Introduction to Abstract Algebra
4CCM121A/5CCM121B

Lecturer: Prof. Payman Kassaei

Semester 1, 2020-21

Contents

Chapter 1

Introduction

1.1 What is abstract algebra?

Abstract algebra is the study of algebraic structures. What then are "algebraic structures"? These are sets with binary operations satisfying certain properties (or "axioms"). Recall that a *binary operation* on a set S is a rule that assigns an element of S to each ordered pair of elements of S. The binary operation might be denoted $*$, and then the element it assigns to the pair a and b would be denoted $a * b$. (We say *ordered* pair because the order may matter; $a*b$ might not be the same as $b*a$.) We'll return to binary operations later. First let's consider some examples of sets with binary operations that turn out to be nice algebraic structures:

- $\mathbb{Z}$ (the set of integers) equipped with the operations $+$ (addition) and $\cdot$ (multiplication). Similarly, we could consider other familiar number systems: $\mathbb{Q}$ (rational numbers), $\mathbb{R}$ (real numbers), $\mathbb{C}$ (complex numbers) with their addition and multiplication operations.

- the set of 2×2 real matrices (denoted $M_2(\mathbb{R})$) with the operations of matrix addition and matrix multiplication. Recall that matrix addition for 2×2-matrices is given by the formula:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} a+a' & b+b' \\ c+c' & d+d' \end{pmatrix},$$

  and matrix multiplication by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} aa'+bc' & ab'+bd' \\ ca'+dc' & cb'+dd' \end{pmatrix}.$$

  (Note that there's no explicit symbol for matrix multiplication. We just write AB for the output of this operation applied to A and B.)

- The set of symmetries of an equilateral triangle, with the operation of "composition." To compose two symmetries (e.g., a 120° rotation, and reflection in an axis through a vertex), you apply one symmetry and then the other. This always gives you another symmetry (in this case, reflection in the axis through a different vertex— draw the picture), so defines a binary operation. More on this example later...

The first two above are examples of rings; the third is a group. These are two particular types of algebraic structures. There are other types of algebraic structures, but these two are the most fundamental. We'll focus mainly on groups this semester. We'll start with the precise definition of a group in the abstract; i.e., we'll give the axioms which must be satisfied by the binary operation on S in order for S to be called a *group*. Then we'll deduce general consequences from the axioms, i.e., we'll prove theorems about groups. This is the *abstraction*, an important tool in mathematics, but almost worthless without examples. So along the way we'll consider lots of examples and see how the theorems apply to them.

Before studying groups though, we'll study the most familiar and fundamental ring, namely $\mathbb{Z}$, the set of integers. We'll be considering its algebraic properties, i.e., how the integers behave with respect to addition and multiplication. I'll assume the most basic properties of the integers. The rest of this chapter and the beginning of the next comprise a review of material I'm assuming you've seen before, including the fact (see Theorem 2.2.2) that the gcd of a and b is also the least positive integer of the form $ax + by$ for $x, y \in \mathbb{Z}$. Recall also that in practice you can compute $\gcd(a, b)$ and find x and y using the *Euclidean Algorithm*; see Examples 2.2.4 and 2.2.5.

1.2 Sets

A **set** is a collection of things, called **elements** or **members** of the set. Here are some familiar sets with special names:

- The set of **integers** (or whole numbers) is denoted by $\mathbb{Z}$.
- The set of **natural numbers** (or positive integers) is denoted by $\mathbb{N}$. (Note: Some texts include 0 in the set $\mathbb{N}$. We will not.)
- The set of **rational numbers** is denoted by $\mathbb{Q}$.
- The set of **real numbers** is denoted by $\mathbb{R}$.

- The set of **complex numbers** is denoted by $\mathbb{C}$. (Recall that a complex number has the form $x + iy$ where x and y are real numbers and i is a square root of -1. Such a number is represented in the "complex plane" by the point in $\mathbb{R}^2$ with coordinates (x, y).)
- The set with no elements is called the **empty set** and denoted $\emptyset$.

You can define a particular set by describing it in words, or by listing its members in braces { }. For example, you could say B is the set of even integers, or you could write:

$$B = \{\ldots, -4, -2, 0, 2, 4, 6, \ldots\}.$$

The notation $x \in A$ means that x is an element of A (for example, $7 \in \mathbb{Z}$), and $x \notin A$ means that x is not an element of A (for example, $1/2 \notin \mathbb{Z}$).

