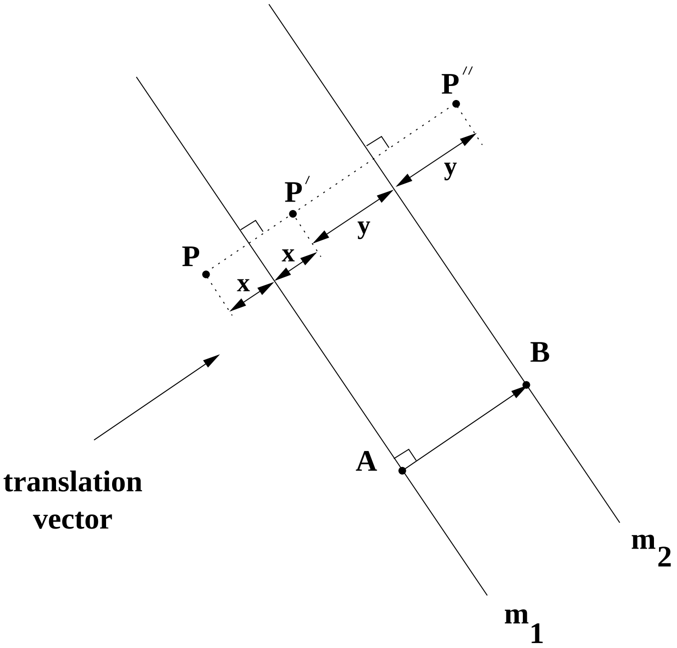
2. **Exercise**. In this picture, each $m_j$ is a line along an (extended) side of equilateral $\triangle ABC$:



   (a) Let $r_j$ be reflection in mirror $m_j$. Compute $r_2 r_1$, $r_1 r_2$, $r_1 r_2 r_1$ and $r_2 r_1 r_2$.

   (b) Let $s = 60°$ rotation at $C$ and $\tilde{s} = 60°$ rotation at $B$. Compute $s\tilde{s}$, $\tilde{s}s$ and $s\tilde{s}^{-1}$.

3. **Theorem** Suppose lines $m_1, m_2$ are parallel and the vector running perpendicularly *from* $m_1$ *to* $m_2$ is $\overrightarrow{AB}$:



   Let $r_j$ be the reflection with mirror $m_j$. Then $r_1 r_2$ is the translation with vector $2\ \overrightarrow{AB}$.

   **Proof**. Easy. □

4. **Exercise**. Return to the equilateral triangle in item (2) just above. Compute and classify the isometry $r_1 r_2 r_3$.

## 7.1 Addition of Angles.

1. **Theorem**. Suppose $p$ and $q$ are rotations with angle $\alpha$, centre $A$ and angle $\beta$, centre $B$, respectively. Then $s = pq$ is a rotation (or translation) with angle $\gamma = \alpha + \beta$; its centre can be determined using the method implicit in the proof.

   In particular, if $\alpha + \beta$ is an integral multiple of 360° ($0°, \pm 360°$, etc.), then $s = pq$ is a translation, whose vector can likewise be deduced from the proof (or otherwise).

   **Proof**. Done in class. Or see *Geometry in a Nutshell*, pages 138ff. □

2. **Corollary**. The product of two rotations $p$ and $q$ are rotations with angles $\alpha$ and $\beta$, respectively, and the *same* centre $A$, is the rotation with angle $\gamma = \alpha + \beta$ and also centre $A$. Only in such cases do rotations commute:

$$pq = qp \, .$$

   **Remark**. This is intuitively clear anyway, as we have earlier observed in §1.5.1 (g)

3. **Corollary**. The product of half-turns $h_M h_N$ is the translation with vector $2\vec{MN}$.

## 7.2 Justifying our Intuition.

1. **Proposition 7.1**

   If $ABC$ is any triangle in $\mathbb{R}^2$, then each point $P \in \mathbb{R}^2$ is uniquely determined by its distances (in order) to $A, B, C$.

   **Proof**. Use the Proposition 6.1 on the perpendicular bisector. □

2. **Proposition 7.2** (The action of an isometry $f$ on any specific triangle)

   Suppose

   $$f : ABC \to A'B'C'.$$

   Then

   $$\triangle ABC \equiv \triangle A'B'C'.$$

   **Proof**. Use SSS. □

3. **Proposition 7.3**

   Suppose $ABC$ are the vertices of a triangle. If $f : ABC \to ABC$, then $f = 1$.

   (We mean of course that $f$ maps the vertices in order; thus $f$ fixes each vertex of $\triangle ABC$ .)

   **Proof**. Again use Proposition 6.1 on the perpendicular bisector. □

4. **Proposition 7.4**

   If both $f, g = ABC \to A'B'C'$, then $f = g$.

   **Proof**. Use the previous result! □

   **Meaning**: Each isometry is completely determined by its effect on one particular triangle. There may be a convenient triangle which makes the calculations easy.

5. **Exercise**. Reprove the theorem that a product of reflections in intersecting mirrors is a particular rotation. Ditto when the mirrors are parallel.

6. **Theorem 7.1 (The Three Reflections Theorem)**

   Any isometry $f : \mathbb{R}^2 \to \mathbb{R}^2$ is a product of at most three reflections.

   **Remark**: This is an absolute theorem. It holds just as well in the non-Euclidean plane $\mathbb{H}^2$. However, the details for the various species there play out in a slightly different way. Indeed in $\mathbb{H}^2$ there are 5 rather than 4 species of isometry.

   **Proof**. Pick any one triangle $\triangle ABC$ to work with. Suppose $f : ABC \to A'B'C'$, so that $AB = A'B'$, $AC = A'C'$ and $BC = B'C'$.

   (a) If necessary, i.e. if $A \neq A'$, apply to $\triangle ABC$ reflection $r_1$ in the perpendicular bisector of segment $AA'$. Thus $r_1$ maps $\triangle ABC$ to the congruent triangle $\triangle A'B''C''$. Thus $A'B' = AB = A'B''$, so that $A'$ is on the perpendicular bisector of segment $B'B''$.

   (b) If necessary, i.e. if $B' \neq B''$, apply reflection $r_2$ in the perpendicular bisector of segment $B'B''$. Thus $r_2$ fixes $A'$ and maps $\triangle A'B''C''$ to the congruent triangle $\triangle A'B'C'''$.

   (c) Now both $A'$ and $B'$ are on the perpendicular bisector of segment $C'C'''$. If necessary, i.e. if $C''' \neq C'$, apply reflection $r_3$ in the perpendicular bisector of segment $C'C'''$. Then $r_3$ maps $\triangle A'B'C'''$ to the congruent triangle $\triangle A'B'C'$.

   In summary, the product $r_1r_2r_3$ (with the unnecessary reflections deleted) maps $\triangle ABC$ to $\triangle A'B'C'$. Thus $f = r_1r_2r_3$. □

7. (a) **Corollary 7.1**

   If $f$ fixes a point $O$, at most 2 reflections are required. Thus every isometry fixing a point $O$ is a rotation centred at $O$, or a reflection in some line through $O$.

   (b) **Corollary 7.2**

   If $f$ fixes two points $A \neq B$, then either $f = 1$ or $f$ is the reflection in the line $AB$.

   (c) **Remarks**: From these results, it is now a routine matter to classify all isometries acting on the Euclidean plane: there are just four species. Similarly, in Euclidean $n$-space $\mathbb{R}^n$, every isometry is the product of at most $n + 1$ reflections; one can classify all isometries there, too, but the details are rather more complicated.

## 7.3 Isometry Groups

Let us gather together all possible isometries of the plane.

1. **Definition** Let $\mathcal{I}$ be the collection of all isometries on the plane $\mathbb{R}^2$. Thus $\mathcal{I}$ is a very infinite set of things, which we know how to multiply. Indeed, we have already verified these properties for isometries $f, g, h \in \mathcal{I}$:

   **closure** The product of two isometries $f$ , $g$ is another isometry $fg$.

   **associativity** Always $(fg)h = f(gh)$.

   **identity** The identity 1 is an isometry.

   **inverses** Each isometry $f$ has an inverse $f^{-1}$.

   Accordingly we say that $\mathcal{I}$ is a *group*.

2. Since isometries preserve distances betweem pairs of individual points, we expect that they preserve the overall shape and structure of all 'macroscopic' objects. Convince yourself that any isometry

   - maps a circle to a (congruent) circle of the same radius
   - maps a straight line to a straight line
   - maps a triangle $\triangle ABC$ to a congruent $\triangle A'B'C'$

3. Interesting subgroups of $\mathcal{I}$.

   **Definition**. By a figure $K$ in the plane we mean any subset $K \subseteq \mathbb{R}^2$.

   (a) The *symmetry group* of $K$ is

   $$\text{Sym}(K) = \{ \text{ isometries } f \text{ which map } K \text{ onto itself (globally)}\}$$

   Note that although $(K)f = K$ for all $f \in \text{Sym}(K)$, it is quite possible for the constituent points $P \in K$ to move 'internally'.

   Why is a Sym$(K)$ a group, more precisely, a subgroup of $\mathcal{I}$?

   (b) In more restricted fashion, we define Fix$(K)$ to be the collection of all isometries which fix each point of $K$.

   Verify that Fix$(K)$ is a subgroup of $\mathcal{I}$.

   (c) In fact we have these subgroup relationships:

   $$\text{Fix }(K) \subseteq \text{ Sym }(K) \subseteq \mathcal{I}.$$

4. Examples. Clearly describe all isometries in Fix $(K)$ and Sym $(K)$ when

   (a) $K = \{A, B\}$ (two distinct points).

   (b) $K$ is a line $m$.

   (c) $K$ is a circle.

(d) $K = \emptyset$ (the empty set).

5. Let's focus on the key example with $K = \{O\}$, one specific point.
   In this case Fix $(O) =$ Sym $(O)$ !

   **Definition**. The resulting infinite group is called the *orthogonal group* for the plane. We sometimes denote it by $O(\mathbb{R}^2)$ or $O_2(\mathbb{R})$.

   Convince yourself that the orthogonal group $O_2(\mathbb{R})$ consists of all rotations with centre $O$ (including 1), together with all reflections in lines through $O$.

   The rotations alone constitute a subgroup of index 2 in $O(\mathbb{R}^2)$; this subgroup is called the *special orthogonal group* and is denoted $SO_2(\mathbb{R})$.

6. **Note**: Rather similar things happen in $n \geq 3$ dimensions, though the details are somewhat more intricate. For example, in Euclidean 3-space there are 6 species of isometry.

## 7.4 A detour into patterns and orbifold notation

1. **Rosettes**. Conway and his coauthors use this term as a convenient way to describe all plane figures $K$ which have *finitely* many symmetries. Here are some examples

   - The square is a a rather plain rosette with 8 symmetries. Both pattern and group are indicated by the symbol $*4\bullet$.
   - The equilateral triangle is a rosette of type $*3\bullet$.
   - A perfect letter H is a rosette of type $*2\bullet$.
   - A perfect letter Z is a rosette of type $2\bullet$. It has no reflection symmetry.

2. **The Theorem of Leonardo da Vinci**. In his book *Symmetry*, Hermann Weyl claims that Leonardo discovered that

   *The only finite groups of isometries in the plane are the cyclic groups of order $p$ and the dihedral groups of order $2p$, $(p = 1, 2, 3, \ldots)$.*

   **Notation** The cyclic group of order $p$ is commonly denoted $C_p$ or $Z_p$; Conways orbifold notation is $p\bullet$ (let's say "$p$" point or "$p$ dot"). This group of plane isometries is comprised of the powers of a single rotation generator. Hence it contains only rotations.