If A and B are sets, then A is a **subset** of B is every element of A is also an element of B. The notation for this is $A \subseteq B$. For example, $\mathbb{N} \subseteq \mathbb{Z}$. We say is a **proper subset** of B (denoted $A \subsetneq B$) if $A \subseteq B$ but $A \neq B$. (I'm also used to writing $A \subset B$ to mean that A is a subset of B, but some texts use this to mean that A is a *proper* subset of B. So I'll try to avoid this notation altogether, but I might sometimes forget.)

Another way to describe a set involves specifying properties of its members. For example,

$$\{\, n \in \mathbb{N} \,|\, n \leq 3 \,\}$$

means the set of natural numbers $\mathbb{N}$ such that $n \leq 3$. The symbol | is used here to mean "such that." (Some people use : instead of |.)

Let's recall a few more definitions. Suppose that A and B are subsets of a set C. The **intersection** of A and B, denoted $A \cap B$, is the set of elements of C which belong to both A and B. The **union** of A and B, denoted $A \cup B$ is the set of elements of C which belong to A or B (or both). The **complement** of A in C is the set of elements of C which do not belong to A. This is written as $C \setminus A$. More generally we write $B \setminus A$ for the set of elements of B which do not belong to A. For example:

$$\mathbb{N} \setminus \{\, n \in \mathbb{Z} \,|\, -3 \leq n \leq 3 \,\} = \{\, 4, 5, 6, \ldots \,\}.$$

1.3 Logic and proofs

In mathematics, we try to state general facts about the objects we're studying (for example, the integers), and *prove* them by a logical sequence of steps starting from our definitions and basic principles and facts we've already

proved. We record such a fact as a **Proposition**. A proposition of particular importance is called a **Theorem**. A **Corollary** is a proposition which follows immediately from another. A **Lemma** is a proposition whose role is mainly as a tool to prove others.

Here is some standard terminology and notation from the language of logic and proofs:

- $\mathbf{P} \Rightarrow \mathbf{Q}$ means "If **P**, then **Q**" (or "**P** implies **Q**"). For example, $n \in \mathbb{Z} \Rightarrow n^2 + 1 \in \mathbb{N}$.
- $\mathbf{P} \Leftrightarrow \mathbf{Q}$" means "**P** if and only if **Q**," i.e., *both* of the following hold: $\mathbf{P} \Rightarrow \mathbf{Q}$, *and* $\mathbf{Q} \Rightarrow \mathbf{P}$.
- $\forall$ means "for all." For example, $n^2 + 1 \in \mathbb{N} \forall n \in \mathbb{Z}$.
- $\exists$ means "there exists."
- $\exists!$" means "there exists a unique." For example, $\exists! n \in \mathbb{N}$ such that $n^2 < 3$.

I'll sometimes use symbols such as $\Rightarrow$ or $\forall$ as shorthand during lectures, but not often when writing out lecture notes.

Mathematical proofs may be short and sweet, or long and complicated. There's no simple set of rules to follow when trying to come up with a proof. You might be able to arrive at a proof by "following your nose" from the definitions, or the proof might require some creative or clever ideas, or the assertion you're trying to prove might just be false. It's very important to be able to recognize a complete and correct proof. Beware of gaps and mistakes—be skeptical. *It's always useful to consider examples of the statement you're trying to prove.* This might give you an idea of how to prove the general statement, or they might provide a counterexample showing the general statement is false.

There are some standard "techniques of proof." One such technique is "proof by induction" reviewed below. Another is "proof by contradiction." You start by assuming the assertion you want to prove is *false*, proceed to deduce a contradiction, and thereby conclude that the assertion you want to prove must be *true*. Here's an example:

Proposition 1.3.1 *The real number $\sqrt{2}$ is irrational.*

Proof. Recall that $\sqrt{2}$ is defined as the positive real number x such that $x^2 = 2$. We want to prove that $x \notin \mathbb{Q}$. Let us suppose this is false, i.e., that $x \in \mathbb{Q}$. This means that there exist $m, n \in \mathbb{Z}$ with $n \neq 0$ such that $x = m/n$.

Changing the sign of m and n if necessary, we may suppose that $n \in \mathbb{N}$. We may further assume that n is the least positive integer such that $x = m/n$ for some $m \in \mathbb{Z}$. Since $(m/n)^2 = x^2 = 2$, it follows that $m^2 = 2n^2$. In particular m^2 is even, and therefore m must also be even. (If m were odd, then m^2 would also be odd, contradicting that m^2 is even.) So we can write $m = 2a$ for some $a \in \mathbb{Z}$. Substituting this into $m^2 = 2n^2$ gives $4a^2 = m^2 = 2n^2$, so $2a^2 = n^2$, So n^2 is even, and therefore so is n. Writing $n = 2b$ for some $b \in \mathbb{Z}$, we see that $m/n = 2a/2b = a/b$. Moreover $0 < b < n$, contradicting our assumption that n is the least positive integer such that $x = m/n$ for some $m \in \mathbb{Z}$. This contradiction shows that our initial assumption that $x \in \mathbb{Q}$ must be false. Therefore x is irrational. □

(The symbol □ is used to mark the end of a proof.)

1.4 Mathematical induction

We take for granted some basic properties of the integers $\mathbb{Z}$. In particular, we assume the existence and basic properties of the multiplication and addition operations; for example, they satisfy the *distributive law* $a(b+c) = ab+ac$ for all $a, b, c \in \mathbb{Z}$. The integers also have an ordering: for any $a, b \in \mathbb{Z}$, exactly one of the following three things is true: $a < b$, $a = b$, or $a > b$. Recall that $a \leq b$ means "$a < b$ or $a = b$." We assume basic properties of the ordering as well, for example: If $a \leq b$ and $b \leq c$, then $a \leq c$. So far, everything I've said is true for the real numbers $\mathbb{R}$ as well, but $\mathbb{Z}$ has the following special property:

The Well-ordering Principle[1]: Every non-empty subset of $\mathbb{Z}$ that is bounded below has a least element.

Recall that a subset $A \subseteq \mathbb{Z}$ is **bounded below** if there exists $n \in \mathbb{Z}$ such that $n \leq a$ for all $a \in A$. Such an n is called a **lower bound** for A, and a **least element** of A is an element of A which is also a lower bound for A. For example, if A is any subset of $\mathbb{N}$, then 1 is a lower bound for A, so A has a least element. We see also that if A is a subset of $\mathbb{Z}$ which is bounded *above*, then it has a *greatest* element. (Consider the set $B = \{\, b \in \mathbb{Z} \mid -b \in A \,\}$. Check that B is bounded below, hence has a *least* element, the negative of which is a *greatest* element for A.)

The principle of mathematical induction is based on the Well-Ordering Principle.

Theorem 1.4.1 *Suppose that $P(n)$ is an assertion for each $n \in \mathbb{N}$. If*

[1] We actually already used this and other basic properties of $\mathbb{Z}$ in the proof that $\sqrt{2}$ is irrational.

1. *$P(1)$ is true, and*

2. *for every integer $n > 1$, we have*

$$P(1), P(2), \ldots, P(n-1) \Rightarrow P(n),$$

then $P(n)$ is true for all $n \in \mathbb{N}$.

Proof. Let A be the set of positive integers n for which $P(n)$ is false. We want to show that $A = \emptyset$. Suppose instead that A is non-empty. Then, by the well-ordering principle, it has a least element m. Since $P(1)$ is true, we know that $m > 1$. Since m is the least element of A, we know that none of $1, 2, \ldots, m-1$ are elements of A. This means that $P(1), P(2), \ldots, P(m-1)$ are all true. Therefore $P(m)$ is true as well, a contradiction to our assumption that $m \in A$. Therefore A is empty, i.e., $P(n)$ is true for all $n \in \mathbb{N}$. □

A "proof by induction" is one which uses the above theorem. In order to apply the theorem, we need to show that the hypotheses apply in our situation, namely 1) that P(1) is true, and 2) that $P(1), \ldots, P(n-1) \Rightarrow P(n)$ for all $n > 1$. It is important be very clear about what the assertion $P(n)$ is. (It may be the actual statement of the proposition, or it might be some intermediate or partial result instead.) Often we just prove the *stronger* hypothesis that $P(n-1) \Rightarrow P(n)$ for all $n > 1$. Since this implies 2) above, we can still apply the theorem to conclude $P(n)$ for all $n \in \mathbb{N}$. Note also that we could just as well prove an assertion $P(n)$ for all $n \geq 0$ by showing 1) that $P(0)$ is true and 2) that $P(0), \ldots, P(n-1) \Rightarrow P(n)$ for all $n > 0$. This version of the induction principle can be proved in exactly the same way as the preceding theorem, or deduced from it by applying the theorem to $Q(n)$, where $Q(n)$ is the assertion $P(n-1)$. In fact there's nothing special about 0 or 1. For example, you could use mathematical induction to prove that an assertion holds for all integers $n \geq -27$.