   Conway labels this as *gyrational point symmetry*, typically marked up in blue. For blue think 'preserve the true orientation'.

   The dihedral group of order $2p$ is often denoted $D_p$ (or a variant of this). The orbifold notation is $*p\bullet$ (" star $p$ point"). This groups contains the $p$ rotations of $p\bullet$ together with the $p$ reflections in symmetrically placed mirrors through thethe Lemma centre of the pattern ("the point").

   Conway labels this as *kaleidoscopic symmetry*, typically marked up in red. For red think reflect.

   **Proof**. Perhaps done in class. Or see Coxeter, *Regular Complex Polytopes*, page 3. □

3. Some special instances of orbifold notation.

   (a) A regular $p$-sided polygon has $*p\bullet$ symmetry.

   (b) A $*$ alone indicates *bilateral symmetry*. It makes sense to agree that $*1\bullet$ and $*\bullet$ denote the same thing.

   (c) The symbol $1\bullet$ denotes *trivial symmetry*. There is just one symmetry in such a group, necessarily the identity 1 (since every group at a minimum contains an identity).

   If you think about it, trivial symmetry is the norm, since a random figure will have imperfections or odd features which prevent any non-trivial symmetry. Of course, we typically deal with idealized figures in which imperfections do not exist or are simply ignored.

# 8 Representing isometry groups by permutations

1. Let $K$ be any subset of $\mathbb{R}^2$, infinite or not. In many cases of interest there are *finitely many points* $P_1, \ldots, P_n \in K$ which are permuted amongst themselves by all isometries $f \in \text{Sym}(K)$. (Think of a square $K$ which does have infinitely many points; but its four vertices $P_1, P_2, P_3, P_4$ play a special role and will serve to keep track of how any isometry operates on the whole square.) In short, we have

$$\text{Sym}(K) \subseteq \text{Sym}(\{P_1, \ldots, P_n\}) .$$

In such cases, we can define a function

$$\begin{aligned} \varphi : \text{Sym}(K) &\rightarrow \mathbb{S}_n \\ f &\mapsto f\varphi \end{aligned}$$

as follows. For visual ease, let's temporarily let $f\varphi = \pi$. Thus $\pi$ should be a permutation on $[n] = \{1, \ldots, n\}$. For $i \in \{1, \ldots, n\}$, we will set $i\pi = j$ if $f$ maps $P_i$ to $P_j$. In short $\pi$ permutes subscripts in the same way that the isometry $f$ permutes the special points $P_i$.

Briefly then, the permutation $f\varphi$ is determined by

$$(P_i)f = P_{i(f\varphi)}, \quad 1 \leq i \leq n. \tag{2}$$

This is how we track isometries by their *action* on a key set of points.

**Caution-rethink this!**. Our description above is correct but a little glib. In order for it to make sense, we have to agree to a couple of things:

- the isometries $f$ are stationary in space. Although the object $K$ may move when we apply $f$, the isometry $f$ stays put. It is useful to think of the isometries as sitting in the background, with the object $K$ moving in the foreground.

  For example, say $r$ is reflection in a 'vertical' line of symmetry for a square and suppose $s$ is a 90° rotation. Then applying $s$ does not rotate the mirror for $r$; the mirror stays vertical in the background.

- Likewise, the points $P_i$ aren't attached to the object $K$. Instead, they refer to stationary positions in the background. One should not here think of the $P_i$'s as labels attached to points on the object $K$.

  For example, returning to the square, we should think of $P_1, P_2, P_3, P_4$ as fixed background positions, as in NW, NE, SE, SW. Then the 'vertical' reflection $r$ does swap the vertices of the square in the NW and NE positions, as well as those in the SW and SE positions. Hence,

  $$r\varphi = (1, 2)(3, 4) .$$

  Likewise, the rotation $s$ (which is 90° anticlockwise), maps the vertex in the background NW position to the vertex in the SW position, and the latter vertex to that in the SE position, etc. Thus,

  $$s\varphi = (1, 4, 3, 2) .$$

2. **Way to think**: isometry $f$ yields a permutation mapping $i$ to $j$, if it maps the vertex in background position $i$ to the vertex in background position $j$.

   The same methodology works if $\mathrm{Sym}(K)$ permutes some $n$ features of the object $K$ amongst themselves. For example, we could see how the symmetries permute the 4 *edges* of a square, rather than the 4 vertices. In fact, we will get essentially the same group of permutations as with vertices.

   But consider a cube $K$ in space: there we must get different kinds of permutations if we permute the 8 vertices versus the 6 square facets (since $8 \neq 6$, after all!). However, in both cases we will get a group of order 48, and these two groups must be isomorphic since each faithfully represents the 48 symmetries of the cube.

3. **Theorem 8.1**

   Suppose $K$ is a figure in $\mathbb{R}^m$ with $n$ distinguished points $P_1, \ldots, P_n$ permuted amongst themselves by $\mathrm{Sym}(K)$. Then the map

   $$\varphi : \mathrm{Sym}(K) \to \mathbb{S}_n$$

   described above is a group homomorphism.

   Furthermore, $\varphi$ is 1–1 if the $n$ points do not all lie in one hyperplane. (For example, in the plane (ambient dimension $m = 2$), the points should not all lie on one line.) In such cases, we have a *faithful* representation of the symmetry group by permutations.

   A similar representation holds if $\mathrm{Sym}(K)$ permutes some set of $n$ features of the object $K$; however, assessing faithfulness can be trickier.

   **Proof**. This is a routine check.

   In the planar case $\mathbb{R}^2$, if $P_1, \ldots, P_n$ *do not all lie on one line*, then $\varphi$ is $1 - 1$, by Proposition 7.3 just above. □

   **Remarks**. If $\varphi$ is 1–1, $\mathrm{Sym}(K)$ can be identified with a subgroup of $\mathbb{S}_n$, the symmetric group on $n$ symbols. In particular, $\mathrm{Sym}(K)$ is finite. For example, the symmetry group of the square, which has order 8, is a subgroup of $\mathbb{S}_4$ (order 24).

   It is clear that there is nothing special about dimension 2. A similar result must hold for figures in higher dimension. For example, in $\mathbb{R}^3$ we obtain a faithful representation of $\mathrm{Sym}(K)$ if the points $P_1, \ldots, P_n$ *do not all lie on one plane.*

# 9 Coordinates

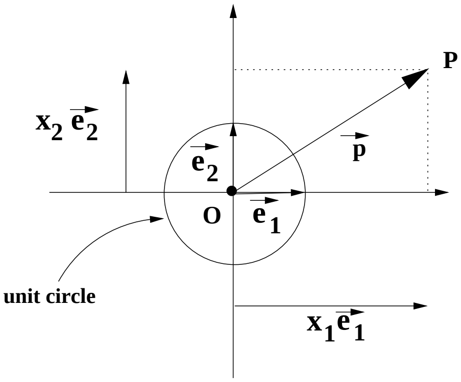
Let $O$ be any base point (origin). We thus know that the orthogonal group $O(\mathbb{R}^2) = \mathrm{Fix}(O)$ consists of all rotations centred at $O$ together with all reflections in lines through $O$.

**Project**. Describe $O(\mathbb{R}^2)$ algebraically using coordinates and matrices.

1. Having fixed an origin $O$, let's introduce the standard basis vectors

   $$\mathbf{e_1} = [1, 0] \quad \text{and} \quad \mathbf{e_2} = [0, 1] \ .$$

   These vectors are *orthonormal*: mutually perpendicular and each of length 1.



2. Any point $P \in \mathbb{R}^2$ can be located by its position vector $\mathbf{p} = \overrightarrow{OP}$. Since $\{\mathbf{e_1}, \mathbf{e_2}\}$ is a basis, *there exist unique* real numbers $x_1, x_2$ such that

   $$\mathbf{p} = x_1\mathbf{e_1} + x_2\mathbf{e_2} \ .$$

   Observe that $(x_1, x_2)$ are the usual *rectangular coordinates* for $P$.

   **Remarks**

   (a) The fact that $\{\mathbf{e_1}, \mathbf{e_2}\}$ is a basis is equivalent to unique coordinates existing for each point $P \in \mathbb{R}^2$. This in turn is equivalent to having a *linearly independent spanning set* of vectors (here $\{\mathbf{e_1}, \mathbf{e_2}\}$).

   (b) In the plane, any two vectors, neither of which is a multiple of the other, will serve as an alternate basis. Sometimes calculations are greatly simplified by working with a non-standard basis.

   (c) For the purposes of calculations to come, it is very helpful to assemble the coordinates into a 1 × 2 row vector. Let us simply write

   $$\mathbf{p} = [x_1, x_2] \ .$$

   It turns out that rows serve better than columns, since we compose functions left to right. If one composes right to left (the usual way in Calculus), then column vectors are better.

(d) As examples, note that

$$\mathbf{e_1} = [1,0]\,,\quad \mathbf{e_2} = [0,1]\,,\quad \mathbf{0} = \overrightarrow{OO} = [0,0] \quad .$$

(e) The geometrical linear combination

$$\mathbf{p} = x_1\mathbf{e_1} + x_2\mathbf{e_2}$$

becomes this component-wise calculation with row vectors:

$$[x_1, x_2] = x_1\,[1,0] + x_2\,[0,1] \quad .$$

(f) Having made these connections, we now understnad how the plane is coordinatized by the 2-dimensional vector space $\mathbb{R}^2$.

3. A *linear transformation* on $\mathbb{R}^2$ is a function which 'respects the vector' operations:

$$f : \mathbb{R}^2 \to \mathbb{R}^2$$

where

$$\begin{aligned} (\mathbf{p}+\mathbf{q})f &= \mathbf{p}f + \mathbf{q}f \\ (t\mathbf{p})f &= t(\mathbf{p}f) \end{aligned}$$

for all vectors $\mathbf{p}, \mathbf{q} \in \mathbb{R}^2$ and all scalars $t \in \mathbb{R}$.

**Remarks**:

(a) In general, $f$ need not be $1-1$ or onto, although orthogonal isometries, being bijections, do have these properties.

(b) More generally, the domain and range could be different vector spaces, as in $f : V \to W$.

(c) Taking $t = 0$, we conclude that

$$(\mathbf{0})f = \mathbf{0}$$

is forced.

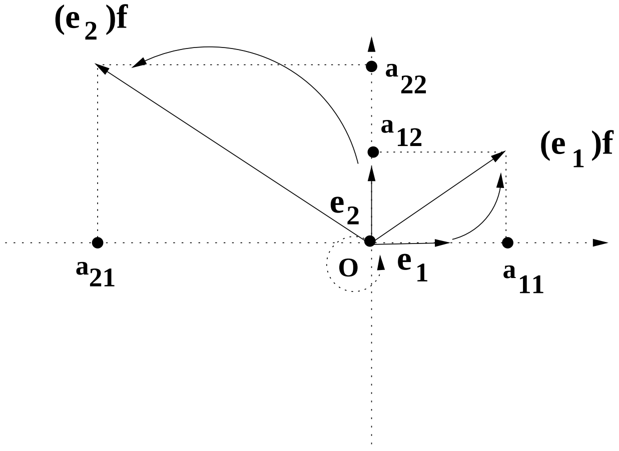
Now $(\mathbf{e_1})f$ and $(\mathbf{e_2})f$ are specific vectors. Suppose

$$(\mathbf{e_1})f = [1,0]f = [a_{11}, a_{12}]$$

and

$$(\mathbf{e_2})f = [0,1]\,f = [a_{21}, a_{22}] \quad .$$

Thus in the scalar $a_{ij}$, the subscript $i$ indicates the input number and the subscript $j$ indicates the coordinate position.



(The transformation $f$ suggested in the figure definitely distorts distances and so could not represent an isometry.)

In general, if we apply $f$ to

$$\mathbf{p} = [x_1, x_2] = x_1\,[1, 0] + x_2\,[o, 1]\ \ ,$$

we obtain

$$\begin{aligned}(\mathbf{p})f &= (x_1\mathbf{e_1} + x_2\mathbf{e_2})f \\ &= (x_1\mathbf{e_1})f + (x_2\mathbf{e_2})f \\ &= x_1(\mathbf{e_1}f) + x_2(\mathbf{e_2}f) \\ &= x_1\,[a_{11}, a_{12}] + x_2\,[a_{21}, a_{22}] \\ &= [\,x_1a_{11} + x_2a_{21}\ ,\ x_1a_{12} + x_2a_{22}\,] \\ &= [x_1, x_2]\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22}\end{bmatrix} \\ &= \mathbf{p}A\ \ ,\end{aligned}$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22}\end{bmatrix}$$

is the fixed coefficient matrix for the linear transformation $f$. Note the natural roles of matrix addition, scalar multiplication and matrix multiplication.

4. In general, having chosen explicit bases, every linear transformation yields a matrix. The algebraic interaction of the linear transformations is exactly paralleled by the algebraic interaction of the matrices. (Technically, we have an *algebra isomorphism.*) For us, the key things to note are that

   (a) If $f$ and $g$ are linear transformations on $\mathbb{R}^2$, with $2 \times 2$ matrices $A$ and $B$, respectively, then $fg$ is also a linear transformation; and its matrix is the product $AB$.

(b) The identity $1 = 1_{\mathbb{R}^2}$ is a linear transformation; and its matrix is the identity matrix

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} .$$

(c) If the linear transformation $f$ is bijective, then the inverse function $f^{-1}$ is also a linear transformation; and if $f$ has matrix $A$, then $f^{-1}$ has matrix $A^{-1}$ (the matrix inverse).

You should ponder and prove these claims.

One conclusion to be made from these observations is that the collection of invertible linear transfromations from a vector space $V$ to itself forms a group (as usual with compostion of functions for the operation).

**Definition**: This group is denoted $GL(V)$. Likewise, the collection of all invertible $2 \times 2$ real matrices forms a non-abelian group, denoted

$$GL_2(\mathbb{R}) .$$

5. **Keep in mind**: we could keep the same transformations, but change from the usual orthonormal basis $\{\mathbf{e_1}, \mathbf{e_2}\}$ to any other basis. The resulting matrices would very likely change, yet still describe the same geometric situation.

   Thus, we may guess that a wise, even unconventional, choice of basis may greatly simplify the matrix calculations.

6. We have strayed a bit from isometries into much more general territory. Let's return to isometries.

   **Theorem 9.1**

   Any isometry $f$ on $\mathbb{R}^2$ which fixes a point $O$ is an invertible linear transformation. With respect to the usual orthonormal basis $\{\mathbf{e_1}, \mathbf{e_2}\}$, each isometry $f$ is represented by an *orthogonal* matrix $A$, namely a matrix satisfying

$$AA^T = I .$$

   (Thus, very simply, $A^{-1} = A^T$.)

   The orthogonal group $O(\mathbb{R}^2)$, consisting of all isometries fixing an origin $O$, is isomorphic to the group $O_2(\mathbb{R})$ of all orthogonal $2 \times 2$ matrices. The direct isometries (rotations centred at $O$) correspond to orthogonal matrices with determinant $+1$. The opposite isometries (reflections in mirrors through $O$) correspond to orthogonal matrices with determinant $-1$.

   **Proof**: The key is to remember that isometries preserve shapes, such as the shape of the parallelogram that underlies vector addition. Also an isometry preserves the ratios of lengths that underly scalar multiplication. One checks then that an isometry fixing $O$ induces a linear transformation on $\mathbb{R}^2$.

In short, any isometry $f$ fixing a point $O$ is represented by some sort of $2 \times 2$ real matrix $A$. But what special property of $A$ comes from $f$ being an isometry (= shape-preserving)?

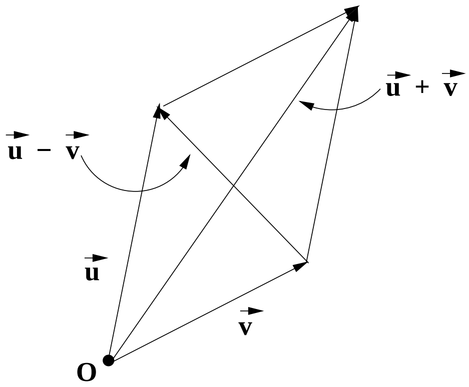
First, we note that $f$ must preserve inner products To see this, suppose

$$\mathbf{u} = [u_1, u_2]\,, \mathbf{v} = [v_1, v_2]\ .$$

Then

$$\begin{aligned} \mathbf{u} \cdot \mathbf{v} &= u_1 v_1 + u_2 v_2 \\ &= \frac{1}{4}[\ (u_1 + v_1)^2 + (u_2 + v_2)^2 - (u_1 - v_1)^2 - (u_2 - v_2)^2\ ] \\ &= \frac{1}{4}(\ \ \|\mathbf{u} + \mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2\ \ )\ . \end{aligned}$$

This *polarization identity* says that the inner product can be expressed in terms of the side and diagonal lengths in a suitable parallelogram:



Since any isometry $f$ preserves the shape and size of such parallelograms, it must also preserve inner products:

$$(\mathbf{u}f) \cdot (\mathbf{v}f) = \mathbf{u} \cdot \mathbf{v}, \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^2\ .$$

But

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}\mathbf{v}^T\ ,$$

where the $1 \times 1$ matrix product on the right is treated as a simple scalar. Hence, *for all* vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$, we have

$$\begin{aligned} (\mathbf{u}A)(\mathbf{v}A)^T &= \mathbf{u}\mathbf{v}^T \\ \mathbf{u}AA^T\mathbf{v}^T &= \mathbf{u}I\mathbf{v}^T \end{aligned}$$

where $I$ is the $2 \times 2$ identity matrix. Since $\mathbf{u}, \mathbf{v}$ are arbitrary, we conclude that $AA^T = I$. □

7. **Exercises**

(a) Give orthogonal matrices which describe

- reflection in the $x$-axis
- reflection in the $y$-axis
- the half-turn $h_O$ centred at the origin
- the identity 1

(b) Rotation matrices.

- Give the *rotation matrix* $A_\alpha$ for the rotation $s_\alpha$ centred at $O$, through angle $\alpha$.
- What is $A_0$ ?
- What is $A_{-\alpha}$?
- Describe – on geometrical grounds – the product of isometries $s_\alpha s_\beta$.
- Use the previous part to rewrite

  $$A_\alpha A_\beta$$

  as a single rotation matrix.
- Look at the entries in the resulting matrix equation. What important facts have you proved?

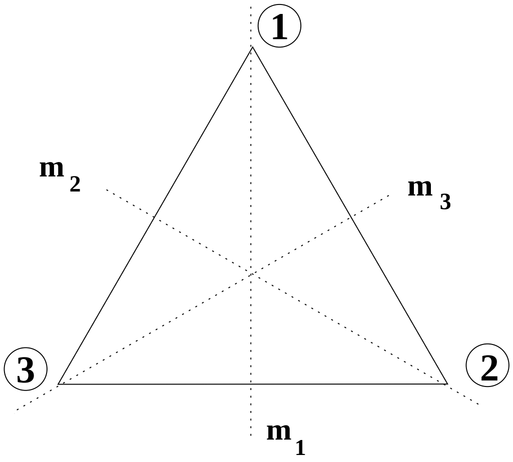
# 10 One Group from Several Points of View — Abstraction

1. **Geometrical Symmetry**: let $G$ be the group of symmetries for an equilateral triangle. We know that there are three rotations, including the identity, say $1, s_1, s_2$, together with three reflections $r_1, r_2, r_3$. Thus

   $$G = \{1, s_1, s_2, r_1, r_2, r_3\} \ ,$$

   with left to right composition as usual.

   Of course, $|G| = 6$.

   **Exercise**. Write out the multiplication table for $G$. Remember that $fg$ means first apply the isometry $f$ to the triangle, then the isometry $g$.

2. **Permutations**: label the vertices of the triangle $1, 2, 3$. Since each isometry of the plane is determined by its effect on this triangle, we can unambiguously track the isometries via permutations of $\{1, 2, 3\}$. We obtain the permutation group

   $$\mathbb{S}_3 = \{(\,), (1,3,2), (1,2,3), (2,3), (1,3), (1,2)\}$$

   (again composed left to right as functions).

   We have seen that $G \simeq S_3$. Explicitly, there is an isomorphism mapping

   $$\begin{array}{rcl} G & \to & \mathbb{S}_3 \\ 1 & \mapsto & (\,) \\ s_1 & \mapsto & (1,3,2) \\ s_2 & \mapsto & (1,2,3) \\ r_1 & \mapsto & (2,3) \\ r_2 & \mapsto & (1,3) \\ r_3 & \mapsto & (1,2) \end{array}$$

3. **Matrices (version 1) : orthogonal**. Place the origin $O$ at the centre of the triangle. Thus every symmetry of the triangle fixes $O$.

   Compute relative to the usual **orthonormal** basis. After rescaling the triangle, we may assume that the top vertex is

$$\mathbf{e_2} = [0, 1] \quad .$$

   As usual,

$$\mathbf{e_1} = [1, 0]$$

   is the unit vector pointing east. (It extends outside the triangle a little.) We then get a matrix group $M1$ in which the above isometries are represented in order as

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} -1/2 & \sqrt{3}/2 \\ -\sqrt{3}/2 & -1/2 \end{bmatrix}, \begin{bmatrix} -1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{bmatrix},$$

$$\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1/2 & -\sqrt{3}/2 \\ -\sqrt{3}/2 & -1/2 \end{bmatrix}, \begin{bmatrix} 1/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{bmatrix} \quad .$$

   Here each matrix is orthogonal: to get the inverse, simply transpose.

4. **Matrices (version 2) : nice but not orthogonal**

   We can actually employ any basis that we want. But it makes sense to choose a 'nice' basis. So let's take vertices 1 and 2 of the triangle as the new basis vectors $\mathbf{d_1}$ and $\mathbf{d_2}$. Because the triangle is equilateral, we see that vertex 3 is given by $-\mathbf{d_1} - \mathbf{d_2}$. A little computation gives a new set $M2$ of matrices for the original isometries, again in the original order:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} -1 & -1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix},$$

$$\begin{bmatrix} 1 & 0 \\ -1 & -1 \end{bmatrix}, \begin{bmatrix} -1 & -1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad .$$

   Thus the entries of these new matrices are a little nicer to work with.

   We have the same group, of course; but since the basis is non-standard, the corresponding coordinates are non-standard and measurement works differently. For example, the usual inner product $x_1y_1 + x_2y_2$ using new coordinates *does not* usefully measure anything.

5. The **trace** of a square matrix $A$ is the sum of its diagonal entries, say

$$\mathrm{tr(A)} := \sum_{\mathrm{j}} \mathrm{a_{jj}} \quad .$$

   Thus the trace of a matrix is a very special scalar.

   Notice that corresponding matrices in the above groups have identical traces. Why is this so?