1.5 Functions

Finally recall that a **function** from a set A to a set B is a rule that assigns exactly one element of B to each element of A. We write $f : A \to B$ to mean that f is a function from A to B, and for $a \in A$, we write $f(a)$ for the element of B that the function assigns to a. A function is often defined by a formula. Consider for example the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$. Note that some elements of B may be assigned to more than one element of A, or to none at all. In the example of $f(x) = x^2$, we have $f(-3) = f(3) = 9$, but there is no $x \in \mathbb{R}$ such that $f(x) = -1$.

Recall that if A and B are sets, a **function** from A to B is a rule that assigns exactly one element of B to each element of A. We write $f : A \to B$ to indicate that f is a function from A to B. The set A is called the **domain** of f; the set B is called the **codomain** (or **target**) of f. We write $f(a)$ (read "f of a") for the element of B which f assigns to a; $f(a)$ is called the **value** of f at a. Here are a few examples:

- the function $f : \mathbb{Z} \to \mathbb{Z}$ defined by $f(n) = 2n + 1$;
- the sine function $\sin : \mathbb{R} \to \mathbb{R}$;
- the sign function $\text{sign} : \mathbb{R} \to \{1, -1\}$ defined by

$$\text{sign}(x) = \begin{cases} 1, & \text{if } x > 0; \\ -1, & \text{if } x < 0; \end{cases}$$

- the *modulus* function $|\ |$ from $\mathbb{C}$ to $\mathbb{R}$ defined by $|x + iy| = \sqrt{x^2 + y^2}$.

You can think an element a in the domain A as "input" for the function a, and the value $f(a) \in B$ as "output." You might also sometimes think of f as a "mapping" from A to B and represent it in a diagram by an arrow from A to B.

Note that different elements of A can be assigned the same value in B; for example $\sin 0 = \sin \pi = 0$. Also, not every element of b needs to be a value of the function; for example, there is no $x \in \mathbb{R}$ such that $\sin x = 2$. What *is* required is that for each $a \in A$, there is *exactly one* element of B which we call $f(a)$. So some *non-examples* of functions are

- $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 1/x$, as it is not defined for $x = 0$;
- $g : \mathbb{R} \to \mathbb{Z}$ defined by $g(x)$ is the nearest integer to x, since the nearest integer to $x = 1/2$, for example, is not unique.

Some functions *do* have the property that no two distinct elements of the domain are assigned the same value; i.e., if a and a' are assigned the same value by f, then $a = a'$. Such functions are called *injective* is *one-to-one*.

Definition 1.5.1 Suppose that f is a function from A to B. We say f is **injective** if it has the following property:

$$a, a' \in A,\ f(a) = f(a') \quad \Rightarrow \quad a = a'.$$

For example, the function $f : \mathbb{Z} \to \mathbb{Z}$ defined by $f(n) = 2n + 1$ is injective since if $m, n \in \mathbb{Z}$ and $2m+1 = 2n+1$, then $m = n$. The function $\sin : \mathbb{R} \to \mathbb{R}$ is not injective.

Likewise, some functions have the property that every element of the codomain B is a value of the function. Such functions are called *surjective* or *onto*.

Definition 1.5.2 Suppose that f is a function from A to B. We say f is **surjective** if it has the following property:

$$b \in B \quad \Rightarrow \quad b = f(a) \text{ for some } a \in A.$$

For example, the function $\text{sign} : \mathbb{R} \to \{\pm 1\}$ is surjective, but $\sin : \mathbb{R} \to \mathbb{R}$ is not. We can also express this in terms of the set of values of the function, called its *range*.

Definition 1.5.3 Suppose that f is a function from A to B. The set

$$f(A) = \{\, b \in B \,|\, b = f(a) \text{ for some } a \in A \,\} = \{\, f(a) \,|\, a \in A \,\}$$

is called the **range** (or **image**) of f.