Well, we have changed basis according to this rule:

$$\mathbf{d_1} = \mathbf{e_2} = 0\mathbf{e_1} + 1\mathbf{e_2} \ \ , \ \ \mathbf{d_2} = (\sqrt{3}/2)\mathbf{e_1} + (-1/2)\mathbf{e_2} \ .$$

Thus the corresponding *basis change matrix* is

$$B = \begin{bmatrix} 0 & 1 \\ \sqrt{3}/2 & -1/2 \end{bmatrix} .$$

Symbolically we should think

$$\text{(new basis } \mathbf{d_1}, \mathbf{d_2} \text{ in a column)} = B \text{ (old basis } \mathbf{e_1}, \mathbf{e_2} \text{ in a column).}$$

It follows that if $A$ is one of the six 'old' matrices in $M1$, then the corresponding 'new' matrix in $M2$ is

$$BAB^{-1} \ .$$

**Remark**: the exact arrangement of matrices here is a little tricky. Of course, much the same procedure works in $n$ dimensions.

Let's return to the traces. It is easy to check for square $n \times n$ matrices $A$ and $C$ that

$$\mathrm{tr(AC)} = \mathrm{tr(CA)} \ .$$

(Do this as an exercise.) Thus

$$\begin{aligned} \mathrm{tr((BA)B^{-1})} &= \mathrm{tr(B^{-1}(BA))} \\ &= \mathrm{tr((B^{-1}B)A)} \\ &= \mathrm{tr(IA)} \\ &= \mathrm{tr(A)} \ . \end{aligned}$$

In short, basis change does not change the trace values for matrix group representations of the original group $G$.

These trace values are called the **character values** for the matrix representation. Indeed, they serve to classify and distinguish essentially different matrix representations for one and the same group $G$.

In a sense, the character values (traces) contain just enough numerical information to completely determine the matrix group (up to a change in basis). All other numerical data in the matrices is clutter.

6. **Exercise**. Prove that conjugate elements in $G$ must have identical character values.

   The upshot, which is quite hard to prove, is that a matrix group is determined by $k$ scalars, where $k$ is the **class number** = number of conjugacy classes in $G$.

7. **The first level of abstraction: the mutiplication table of $G$**: In a basic way, the multiplication table alone completely defines $G$, though we must of course inspect the table to root out the interesting properties of $G$. In this abstract point of view, we forget all concrete representations such as isometries, permutations, matrices, etc. and think merely of $|G|$ symbols combined according to the table.

| | **1** | $\mathbf{s_1}$ | $\mathbf{s_2}$ | $\mathbf{r_1}$ | $\mathbf{r_2}$ | $\mathbf{r_3}$ |
|---|---|---|---|---|---|---|
| **1** | **1** | $\mathbf{s_1}$ | $\mathbf{s_2}$ | $\mathbf{r_1}$ | $\mathbf{r_2}$ | $\mathbf{r_3}$ |
| $\mathbf{s_1}$ | $\mathbf{s_1}$ | $\mathbf{s_2}$ | **1** | $\mathbf{r_3}$ | $\mathbf{r_1}$ | $\mathbf{r_2}$ |
| $\mathbf{s_2}$ | $\mathbf{s_2}$ | **1** | $\mathbf{s_1}$ | $\mathbf{r_2}$ | $\mathbf{r_3}$ | $\mathbf{r_1}$ |
| $\mathbf{r_1}$ | $\mathbf{r_1}$ | $\mathbf{r_2}$ | $\mathbf{r_3}$ | **1** | $\mathbf{s_1}$ | $\mathbf{s_2}$ |
| $\mathbf{r_2}$ | $\mathbf{r_2}$ | $\mathbf{r_3}$ | $\mathbf{r_1}$ | $\mathbf{s_2}$ | **1** | $\mathbf{s_1}$ |
| $\mathbf{r_3}$ | $\mathbf{r_3}$ | $\mathbf{r_1}$ | $\mathbf{r_2}$ | $\mathbf{s_1}$ | $\mathbf{s_2}$ | **1** |

8. **The second and universal level of abstraction: a presentation for $G$.** Intuitively, a *presentation* for a group $G$ is a 'concise' summary of the multiplication table, basically a minimal amount of information which would suffice to reconstruct the whole table. Note that this means that

   - we should be able to reconstruct all elements of the group; and
   - we should be able to say how all elements multiply.

   Now let's be more precise. What we require in a presentation is

   (a) a (preferably small) set of **generators** $a, b, c, \ldots$ for the group $G$. This means that *every* element $g \in G$ is a product of these generators or their inverses, allowing repeats. Such a product is often called a **word** in the generators. Examples are $a, aa^{-1}, abaaab^{-1}b^{-1}cc$ etc. Of course, these can sometimes be simplified using the basic laws of exponents valid for *all* groups:

   $$aa^{-1} = 1 \ , \ abaaab^{-1}b^{-1}cc = aba^3b^{-2}c^2 \quad .$$

   But there could well be other simplifications possible due to special features of the group $G$ in question. These peculiarities are given by

   (b) a set of **relations** (a.k.a. relators) satisfied by the given generators and from which all valid relations in $G$ follow by algebraic manipulations in the group. This is a little hard to define more precisely, so here we will just sketch a few examples and state the key theorems.

9. **Example**. Suppose in the calculation just above, we do know that $ab = ba$, which can be rewritten as $aba^{-1}b^{-1} = 1$. Then we achieve a further simplification:

$$abaaab^{-1}b^{-1}cc = a^4b^{-1}c^2 \quad .$$

10. **Example**. Suppose $G$ is generated by *two* elements $a, b$ which satisfy the relations

$$a^2 = b^2 = (ab)^3 = 1 \qquad (**)$$

Various different groups have these generators and satisfy the relations!!

(a) $a = b = 1$ (say the integer 1); so $G = \{1\}$ has order 1.

(b) $a = b = -1$ (again integers ). Check that the relations (**) are satisfied. What now is the order of $G$?

(c) Another possibility using ordinary integers? $a = 1$ and $b = -1$. Are all the relations (**) above satisfied?

(d) Now try the symmetry group of the equilateral triangle above. Let $a =$ ? and $b =$ ? be carefully chosen symmetries. Do they generate the full symmetry group? Do they satisfy the relations (**)?
Hint: your choices for $a$ and $b$ will be closely guided by the relations to be satisfied.

(e) Thus the order of $G$ could be as big as 6. Could it be larger still? Try to compute the possibilities!! Take all possible combinations of $a, b, a^{-1}, b^{-1}$, subject to the relations (**), and determine how many truly different elements you can get. For example, $a^2 = 1$ implies $a^2a^{-1} = 1a^{-1}$, so that $a = a^{-1}$. In short, *in this example*, negative poweres of the generators are unnecessary, and at the outset, we can restrict only to positive integral exponents.

(f) In fact, there is a *largest such group* satisfying (**) !! And its order is __________

**Remark**: the peculiar structure of the relations in (**) means that the symmetry group of the equilateral triangle is the *Coxeter group* of type $A_2$.

11. **Theorem 10.1**

Consider all groups generated by generators

$$a, b, c \ldots$$

satisfying specified relations

$$w_1 = w_2 = \ldots = 1$$

(namely certain special words in the generators).

Then there exists a 'largest' such group, denoted

$$G = \langle a, b, c \ldots \mid w_1 = w_2 = \ldots = 1 \rangle$$

(This is called a *presentation* for the group $G$.)

More precisely, if $H$ is any other group with corresponding generators $\tilde{a}, \tilde{b}, \tilde{c} \ldots$ satisfying the corresponding relations $\tilde{w}_1 = \tilde{w}_2 = \ldots = \tilde{1}$, then there exists a unique homomorphism

$$\varphi : G \to H$$

which explicitly sends $a$ to $\tilde{a}$, $b$ to $\tilde{b}$, etc.

12. **Remarks**.

(a) This is a very powerful theorem. For example, it says that we can construct groups at will, choosing random symbols for generators, random equations for relations. Of course, the resulting groups could be trivial (order 1), could be infinite, could be uninteresting.

(b) Recall that $H \simeq G/\ker\varphi$. Hence,

$$|G| = |H|\,|\ker\varphi|\ .$$

Since $|H|$ divides $|G|$, we do indeed find that $|G| \geq |H|$. In this sense, $G$ is the largest group satisfying the relations. (It could be infinite.)

(c) It is a nice exercise to use the theorem to prove that $G$ is uniquely defined up to isomorphism.

13. **Exercises on presentations**. Compute the orders of these groups and describe each in more familiar terms (e.g. symmetry group of equilateral triangle).

    (a) $G = \langle a \,|\, a^2 = 1 \rangle$

    (b) $G = \langle b \,|\, b^3 = 1 \rangle$

    (c) $G = \langle a, b \,|\, a^2 = b^2 = (ab)^4 = 1 \rangle$

    (d) $G = \langle a, b \,|\, a^2 = b^4 = aba^{-1}b^{-1} = 1 \rangle$

    (e) $G = \langle a, b, c \,|\, a^2 = b^2 = c^2 = (ab)^3 = (bc)^3 = (ac)^2 = 1 \rangle$

    (f) $G = \langle a, b \,|\, a^2 = b^2 = (ab)^2 \rangle$

    Warning: we aren't saying $a^2 = 1$ here; rather, the relations are just clean ways of writing

    $$a^2 b^{-2} = a^2 (ab)^{-2} = 1 \ .$$

    It is still possible, for example, that $a$ has infinite period!!

    (g) $G = \langle a, b \,|\, a^2 = b^2 = 1 \rangle$

    (h) $G = \langle a, b \,|\, a^3 = b^3 = (ab)^3 = 1 \rangle$

# 11 Conjugacy and Characters

Here are several exercises to bolster your understanding of conjugacy and characters. Try to construct your own proofs before consulting a standard text.

Usually

**$G$ will be a general finite group, its order denoted by $|G|$. Its identity will usually be $e$.**

1. Recall that $b$ **is conjugate to** $c$ in a group $G$ if $b = g^{-1}cg$ for some $g \in G$. Let us indicate this by

$$b \sim c$$

**Proposition 11.1**

$\sim$ is an equivalence relation.

**Proof**. See any group theory text. □

2. Thus each element $b \in G$ belongs to a unique equivalence class, called naturally a **conjugacy class**. Let's denote this class by $\mathrm{Cl}(b)$. Thus, $|\mathrm{Cl}(b)|$ is the number of elements of $G$ which are conjugate to $b$ (including $b$ itself, of course).

3. **Remarks for Discussion**: when $G$ is a geometrical group, the conjugacy classes correspond to 'geometrically distinct' kinds of isometries. For example, for an ordinary square the rotational symmetries and the reflections lie in different conjugacy classes. In fact, the reflections themselves split into two distinct conjugacy classes. What are they?

In the full permutation group $\mathbb{S}_n$, the conjugacy classes correspond to the essentially different ways of writing $n$ as a sum of positive integers. These ways of writing $n$ are called *partitions* of $n$. The study of partions is a deep area of number theory.

**Example**. $n = 5$ can be partitioned in these 7 ways:

$$\begin{aligned} 5 &= 5 \\ &= 4+1 \\ &= 3+2 \\ &= 3+1+1 \\ &= 2+2+1 \\ &= 2+1+1+1 \\ &= 1+1+1+1+1 \end{aligned}$$

Thus $\mathbb{S}_5$ has 7 conjugacy classes; here are typical representatives of these classes, corresponding in order to the above partitions:

$$(1,2,3,4,5),\ (1,2,3,4)\ ,(1,2,3)(4,5),\ (1,2,3),\ (1,2)(3,4),\ (1,2),()\ \ .$$

4. Suppose a group $G$ has $k$ conjugacy classes. Choose at random an element $b_j$ in the $j$th class.

   **Exercises**.

   (a) One such class representative is forced. Which is it and how big is that conjugacy class?