Observe the notation: instead of putting an element a of the domain in the parentheses following f, we use the *set* A. While the value $f(a)$ is an *element* of the codomain B, the range $f(A)$ is a *subset* of B. For example, if $f : \mathbb{Z} \to \mathbb{Z}$ is defined by $f(n) = 2n+1$, then the range of f is the set of odd integers. The range of $\sin : \mathbb{R} \to \mathbb{R}$ is the closed interval $[-1, 1] = \{\, x \in \mathbb{R} \,|\, -1 \leq x \leq 1 \,\}$. A function $f : A \to B$ is surjective if and only if $f(A) = B$, so neither of these functions is surjective.

Definition 1.5.4 A function $f : A \to B$ is **bijective** if it is both injective and surjective.

An example of a bijective function if the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 2x + 1$. It is injective since $2x + 1 = 2y + 1 \Rightarrow 2x = 2y \Rightarrow x = y$. It is surjective since if $y \in \mathbb{R}$, then $y = f((y - 1)/2)$ is in the range of f.

Functions are sometimes described by formulas without specifying their domain and codomain; for example, we might just write $f(x) = 1/x$. In that case, it's implicitly understood that the domain is the set of $x \in \mathbb{R}$ such that $f(x)$ is defined ($\mathbb{R} \setminus \{0\}$ in this example) and the codomain is $\mathbb{R}$. But for the notions we'll be discussing, such as *bijectivity*, it's important to be clear about the domain and codomain of the function.

Chapter 2

The integers

2.1 The Division Algorithm

Recall the definition of divisibility: Suppose that $m, n \in \mathbb{Z}$. We say m is **divisible** by n (written $n|m$) if $m = nk$ for some $k \in \mathbb{Z}$. We say also that n is a **divisor** of m, and m is a **multiple** of n. The proof of the following basic properties of divisibility is left as an exercise:

Proposition 2.1.1 *Suppose that $a, b, c, d \in \mathbb{Z}$ and $n \in \mathbb{N}$.*

1. *If $a|b$ and $b|c$, then $a|c$.*

2. *If $a|n$, then $a \leq n$.*

3. *If $n|a$ and $n|b$, then $n|(ac + bd)$.*

The following theorem is known as the **Division Algorithm**[1]:

Theorem 2.1.2 *Suppose that $m \in \mathbb{Z}$ and $n \in \mathbb{N}$. Then there exist $q, r \in \mathbb{Z}$ such that both of the following hold:*

1. *$m = qn + r$, and*

2. *$0 \leq r < n$.*

Moreover, q and r are the unique pair of integers such that these both hold.

[1] An "algorithm" for finding quotients and remainders for integer division is embedded in its proof. As an algorithm though it's rather inefficient. You learned a more practical method of finding q and r in primary school.

Proof. Let $P(m)$ be the "existence" assertion, i.e., that there exist $q, r \in \mathbb{Z}$ such that 1) $m = qn+r$, and 2) $0 \leq r < n$. We first prove $P(m)$ for $m \geq 0$ by induction on m (where n is now an arbitrary fixed positive integer). We start the induction argument by noting not only that $P(0)$ is true, but indeed that $P(0), P(1), \ldots, P(n-1)$ are all true, for if $0 \leq m < n$, then we can take $q = 0$ and $r = m$. Now suppose that $m \geq n$ and that $P(0), P(1), \ldots, P(m-1)$ are all true. Let $m' = m - n$. Then $0 \leq m' < m$, so $P(m')$ is true. This means we can write:

$$m' = q'n + r' \quad \text{for some } q', r' \in \mathbb{Z} \text{ with } 0 \leq r' < n.$$

Since $m = n + m'$, we therefore have

$$m = n + q'n + r' = (1 + q')n + r' = qn + r,$$

where $q = q' + 1$ and $r = r'$. Since $q, r \in \mathbb{Z}$ and $0 \leq r = r' < n$, we conclude that $P(m)$ is true. The principle of induction now shows that $P(m)$ is true for all integers $m \geq 0$.

Now suppose $m < 0$. Then $-m > 0$, so we have already proved that $P(-m)$ is true. This means that there are integers q', r' such that

$$-m = q'n + r' \quad \text{and } 0 \leq r' < n.$$

If $r' = 0$, then we have $m = qn + r$ where $q = -q'$ and $r = 0$, so $P(m)$ is true. On the other hand if $0 < r' < n$, then we have

$$m = -q'n - r' = -n - q'n + n - r' = qn + r$$

where $q = -1 - q'$ and $r = n - r'$. Note that $q, r \in \mathbb{Z}$ and $0 \leq r < n$, so $P(m)$ is true in this case as well. We have now shown that $P(m)$ is true for all $m \in \mathbb{Z}$.

We still need to prove the "uniqueness" assertion. Suppose we have two pairs of integers, say q_1, r_1 and q_2, r_2, so that 1) and 2) hold; i.e.,

$$m = q_1 n + r_1 = q_2 n + r_2 \quad \text{and} \quad 0 \leq r_1, r_2 < n.$$

We have to prove that $q_1 = q_2$ and $r_1 = r_2$. We first show that $r_1 = r_2$. Suppose that $r_1 \neq r_2$. Without loss of generality, we can assume $r_1 < r_2$. Then $0 < r_2 - r_1 \leq r_2 < n$, but $r_2 - r_1 = (q_1 - q_2)n$ is a multiple of n, a contradiction. Therefore $r_1 = r_2$. Now it follows that $(q_1 - q_2)n = 0$, and since $n \neq 0$, we conclude that $q_1 - q_2 = 0$, so $q_1 = q_2$. $\square$

2.2 The Euclidean Algorithm

Definition 2.2.1 Suppose a and b are integers, not both zero. Then the **greatest common divisor** of a and b, denoted $\gcd(a, b)$, is the greatest integer which divides both a and b. Thus $g = \gcd(a, b)$ if the following hold:

1. $g|a$ and $g|b$, and
2. if $d|a$ and $d|b$, then $d \leq g$.

For example, $\gcd(114, 42) = 6$ since the (positive) divisors of 42 are 1, 2, 3, 6, 7, 14, 21 and 42, and those which also divide 114 are 1, 2, 3 and 6.

The following theorem gives two important properties of the gcd:

Theorem 2.2.2 *Suppose that $a, b \in \mathbb{Z}$ and that a and b are not both zero, and let $g = \gcd(a, b)$. Then*

1. *g is the least positive integer of the form $ax + by$ with $x, y \in \mathbb{Z}$;*
2. *if $d|a$ and $d|b$, then $d|g$.*

Proof. Let S be the set of positive integers n such that $n = ax + by$ for some $x, y \in \mathbb{Z}$. Then S is non-empty; for example $|a| + |b| \in S$. So by the Well-Ordering Principle, S has a least element, say m. Thus m is the least positive integer of the form $ax + by$ with $x, y \in \mathbb{Z}$.

We will show that 1) $m|a$ and $m|b$, and 2) $d|a$ and $d|b \Rightarrow d|m$. Since $d|m \Rightarrow d \leq m$ (as $m > 0$), it follows that m is in fact $\gcd(a, b) = g$, and we will have proved both parts of the theorem.

1) Let us prove that $m|a$. By Thm. 2.1.2[2], $a = mq + r$ for some $q, r \in \mathbb{Z}$, and $0 \leq r < m$. We will prove by contradiction that $r = 0$. Suppose that $r > 0$. Since $m = ax + by$ for some $x, y \in \mathbb{Z}$, we have

$$\begin{aligned} r &= a - mq \\ &= a - (ax + by)q \\ &= a(1 - xq) + b(-yq). \end{aligned}$$

We have now written r in the form $ax' + by'$ for some $x', y' \in \mathbb{Z}$. So if r is positive, then $r \in S$. But then $r < m$, contradicting that m is the least element of S. We therefore conclude that $r = 0$ and $a = mq$ is divisible by m. The proof that $m|b$ is similar.

[2]This is a standard trick for proving divisibility: apply the Division Algorithm and show the remainder has to be 0.

2) Since $m = ax + by$ for some $x, y \in \mathbb{Z}$, it follows that if $d|a$ and $d|b$, then $d|m$ (Prop. 2.1.1, part 3). □

In the example of $a = 114$, $b = 42$, we could take $x = 3$, $y = -8$. Note that there are other possible values of x and y; for example, $x = -4$, $y = 11$. We'll see later how to find all possible values of x and y.

You may have seen a different proof of Thm. 2.2.2 using the Euclidean Algorithm. The Euclidean algorithm is a nice efficient way of computing $\gcd(a, b)$ without having to find *all* the divisors of a and b. It works by repeatedly applying the Division Algorithm (Thm. 2.1.2). Let's assume that $b > 0$ (we can swap a and b or change their signs without changing their gcd).