   (b) Simplify

   $$\sum_{j=1}^{k} |\mathrm{Cl}(b_j)| = \underline{\hspace{3cm}}$$

5. Fix an element $b \in G$. Then the **centralizer** of $b$ in $G$ is the set of all elements of $G$ which commute with $b$:

   $$C(b) := \{g \in G \,|\, gb = bg\} \ \ .$$

   **Theorem 11.1**

   $C(b)$ is a subgroup of $G$.

   **Proof**. Supply details concerning identity $e$, closure under inverse, products.

   □

6. **Exercises**.

   (a) What is $C(e)$ ?

   (b) Show that always $b \in C(b)$. Ditto for $b^{-1}$, in fact for any $b^n$, where $n \in \mathbb{Z}$.

7. **Theorem 11.2**

   Suppose $b = g^{-1}cg$ (so that $b \sim c$). Then

   $$C(b) = g^{-1}\, C(c)\, g \ \ .$$

   **Remark**: thus conjugate elements have conjugate centralizers. Two such groups must be isomorphic, and so have the same size.

   **Proof**. See any text on group theory. □

Now for the really neat theorem!!

8. **Theorem 11.3**

The number of conjugates of $b$ in $G$ equals the index of the centralizer of $B$:

$$|\mathrm{Cl}(b)| = [G : C(b)] \ .$$

Hence, the number of elements in a conjugacy class divides the order of the group.

**Proof**. Fix $b \in G$. Put the right cosets of $C(b)$ into a set

$$RC := \{ C(b)g : g \in G \}$$

Remember that in a set we count only distinct elements. Thus

$$|RC| = \frac{|G|}{|C(b)|} = t \ \text{(say)}.$$

It is not necessary here, but one could choose explicit coset representatives $g_1, \ldots, g_t$, so that the different cosets in $RC$ would then be $C(b)g_1, \ldots, C(b)g_t$.

Define

$$\begin{aligned} \varphi : \mathrm{Cl}(b) &\quad \to \quad RC \\ g^{-1}bg &\ \mapsto C(b)g \end{aligned}$$

Supply details that $\varphi$ is well-defined, onto and 1–1. □

9. **Exercises**.

(a) Let $G = \mathbb{S}_4$, with standard generators $r_1 = (1\,2), r_2 = (2\,3), r_3 = (3\,4)$. Fill in the data in the following table:

| Class Rep. | Factorization of $b$ | Size | $\lvert G\rvert$ | Explicit list of conjugates |
|---|---|---|---|---|
| $b$ | in terms of the $r_j$'s | $C(b)$ | $\lvert C(b)\rvert$ | |
| () | | | | |
| (1 2) | | | | |
| (1 2)(3 4) | | | | |
| (1 2 3) | | | | |
| (1 2 3 4) | | | | |

(b) Do the same thing for $\mathbb{S}_5$, say with generators $r_1 = (1\,2), r_2 = (2\,3), r_3 = (3\,4)$ and $r_4 = (4\,5)$.

You needn't however explicitly list the conjugates.

(c) How many conjugacy classes does $\mathbb{S}_6$ have?

(d) Let $G$ be the symmetry group of a cube. What is $|G|$? How many conjugacy classes does $G$ have and what are their sizes?

# 12 The Ring of Integers

1. The set $\mathbb{Z}$ of integers comes with two *closed* operations. Any two integers $a, b$ have

   - a *sum* $a + b$ which is also an integer; and
   - a *product* $ab$ which is also an integer.

   There are special integers for each operation: 0 is the *additive identity* and 1 is the *multiplicative identity*. Moreover, every integer $a$ has an *additive inverse* (or *negative*) $-a$.

   Since the familiar rules of arithmetic hold for addition (and its close cousin subtraction) and multiplication, we say that $\mathbb{Z}$ is a *ring*.

   However, not every integer has a multiplicative inverse (reciprocal) which is itself an integer; for example, 3 is an integer but $\frac{1}{3}$ is not. Because of this, $\mathbb{Z}$ is not a *field*.

   Another way to phrase this is to say that division of integers is not a closed operation. That deficiency is, however, easily repaired by expanding the integers to the rational numbers $\mathbb{Q}$, which do form a field.

   Not every ring admits such a repair job. An example which we will soon encounter is the residue class ring

   $$\mathbb{Z}_6 = \{0, 1, 2, 3, 4, 5\} ,$$

   equipped with modular addition, subtraction and multiplication. Recall that the convenient symbols $0, 1, \ldots, 5$ are no longer integers; instead they denote residue classes. Thus $3 + 4 = 2$, and $2 \cdot 3 = 0$, even though neither 2 nor 3 equals 0. We say that 2 and 3 are *zero-divisors* in $\mathbb{Z}_6$.

2. For the record, here is a definition for *commutative rings*. Recall that a closed operation on a set $R$ is the usual sort of operation which returns a result in the *same* set. Thus, the dot product on the set $\mathbb{R}^3$ is not closed, since $\mathbf{u} \cdot \mathbf{v}$ is a scalar, not a vector in $\mathbb{R}^3$.

   **Definition 12.1** *A* commutative ring *is a set $R$ equipped with two closed operations on $R$, typically called addition $a + b$ and multiplication $ab$, satisfying these familar properties:*

   **Concerning addition**: *For all $a, b, c \in R$,*

   *(a) (associative law)* $(a + b) + c = a + (b + c)$

   *(b) (commutative law)* $a + b = b + a$

   *(c) (zero) there is a special element* 0 *such that* $0 + a = a$

   *(d) (negatives) each $a$ has its own special negative, denoted $-a$, such that*

   $$a + (-a) = 0$$

   **Concerning multiplication**: *For all $a, b, c \in R$,*

   *(e) (associative law)* $(ab)c = a(bc)$

   *(f) (commutative law)* $ab = ba$

   *(g) (identity) there is a special element* 1 *such that* $1a = a$

**Linking the operations**: *For all $a, b, c \in R$,*

*(h) (distributive law)* $a(b + c) = ab + ac$

**Remarks**. There are many more familiar, and not so familar, algebraic properties; but these must be proved from the above axioms. For example, it can (and must!) be proved that

$$(-1)a = -a, \text{ and } 0a = 0, \text{ for all } a \in R.$$

As a rule, your basic algebraic instincts will carry you forward. Naturally one can learn caution only after experiencing some unpleasant surprises!

*Subtraction* is a subsidiary operation, defined in terms of what we already have. By definition,

$$b - a := b + (-a) .$$

3. There is a huge variety of different kinds of rings, quite unlike the integers $\mathbb{Z}$. Of course, the rationals $\mathbb{Q}$, the reals $\mathbb{R}$ and the complex numbers $\mathbb{C}$ are also rings. But they have the particular virtue of being fields, in which division is possible:

**Definition 12.2** *A* field *is a commutative ring $R$ in which every* non-zero *element $a$ has a multiplicative inverse, denote $1/a$ or $a^{-1}$, and satisfying*

$$a(1/a) = 1 .$$

Thus division is possible in a field. It, too, is a subsidiary operation. By definition, for $a \neq 0$,

$$\frac{b}{a} := b(1/a) .$$

4. **Aside**. Strictly speaking we have described a commutative ring with unit element 1. Some rings do not have a 1. There are lots of other ways of varying the requirements to produce new and interesting kinds of objects.

5. Even though most integers do not have reciprocals, it is still very interesting to study divisibility, factorization and the like. Indeed, that is a central purpose of number theory.

**Definition 12.3** *Suppose $a$ and $b$ are integers. We say $a$* divides *$b$ and write*

$$a|b$$

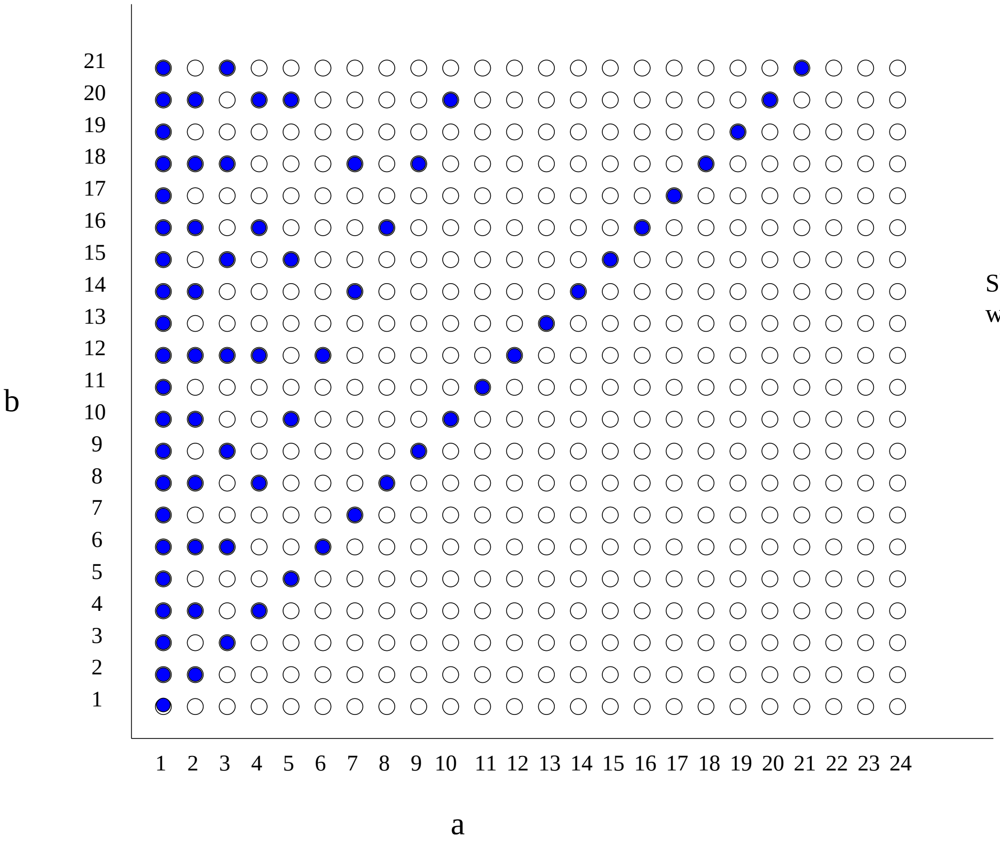
*if $b = az$, for some integer $z$. Of course, we also say that $a$ is a* factor *of $b$ and that $b$ is a* multiple *of $a$, etc.*

When discussing factors and divisibility, multiplication by the integer *unit* $-1$ does no harm:

$$b = (-1)a \cdot (-1)z = (-a)(-z) .$$

Because of this we can focus on the positive integers, or natural numbers $\mathbb{N} = \{1, 2, 3, 4, \ldots\}$.

6. Indeed, we can graph the *divisibility lattice* on $\mathbb{N}$. To do this we lay out $\mathbb{N}^2$, the set of all ordered pairs of natural numbers $(a, b)$, but shade in only those positions in which $a|b$:



Shade in
when a divides b

Thus divisibility, which is a relation between certain natural numbers (or more generally integers), can be viewed in the language of set theory as a subset

$$\mathcal{R} \subset \mathbb{N} \times \mathbb{N}$$

We now have a precise way of defining just what a relation 'really is'. Indeed, 'functions' are a special case of this. However, in practice we merely take comfort in the fact that all this has been done, and (as humans) think more intuitively.

It is quite another matter to decide how a computer should deal with relations and functions!