We begin by dividing a by b to get $a = qb + r$ with $q, r \in \mathbb{Z}$ and $0 \leq r < b$. Since we're going to iterate this process, let $q_1 = q$ and $r_1 = r$. Now divide b by r_1 to get $b = q_2 r_1 + r_2$ with $q_2, r_2 \in \mathbb{Z}$, $0 \leq r_2 < r_1$. Next divide r_1 by r_2 to get $r_1 = q_3 r_2 + r_3$ with $q_3, r_3 \in \mathbb{Z}$, $0 \leq r_3 < r_2$. Keep repeating this as long as $r_i > 0$, so having found $r_1 > r_2 > \cdots > r_{i-1} > r_i$, the $(i+1)^{\text{st}}$ step in the process gives

$$r_{i-1} = q_{i+1} r_i + r_{i+1} \quad \text{with } 0 \leq r_{i+1} < r_i.$$

Since the remainders keep decreasing, we must eventually get $r_n = 0$ for some $n > 0$.

To illustrate the roles of a and b in starting the process, we could let $r_{-1} = a$ and $r_0 = b$, so that our sequence of equations now reads:

$$\begin{array}{rcccl} r_{-1} & = & q_1 r_0 & + & r_1 \\ r_0 & = & q_2 r_1 & + & r_2 \\ r_1 & = & q_3 r_2 & + & r_3 \\ & & \vdots & & \\ r_{n-3} & = & q_{n-1} r_{n-2} & + & r_{n-1} \\ r_{n-2} & = & q_n r_{n-1} & & \end{array}$$

with $r_0 > r_1 > \cdots > r_{n-1} > r_n = 0$.

Then the last non-zero remainder, namely r_{n-1}, turns out to be $\gcd(a, b)$. To see why this works, note the following:

Proposition 2.2.3 *If $a, b, c, k \in \mathbb{Z}$ with $b \neq 0$ and $a = kb + c$, then $\gcd(a, b) = \gcd(b, c)$.*

Proof. If $d|a$ and $d|b$, then $d|c$ since $c = a - bk$. So if d is a common divisor of a and b, then it is a common divisor of a and c as well. Similarly, if $d|b$

and $d|c$, then $d|a$ (and $d|b$), so in fact the common divisors of a and b are the *same* as the common divisors of b and c. Therefore $\gcd(a,b) = \gcd(b,c)$. $\square$

In view of the equations $r_{i-1} = q_{i+1}r_i + r_{i+1}$ for $i = 0, \ldots, n-2$, the proposition shows that

$$\gcd(a,b) = \gcd(r_{-1}, r_0) = \gcd(r_0, r_1) = \cdots = \gcd(r_{n-2}, r_{n-1}).$$

Since $r_{n-1}|r_{n-2}$, we have $\gcd(r_{n-2}, r_{n-1}) = r_{n-1}$.

Example 2.2.4 We work through the Euclidean algorithm for $a = 114$, $b = 42$:

$$\begin{array}{rcl} 114 & = & 2 \cdot 42 + 30 \\ 42 & = & 1 \cdot 30 + 12 \\ 30 & = & 2 \cdot 12 + 6 \\ 12 & = & 2 \cdot 6 + 0. \end{array}$$

So the last non-zero remainder is 6, and indeed $\gcd(114, 42) = 6$.

The Euclidean algorithm also provides an algorithm for finding integers x and y so that $\gcd(a,b) = ax + by$. Note that we can rewrite the first $n-1$ equations of the Euclidean Algorithm, in reverse order, as:

$$\begin{array}{rcl} r_{n-1} & = & r_{n-3} - q_{n-1}r_{n-2} \\ r_{n-2} & = & r_{n-4} - q_{n-2}r_{n-3} \\ & \vdots & \\ r_2 & = & r_0 - q_2 r_1 \\ r_1 & = & r_{-1} - q_1 r_0. \end{array}$$

Since $\gcd(a,b) = r_{n-1}$, the first equation gives $\gcd(a,b)$ in terms of r_{n-3} and r_{n-2}. Using the next equation to substitute for r_{n-2}, we get $\gcd(a,b)$ in terms of r_{n-4} and r_{n-3}. Iterating this, we eventually get it in terms of $r_{-1} = a$ and $r_0 = b$.