7. Have a look at the numbers to the left of rows which contain exactly two shaded dots. We have found the prime numbers

$$\mathbb{P} = \{2, 3, 5, 7, 11, 13, 17, 19, 23, \ldots\} \tag{3}$$

**Definition 12.4** *A positive integer $p$ is* prime *if it has exactly two positive divisors,* namely, 1 and $p$ itself.

8. Our diagram above is really a variant of the famous *Sieve of Erathosthenes* (ca. 230 BCE). Here we sieve out the primes from the set $\{2, \ldots, 24\}$:

   $$\boxed{2}, \boxed{3}, \cancel{4}, \boxed{5}, \cancel{6}, \boxed{7}\ \cancel{8}, \cancel{9}, \cancel{1}0, \boxed{11}, \cancel{1}2, \boxed{13}, \cancel{1}4, \cancel{1}5, \cancel{1}6, \boxed{17}, \cancel{1}8, \boxed{19}, \cancel{2}0, \cancel{2}1, \cancel{2}2, \boxed{23}, \cancel{2}4$$

   Clearly this is the sort of thing a computer is made to do; just how to make Gap perform the trick is another matter (see below).

   We can say a little about when the sieving must stop if we want to find the primes in $\{2, 3, 4, \ldots, n\}$, for some positive integer $n$. Recall that a positive integer $z$ is *composite* if it has two proper factors, say

   $$z = ab\ ,$$

   where $a > 1$ and $b > 1$. Thus the composite positive integers are $\{4, 6, 8, 9, 10, \ldots\}$. We can easily prove that

   **Proposition 12.1** *A composite integer $z > 1$ must have a prime factor $p \leq \sqrt{z}$.*

   Meaning. Sieving will stop when we reach $\lfloor\sqrt{n}\rfloor$. Nevertheless, when $n$ becomes quite large, sieving becomes 'computationally slow'. You can test this using Gap's version of the Sieve:

```
sieve:=function(n) local numbers,p,primes,m;
numbers:=[2..n];# this is the range of all integers from 2 to n
# the programme sieves out the primes from this range
# [ 2 .. n ]
primes:=[]; # we must have an empty box in which we can put things!
#
 for p in numbers do
Add(primes,p);
for m in numbers do
if m mod p = 0 then
Unbind(numbers[m-1]);
fi;
od;
od;
Print("In the interval [2..",n,"] there are ",Size(primes)," primes.","\n");
Print("They are ","\n");
return primes;
end; # end of function sieve
```

   This function was adapted from the Gap tutorial at

   `http://www.gap-system.org/Manuals/doc/htm/tut/CHAP003.htm#SECT005`

9. **How many primes are there?**

It could be that all integers past a point are sifted out, i.e. that there are no primes left over past a certain, presumably big, positive integer $B$. It could be; but in fact Euclid (300 BCE) showed ‘No, that’s not the way it is!’

For if there were no primes after the big integer $B$, then there would be only finitely many primes, say $L$ of them. We could write these in order as follows:

$$p_1 = 2,\ p_2 = 3,\ p_3 = 5,\ p_4 = 7,\ p_5 = 11,\ \ldots, p_L \ ,$$

where $p_L$ is the last and biggest prime. Since there are only finitely many such primes, we can (as Euclid observed) multiply them out and add 1 to get the integer

$$a = p_1 p_2 \cdots p_L + 1 \ .$$

Yes, $a$ is presumably huge; but still it is finite and we can think about it. Now every integer, including $a$, can be factored into primes with enough patience (see Theorem 12.2 below). But each prime factor of $a$ must appear on our exhaustive list, so at least we have some $p_j$ dividing $a$. Thus for some integer $q$ we get $a = qp_j$ so that

$$1 + (p_1 \cdots p_{j-1} p_j p_{j+1} \cdots p_L) = a = p_j q \ ,$$

and thus

$$p_j(q - p_1 \cdots p_{j-1} p_{j+1} \cdots p_L) = 1 \ .$$

The term in brackets is some integer, so that $p_j$ is a factor of 1. But this means $p_j = \pm 1$, a contradiction.

**Theorem 12.1 Euclid’s Theorem on the infinity of primes**. *There are infinitely many prime numbers.*

10. It is easy to believe, and not so hard to prove by induction, that every integer $n \geq 2$ can be factored into primes. In fact, up to a reordering of the terms, there is only one way to do this for a given integer:

**Theorem 12.2 The Fundamental Theorem of Arithmetic**. *Every integer $n \geq 2$ can be uniquely written as a product of primes, with the prime factors written in non-decreasing order.*

We are familiar with this. Of course, we usually group repeated primes using exponential notation:

$$2232 = 2 \cdot 2 \cdot 2 \cdot 3 \cdot 3 \cdot 31 = 2^3 3^2 31 \ .$$

Here is a relevant Gap session:

```
gap> FactorsInt(2232);
[ 2, 2, 2, 3, 3, 31 ]
gap> # For more easy reading try:
gap> PrintFactorsInt(2232);Print("\n");
2^3*3^2*31
```

11. It is easy to manufacture a Gap routine which uses Euclid's idea to generate primes. However, the method does not generate all primes and produces what primes it does in a seemingly 'random' order. Moreover, this method is excruciatingly slow. The reason for this is that here we find primes by *factoring* composite integers (like $a$ in the proof of Theorem 12.1 above); but as far as we know, factoring is 'computationally very hard'. Indeed, this supposition is used to justify the supposed security of computer communications systems, including that used between an ATM and your bank!

    Anyway, here is a Gap version of Euclid's idea:

```
gap> # We start with the first prime we know and put it in a list:
gap> pr:=[2];
[ 2 ]
gap> # Now we produce several primes using Euclid's idea.
gap> for j in [1..10] do x:=Factors(Product(pr)+1)[1];Append(pr,[x]);od;
#I  IsPrimeInt: probably prime, but not proven: 38709183810571
#I  IsPrimeInt: probably prime, but not proven: 420743244646304724409
gap> pr;
[ 2, 3, 7, 43, 13, 53, 5, 6221671, 38709183810571, 139, 2801 ]
```

    What we are doing is selecting out the smallest prime factor of $1 + p_1 \cdots p_k$ as the newest prime $p_{k+1}$.

    In fact, Gap must be using some probabalistic argument to assess whether large integers like 38709183810571 are prime. Presumably we can bypass that uncertainty; but then we might send the computer off for hours of work!

12. Here is a horribly slow (but fail-safe!) routine to test whether a positive integer $n$ is prime. It checks to see whether each number $b$ from 2 to $\lfloor\sqrt{n}\rfloor$ is a factor of $n$. That is easy; just use long division (see below) and check whether the remainder is 0.

    So repeat this for each $b \in [2..\lfloor\sqrt{n}\rfloor]$. If we ever get remainder 0, then $n$ is composite. Otherwise, $n$ is prime:

```
crudefact:=function(n) local b;
for b in [2..RootInt(n)]  # RootInt(n) is the square root
             # of n (rounded down if necessary). Also,
             # n mod b is the remainder when n is divided by b
do
if (n mod b) = 0 then Print("The integer ",n," is composite.","\n");return;
else fi;
od;
Print("The integer ",n," is prime.","\n");
end; # end of function crudefact:
```

    Even this tends to work well for a 10 or 15 digit number $n$, since a 'randomly chosen' $n$ of that magnitude will not be prime. However, my computer will start to struggle to make a determination, if by bad luck $n$ does happen to be a prime of that magnitude.

# 13 Long division and modular arithmetic

1. We will reconsider how it is that we 'divide' an arbitrary integer $z$ by a particular positive integer $d$. To illustrate matters we take $d = 4$; but it will be clear that our reasoning works for

$$\text{any particular integer } d \geq 1.$$

2. Now lay out the integers in $d$ columns, starting with 0 on the left, and returning to the first column after $d$ steps. Thus we start by writing

$$0, 1, \ldots (d-1)$$

in what I'll call the *remainder row*; then we continue the display in a natural way. Here is what we get when $d = 4$:

$$\begin{array}{rrrr}
\vdots & \vdots & \vdots & \vdots \\
-12 & -11 & -10 & -9 \\
-8 & -7 & -6 & -5 \\
-4 & -3 & -2 & -1 \\
0 & 1 & 2 & 3 \\
4 & 5 & 6 & 7 \\
8 & 9 & 10 & 11 \\
12 & 13 & 14 & 15 \\
\vdots & \vdots & \vdots & \vdots
\end{array} \tag{4}$$

Notice that the displayed integers increase by 1 for each step to the right or down, 'wrapping around' when we hit the end of a row. (If you are reading this online, the remainder row appears in red.)

We will label the columns by the integer $r$ occurring in the remainder row. Thus the left-most column corresponding to $r = 0$ contains all the multiples of $d$.

We will label rows by the integer $q$, taking $q = 0$ for the remainder row, with $q$ increasing by 1 for each row-step down, hence, of course, decreasing by 1 for each row-step up. In other words, the row labelled $q$ begins with the integer $qd$.

This means that $q$ is the 'number of $d$'s' found in any integer $z$ in that row. For example, we can take $q = 3$ copies of $d = 4$ out of $z = 13$. How much is left over? Clearly the remainder $r = 1$ is the entry to be found where the remainder row crosses the column containing $z = 13$. In brief,

$$13 = 3 \cdot 4 + 1 \; .$$

We have the essence of long division!

Likewise, $12 = 3 \cdot 4 + 0$ and $15 = 3 \cdot 4 + 3$. Returning to the remainder row, we see that $2 = 0 \cdot 4 + 2$, indicating that $d = 4$ does not go into $z = 2$ and merely leaves the remainder 2.

Let us continue upward into the negative integers, taking care to continue the pattern. Here we must be careful. For example, we find $z = -10$ in the row beginning $-12 = (-3) \cdot 4$. Thus we should agree that 4 goes into $-10$ a total of $q = -3$ times, with remainder $r = 2$:

$$-10 = (-3) \cdot 4 + 2 \ .$$

That way all integers behave uniformly and give a non-negative remainder from 0 to $d - 1$. This is a very useful convention.

Since every integer $z$ fits into such an array in exactly one position, we have more or less proved

**Theorem 13.1 The division algorithm** *Suppose the integer $d \geq 1$. Then for each integer $z$ there exist a unique* quotient *$q$ and* remainder *$r$ such that*

$$z = qd + r, \quad \textit{where} \ \ 0 \leq r \leq d - 1 \ .$$

3. The grade school algorithm for long division implements this process in a familiar and efficient way. The corresponding Gap commands are illustrated here:

```
gap> z:=-10; d:=4;
-10
4
gap> # for the remainder r we use the command
gap> #    z mod d
gap> r:=z mod d;
2
gap> # Thus the quotient must be
gap> q:=(z-r)/d;
-3
gap> z = q*d+r;
true
```

There are related special Gap commands – QuoInt($z$,$d$) and RemInt($z$,$d$) – which unfortunately do not give the quotient and remainder our way when $z$ is negative. These peculiar functions are there for use in other routines; for simplicity, we avoid using them.

4. It is instructive, but less efficient, to programme Gap to do long division by mimicking our reasoning for the infinite array (4) above. Here is one attempt:

   ```
   # Goal: model elementary division of integers
   # as an illustration of recursive programming.
   # The first function ‘divplus’ handles positive integers only.
   # We divide a positive integer n by a positive integer d.
   divplus:=function(n,d) local q,r;
   if d<0 then Print("bad divisor d = ",d,"\n");
   else
   q:=0;r:=n;
   if r<d then return [q,r];
   else return [1,0]+divplus(n-d,d);# this amounts to stepping up
            #one row in the integer grid. Remainder is unchanged;
            # but quotient is decreased by 1.
   fi;
   fi;
   end;# end function divplus
   #
   # The next function for general integers calls the previous function.
   divint:=function(n,d) local a,b,c;
   if d<=0 then Print("bad divisor d = ",d,"\n");
   else if n<0 then a:= divplus(-n,d);
   if a[2]=0 then return -a;
   else return (-a)+[-1,d];fi;
   #
   else return divplus(n,d);
   fi;
   fi;
   end;#end function divint
   #
   ```

   These implementations will stuggle for big inputs because of the recursion in divplus. The are for understanding not efficiency. Another deficiency is that the routines do not make essential use of decimal notation for integers.

5. Let's study more carefully the integer array in (4). Every integer appears in exactly one of the $d$ columns (in this case $d = 4$). This gives $d = 4$ subsets which partition $\mathbb{Z}$.

   **Definition 13.1** *A* partition *of a set $A$ is a collection of non-empty, mutually disjoint subsets $B_j$ whose union is $A$.* (The subscript $j$ runs over some index set $J$.)

   In (4), the set $A$ is the integers $\mathbb{Z}$, and there are four subsets in the partition of interest to us. Just assemble each column into its own subset:

$$\begin{aligned} B_0 &= \{\ldots, -8, -4, 0, 4, 8, \ldots\} \\ B_1 &= \{\ldots, -7, -3, 1, 5, 9, \ldots\} \\ B_2 &= \{\ldots, -6, -2, 2, 6, 10, \ldots\} \\ B_3 &= \{\ldots, -5, -1, 3, 7, 11, \ldots\} \end{aligned}$$

   **Remark**. Thus the index set $J = \{0, 1, 2, 3\}$ is finite, and in fact corresponds exactly to the remainder row. The sets $B_r$ are called *residue classes*; one often writes $\bar{r}$ as a convenient abbreviation for $B_r$.

   In other applications there could be infinitely many blocks $B_j$ in the partition, so that $J$ would be infinite.

   But for divisor $d$, it is clear that the $d$ columns give exactly $d$ residue classes $B_r$, one for each of the possible entries $r$ in the remainder row

$$0, \ldots, d-1 .$$

6. What do the elements in class $B_r$ have in common? Two answers are immediate from the picture in (4):

   **Theorem 13.2** *Two integers $z$ and $w$ lie in the same column*

   - *if and only if $z$ and $w$ have the same remainder $r$ upon division by $d$ (and $r$ then labels their column);*
   - *if and only if their difference $z - w$ is a multiple of $d$:*

$$z - w = qd .$$

   **Definition 13.2** *Suppose $d$ is a positive integer and $z, w \in \mathbb{Z}$. We write*

$$z \equiv w \pmod{d}$$

   *if $z - w$ is a multiple of $d$ (equivalently, $z$ and $w$ have the same remainder upon division by $d$).*

   We often read $z \equiv w \pmod{d}$ as '$z$ is congruent to $w$ modulo $d$'.

7. It is remarkable and very useful that the symbol $\equiv$ behaves rather like the ordinary equality symbol '='. More precisely, it has the three natural looking properties described in the next

   **Theorem 13.3** *For each fixed $d \geq 1$, the relation '$\equiv \pmod{d}$' is an* equivalence relation *on $\mathbb{Z}$. That is, for all $z, w, v \in \mathbb{Z}$, we have*

   *(a)* the reflexive property*: $z \equiv z \pmod{d}$.*

   *(b)* the symmetric property*: If $z \equiv w \pmod{d}$, then $w \equiv z \pmod{d}$.*

   *(c)* the transitive property*: If $z \equiv w \pmod{d}$ and $w \equiv v \pmod{d}$, then $z \equiv v \pmod{d}$.*

   **Proof**. Try it yourself! In part (c), for example, we assume $z - w = qd$ and $w - v = yd$ for certain unknown integers $q$ and $y$. But then add these equations to get $z - v = (q + y)d$. Done! (Really all we are saying is that if $z$ and $w$ lie in the same column in (4) and also $w$ and $v$ lie in the same column, then $z$ and $v$ lie in the same column. □

   **Remark**. Intuitively, this means you can play as fast and loose with '$\equiv \pmod{d}$' as with '='.

   At a more abstract level, we begin to see that any partitioning of a set $A$ lets us think of the blocks in terms of a new variant of 'equality'.

8. There is much more to the algebra of congruences. The interactions between congruences and the ring operations on $\mathbb{Z}$ are most useful tools in number theory and algebra. The following results are easily proved:

   **Theorem 13.4** *For all $z, w, a, b \in \mathbb{Z}$ (integers!), and a fixed modulus $d \geq 1$, we have:*

   *(a) If $z \equiv w \pmod{d}$ and $a \equiv b \pmod{d}$, then $a + z \equiv b + w \pmod{d}$. Likewise, $a - z \equiv b - w \pmod{d}$.*

   *(b) If $z \equiv w \pmod{d}$ and $a \equiv b \pmod{d}$, then $az \equiv bw \pmod{d}$. In particular,*

   *(c) If $z \equiv w \pmod{d}$, then $az \equiv aw \pmod{d}$.*

   We should not be surprised, however, that we cannot 'divide' one congruence by another, at least without a more careful look at what that would entail. For a different take on much the same issues, consult Theorems 13.5 and 13.6 below.

9. **The idea of modular arithmetic a.k.a. clock arithmetic**.

   (a) If it is Wednesday today, then 300 days from now it will be Tuesday, since

   $$300 = 42(7) + 6 .$$

   Likewise, 300 days ago it must have been Thursday, since

   $$-300 = (-43)(7) + 1 .$$

For many mathematical purposes, what counts for an integer $z$ is its remainder (or residue) after division by a fixed positive modulus $d$. (For week-work, $d = 7$.)

We can define a new arithmetic on the standard residues by replacing the ordinary sum and product of integers by the remainder modulo $d$. Basically this is clock arithmetic.

(b) **Definition 13.3** *The residue class ring $\mathbb{Z}_d$.*

*For a fixed integer modulus $d \geq 1$, we denote the set of standard remainders by*

$$\mathbb{Z}_d := \{0, \ldots, d-1\} .$$

*On this set we define two closed operations, temporarily denoted $\oplus$ and $\odot$. For $a, b \in \mathbb{Z}_d$ let*

- *$a \oplus b := r$, if as ordinary integers $a + b = qd + r$, with remainder $r$ satisfying $0 \leq r \leq d-1$ as usual.*
- *$a \odot b := r$, if as ordinary integers $a \cdot b = qd + r$, with remainder $r$ satisfying $0 \leq r \leq d-1$ as usual.*
- *$-a := r$, if as ordinary integers $-a = qd + r$, with remainder $r$ satisfying $0 \leq r \leq d-1$ as usual.*

These operations are closed, since by definition each always produces a result back in the set $\mathbb{Z}_d := \{0, \ldots, d-1\}$.

(c) In brief, then, we operate on $\mathbb{Z}_d$ by replacing, wherever possible, any integer $z$ by its remainder upon division by $d$. This remainder is often called the *residue* of $z$ (mod $d$).

It is not too hard to prove that under these new operations we still have the normal rules of arithmetic, as detailed in Definition 12.1. In short, we have

**Theorem 13.5** *$\mathbb{Z}_d$ with operations $\oplus$ and $\odot$ is a commutative ring.*

**Proof**. We must check each property in Definition 12.1, which is easy. As a test case, let's prove that multiplication is associative.

For any and all $a, b, c \in \mathbb{Z}_d$ our goal is to show $(a \odot b) \odot c = a \odot (b \odot c)$. We need to give integer names to the relevant factors.

First of all, $a \odot b = r$ means simply that $ab = qd + r$, where $0 \leq r \leq d-1$. Thus $(a \odot b) \odot c = r \odot c = s$, means $rc = q_1 d + s$, again with $0 \leq s \leq d-1$. Thus the ordinary integer

$$(ab)c = (qc + q_1)d + s = q_2 d + s ,$$

where $q_2 = qc + q_1$. But the division algorithm Theorem 13.1 produces a unique quotient $q_2$ and remainder $s$ regardless of the involved process by which got us to this result. Since $a(bc) = (ab)c$ as ordinary integers, the same sort of calculation has to give $a \odot (b \odot c) = s$, as well. In other words, we must always have

$$(a \odot b) \odot c = a \odot (b \odot c) .$$

The same sort of verification will work for all ring properties. □

(d) **Calculation in $\mathbb{Z}_d$ in practice**. We usually abuse notation and revert from the tiresome $a \oplus b, a \odot b$ to the usual $a + b, ab$.

For example, in $\mathbb{Z}_6$ we will simply write

$$1 + 3 = 4,\ 2 + 5 = 1,\ 2 \cdot 3 = 0,\ 4 \cdot 4 = 4\ , -1 = 5\ .$$

In this context it is useful to interpret '$=$' as '$\equiv \pmod 6$'. And strictly speaking the symbols $6, 7, -13$ make no sense, although we cannot go wrong by agreeing that

$$6 = 0,\ \ 7 = 1,\ \ -13 = -1 = 5\ .$$

It is important to note a built in deficiency in $\mathbb{Z}_6$. The crucial blemish is that 2, for instance, has no multiplicative inverse (what we would want to label 1/2) in $\mathbb{Z}_6$. For suppose some $a \in \mathbb{Z}_6$ did the job. This means $2 \cdot a = 1$. But then

$$3 = 3 \cdot 1 = 3 \cdot (2 \cdot a) = (3 \cdot 2) \cdot a = 0 \cdot a = 0\ ,$$

a contradiction! (3 and 0 really are different remainders upon division by 6.) Consulting Definition 12.2, we conclude that $\mathbb{Z}_6$ is not a field.

The core problem here is that the modulus $d = 6 = 3 \cdot 2$ is composite. Thus, the ring $\mathbb{Z}_d$ cannot be a field if $d$ is a composite integer. But what happens if $d$ is prime?

(e) **Theorem 13.6** *$\mathbb{Z}_d$ is a (finite) field if $d$ is prime.*

**Proof**. Fix any $a \neq 0$ in $\mathbb{Z}_d$. (We're supposing $d$ is a prime.) In other words, $1 \leq a \leq d - 1$. Now put the multiples of $a$ in the list

$$M = [0 = 0a, 1a, 2a, 3a, \ldots, (d-2)a, (d-1)a]\ ,$$

all taken $\pmod d$. Could two of these be equal? Well suppose $ia = ja$ for $0 \leq i \leq j \leq d - 1$. This means $(j - i)a = 0$ in $\mathbb{Z}_d$, which in turn means by definition that $(j - i)a$ has remainder 0 on division by $d$. In other words, $(j - i)a$ is a multiple of the <u>prime</u> number $d$. But we know (and will later prove in Theorem 14.2(b)) that a prime like $d$ has the special property that it must be a factor of either $j - i$ or of $a$, if it divides their product. But this is impossible since $a$ lies between 1 and $d - 1$, as does $j - i$, unless $i = j$.

We conclude that the elements listed in $M$ are distinct. But there are $d$ such elements, so we must simply have a rearrangement of $[0, 1, \ldots, d - 1]$, with 0 matching up in the first slot. Thus 1 must appear somewhere else in $M$! Hence there exists an integer $i \neq 0$ such that $i \cdot a = 1 \pmod d$. This $i$ will serve duty as $1/a$.

We now see why every $a \neq 0$ has a multiplicative inverse $1/a \in \mathbb{Z}_d$. We conclude that the ring $\mathbb{Z}_d$ is really a field. □

# 14 The greatest common divisor

1. Any two integers $a$ and $b$ have some common divisors, at least $+1$ and $-1$:

   $$a = a \cdot 1 \text{ and } b = b \cdot 1, \text{ so } 1|a \text{ and } 1|b \ .$$

   On the other hand, for <u>any</u> integer $n$, no matter how large, we have $n|0$ and $n|0$, so that 0 and 0 do not have a *greatest* common divisor. (Or perhaps we might call it $\infty$.)

   Finally, we observe that if say $b \neq 0$ and $k|b$, then $k \leq |b|$, so that there is a largest possible divisor for any non-zero integer, namely the positive integer $|b|$.

   These thoughts motivate

   **Definition 14.1** *Suppose $a, b \in \mathbb{Z}$ are not both 0. Then the* greatest common divisor *of $a, b$ is the largest integer $g$ which divides both $a$ and $b$. We write*

   $$g = \gcd(a, b) \ .$$

   Notice that $gcd(a, b)$ is then a positive integer, namely at least 1. The minimal case where we hit 1 is interesting and familiar:

   **Definition 14.2** *Two integers $a, b \in \mathbb{Z}$ are said to be* relatively prime *if* $\gcd(a, b) = 1$.

   For example, we say that a rational number $\frac{a}{b}$ is in *lowest terms* if $\gcd(a, b) = 1$. We often say, a little abusively, that $a$ and $b$ have 'no common factor', by which we really mean that $+1$ and $-1$ are the only common divisors of $a$ and $b$.

2. The *Euclidean Algorithm* lets us compute $g = \gcd(a, b)$ quickly and furthermore write $g$ as a $\mathbb{Z}$-linear combination of $a$ and $b$. This means we can find integers $u, v$ such that

   $$g = ua + vb \ .$$

   In fact, we then have

   $$\mathbb{Z}g = \mathbb{Z}a + \mathbb{Z}b \ .$$

   This is actually a statement concerning *ideals* in the ring $\mathbb{Z}$.

   For example,

   $$6 = \gcd(150, 114)$$

   and

   $$6 = (-3)150 + (4)114 \ .$$

3. To ease our proof, we need to understand some facts about divisibility:

**Proposition 14.1** *For integers $x, y, z$ we have*

*(a)* $\gcd(x, y) = \gcd(y, x)$.

*(b)* $\gcd(x, 0) = |x|$, *if* $x \neq 0$.

*(c)* $\gcd(x, y) = \gcd(x, y - zx)$, *for any integer* $z$.

**Proof**. Part (a) is clear, since when talking of *common* divisors, the order in which we consider $x$ or $y$ is irrelevant. We have already commented on part (b).

In part (c) we need only observe that the two numbers $x, y$ have the same set of common divisors as the two numbers $x, y - zx$. For if $x = q_1 d$ and $y = q_2 d$, so that $d$ is a common divisor of $x$ and $y$, then $y - zx = (q_2 - zq_1)d$. Since $q_2 - zq_1$ is some integer, $d$ is also a divisor of $y - zx$. The reasoning is reversible, since $y = (y - zx) - (-z)x$. □

4. Here then is the procedure. To make the recursive notation easier, we start off by conveniently relabelling $a$ and $b$ as $x_1$ and $y_1$. We furthermore initialize integer vectors $\mathbf{w}_{-1}$ and $\mathbf{w}_0$ whose two components will ultimately be transformed into the coefficients $u, v$ in

$$\gcd(a, b) = ua + vb \, .$$

You can omit the **w**s if you don't need the $\mathbb{Z}$-linear representation.

**Theorem 14.1 Euclidean Algorithm**.
*Suppose $a$ and $b$ are positive integers.*

- Initialize: let $x_1 := a,\ y_1 := b,\ \mathbf{w}_{-1} = [1, 0],\ \mathbf{w}_0 = [0, 1]$.
- Start: by dividing $x_1$ by $y_1$:

$$x_1 = q_1 y_1 + r_1, \text{ where } 0 \leq r_1 < y_1 \, .$$

  Let $\mathbf{w}_1 = \mathbf{w}_{-1} - q_1 \mathbf{w}_0$.

- Repeat as long as $r_j > 0$:

  Update: let $x_{j+1} := y_j$, $y_{j+1} := r_j$ and divide $x_{j+1}$ by $y_{j+1}$ to get

$$x_{j+1} = q_{j+1} y_{j+1} + r_{j+1}, \text{ where } 0 \leq r_{j+1} < r_j = y_{j+1} \, .$$

  Also let $\mathbf{w}_{j+1} = \mathbf{w}_{j-1} - q_{j+1} \mathbf{w}_j$.

- Stop: when the remainder hits 0, say at step $k + 1$, meaning $r_{k+1} = 0$.
- Return $\gcd(a, b) = r_k$ and $[u, v] = \mathbf{w}_k$ .

**Proof**. The sequence of non-negative remainders is decreasing:

$$0 \leq \ldots r_3 < r_2 < r_1 < y_1 = b \, .$$

Thus at some step, say $k + 1$, we must get remainder 0 and stop.

Now since $r_{k+1} = 0$ we have $x_{k+1} = q_{k+1}y_{k+1} + 0 = q_{k+1}y_{k+1}$, so that $r_k = y_{k+1}$ is a divisor of $x_{k+1}$. Hence $r_k = \gcd(x_{k+1}, y_{k+1})$.

But for each $j$, $x_{j+1} = y_j$ and $y_{j+1} = r_j = x_j - q_jy_j$. Now Theorem 14.1(c),(a) imply that

$$\gcd(x_{j+1}, y_{j+1}) = \gcd(y_j, x_j - q_jy_j) = \gcd(x_j, y_j) \ .$$

Tracing all the way back to the start we find that

$$r_k = \ldots = \gcd(x_1, y_1) = \gcd(a, b) \ .$$

Finally we must deal with the $\mathbf{w}_j$s. We start off with $\mathbf{w}_1 = [1, 0] - q_1[0, 1] = [1, -q_1]$, and note that $r_1 = x_1 - q_1y_1 = 1a - q_1b = \mathbf{w}_1 \cdot [a, b]$ (think dot product). This starts an induction in which we assume $r_j = \mathbf{w}_j \cdot [a, b]$. Then

$$\begin{array}{rcl}
\mathbf{w}_{j+1} \cdot [a, b] & = & (\mathbf{w}_{j-1} - q_{j+1}\mathbf{w}_j) \cdot [a, b] \text{ (by defn.)} \\
& = & \mathbf{w}_{j-1} \cdot [a, b] - q_{j+1}\ \mathbf{w}_j \cdot [a, b] \text{ (props. of } \cdot\text{)} \\
& = & r_{j-1} - q_{j+1}r_j \text{ (by inductive hypoth.)} \\
& = & y_j - q_{j+1}r_j \text{ (by defn.)} \\
& = & x_{j+1} - q_{j+1}y_{j+1} \text{ (by defn.)} \\
& = & r_{j+1}.
\end{array}$$

This completes the induction. (There is a small inductive lapse here that I will let you fill in yourself.) In particular, we get $\gcd(a, b) = r_k = \mathbf{w}_k \cdot [a, b]$. □

5. **Corollary 14.1** *The* $2 \times 2$ *matrix*

$$A = \begin{bmatrix} \mathbf{w}_k \\ \mathbf{w}_{k+1} \end{bmatrix}$$

*(with rows* $\mathbf{w}_k, \mathbf{w}_{k+1}$*) is* unimodular*: it has integer entries and determinant* $\pm 1$*. Moreover,*

$$A \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \gcd(a, b) \\ 0 \end{bmatrix} .$$

**Proof**. Note that $\mathbf{w}_{k+1} \cdot [a, b] = r_{k+1} = 0$. □

6. **An important application to relatively prime integers**.

Recall that two integers $a$ and $b$ are *relatively prime* (or *coprime*) if $\gcd(a,b)=1$.

**Theorem 14.2** *(a) Suppose $a$ and $b$ are relatively prime integers and that $a|(bc)$. Then $a$ must divide $c$ (i.e. $a|c$).*

*(b) Suppose the prime number $p$ divides the product $bc$. Then $p$ must divide either $b$ or $c$ (or both).*

*(c) Suppose $d$ is some common divisor of $a$ and $b$. Then $d$ divides* $\gcd(a,b)$.

**Proof**. In part (a) we assume $bc = za$ for some integer $z$. We are also assuming $a$ and $b$ are relatively prime, so from Theorem 14.1 we get (easily computed!) integers $u, v$ such that $1 = ua + bv$. Now here is a simple but very useful trick - multiply by $c$ to get

$$c = c \cdot 1 = uac + bcv = a(uc + zv) \ .$$

Thus $a$ really does divide $c$.

In part (b), $g = \gcd(p,b)$ could only equal $p$ or 1, being a positive divisor of the prime $p$. If $\gcd(p,b) = p$, then $p$ is also a divisor of $b$ and we have the result of the theorem. On the other hand, if $\gcd(p,b) = 1$, then by part (a) we get that $p$ is a divisor of $c$.

In part (c), we have $g = \gcd(a,b) = ua + bv$; we also assume $a = q_1 d$, $b = q_2 d$ for some integers $q_1, q_2$. But then we have $g = d(uq_1 + vq_2)$, so that $d$ really is a divisor of $g$. □

7. **Another important application in number theory**.

There is an extensive set of techniques for solving congruences. We will only dabble in this. Here is a central

**Theorem 14.3 The Chinese Remainder Theorem**.

(a) *Suppose integers $a_1, a_2 \geq 1$ are relatively prime. Then the system of congruences*

$$\left\{ \begin{array}{lll} x & \equiv & k_1 \pmod{a_1} \\ x & \equiv & k_2 \pmod{a_2} \end{array} \right. \tag{5}$$

*has a solution for any integers $k_1, k_2$. Moreover, this solution is unique modulo the product $a_1 a_2$. In detail, all the infinitely many solutions are given by $x + ta_1 a_2$, where $x$ is one particular solution and $t \in \mathbb{Z}$.*

(b) *More generally, suppose $a_1, \ldots, a_n$ are pairwise relatively prime positive integers. Then for any $k_1, \ldots, k_n \in \mathbb{Z}$, the system*

$$\left\{ \begin{array}{lll} x & \equiv & k_1 \pmod{a_1} \\ & \vdots & \\ x & \equiv & k_n \pmod{a_n} \end{array} \right. \tag{6}$$

*has a solution which is unique modulo the product $a_1 \cdots a_n$.*

**Proof**. We prove just part (a). By Theorem 14.1 there are integers $u_1, u_2$ such that

$$1 = u_1 a_1 + u_2 a_2 \ .$$

Let $x = k_1(u_2 a_2) + k_2(u_1 a_1)$. Then

$$\begin{aligned} x &= k_1(1 - u_1 a_1) + k_2 u_1 a_1 \\ &\equiv k_1(1-0) + k_1 0 \pmod{a_1} \\ &\equiv k_1 \pmod{a_1} \ . \end{aligned}$$

Likewise $x \equiv k_2 \pmod{a_2}$. Now if $y \in \mathbb{Z}$ is any other solution, then $y \equiv k_1 \equiv x \pmod{a_1}$. Thus $y - x$ is divisible by $a_1$ and similarly by $a_2$. Say then that

$$y - x = r a_1 = s a_2 \ .$$

Thus $a_2 | (r a_1)$, so $a_2 | r$ by Theorem 14.2(a). Therefore $y - x = t(a_1 a_2)$ for some $t \in \mathbb{Z}$ and $y \equiv x \pmod{a_1 a_2}$. □