Example 2.2.5 Again consider $a = 114$, $b = 42$. "Unwinding" the equations from Example 2.2.4 gives:

$$\begin{array}{rcl} 6 & = & 30 - 2 \cdot 12 \\ 12 & = & 42 - 30 \\ 30 & = & 114 - 2 \cdot 42. \end{array}$$

Substituting each equation into the previous one gives:

$$\begin{array}{rcl} 6 & = & 30 - 2 \cdot 12 \\ & = & 30 - 2 \cdot (42 - 30) \quad = \quad -2 \cdot 42 + 3 \cdot 30 \\ & & \qquad\qquad\quad = \quad -2 \cdot 42 + 3(114 - 2 \cdot 42) = 3 \cdot 114 - 8 \cdot 42. \end{array}$$

So we get $6 = \gcd(114, 42)$ in the form $114x - 42y$ by taking $x = 3$, $y = -8$.

2.3 Relatively prime integers

The following notion is a useful one when working with divisibility:

Definition 2.3.1 Suppose that $a, b \in \mathbb{Z}$ and that a and b are not both zero. We say that a and b are **relatively prime** if $\gcd(a, b) = 1$.

For example, 15 and 28 are relatively prime. The integers 114 and 42 are *not* relatively prime.

We have the following corollaries of Thm. 2.2.2 (recall this says $\gcd(a, b)$ is the least positive integer of the form $ax + by$ with $x, y \in \mathbb{Z}$):

Corollary 2.3.2 *Suppose that $a, b \in \mathbb{Z}$, not both 0. Then a and b are relatively prime if and only if $ax + by = 1$ for some $x, y \in \mathbb{Z}$.*

Before explaining why this is immediate from Thm. 2.2.2, let's read the statement of the corollary carefully. Note that the conclusion is of the form **P** *if and only if* **Q**, so we have to prove two things: 1) if **P** is true (for the integers a and b), then **Q** is also true, and 2) if **Q** is true, then so is **P**. Let's prove 1): **P** is the statement that a and b are relatively prime. According to Def. 2.3.1, this means that $\gcd(a, b) = 1$. Thm. 2.2.2 then says that 1 is the least positive integer of the form $ax + by$. Therefore **Q** is true. Now let's prove 2): If $ax + by = 1$ for some $x, y \in \mathbb{Z}$, then 1 is certainly the *least* positive integer of the form $ax+by$ with $x, y \in \mathbb{Z}$. Applying Thm. 2.2.2 again shows that $\gcd(a, b) = 1$, i.e., that is true.

Corollary 2.3.3 *Suppose that $a, b \in \mathbb{Z}$, not both 0. Let $g = \gcd(a, b)$. Then a/g and b/g are relatively prime.*

Proof. By Thm. 2.2.2, $g = ax + by$ for some $x, y \in \mathbb{Z}$, so $1 = \frac{a}{g} \cdot x + \frac{b}{g} \cdot y$. Now Cor. 2.3.2 shows that $\frac{a}{g}$ and $\frac{b}{g}$ are relatively prime. □

For example, $\gcd(114, 42) = 6$, so $19 = 114/6$ and $7 = 42/6$ are relatively prime.

Corollary 2.3.4 *Suppose that $a, b, c \in \mathbb{Z}$ with a and b relatively prime. If $a|bc$, then $a|c$.*

Proof. Since a and b are relatively prime, we have $ax + by = 1$ for some $x, y \in \mathbb{Z}$. Multiplying through by c gives $acx + bcy = c$, so if $a|bc$, then a divides both terms of $acx + bcy$, so $a|c$. □

Note that we need the assumption that a and b are relatively prime. For example, if $a = 6$, $b = 2$ and $c = 3$, then $a|bc$, but neither b or c is divisible by a.

Corollary 2.3.5 *Suppose that $a, b, c \in \mathbb{Z}$ with a and b relatively prime. If $a|c$ and $b|c$, then $ab|c$.*

Proof. If $b|c$, then $c = bm$ for some $m \in \mathbb{Z}$. If also $a|c$, then $a|bm$. Since a and b are relatively prime, Cor. 2.3.4 implies that $a|m$. This means that $m = an$ for some $n \in \mathbb{Z}$. Therefore $c = abn$, so c is divisible by ab. □

Again, note that it's not always the case that if $a|c$ and $b|c$, then $ab|c$. (Take for example $a = b = c = 2$.) We need a and b to be *relatively prime* to draw this conclusion.

2.4 Linear Diophantine equations

A linear Diophantine equation (in two variables) is an equation of the form

$$ax + by = c,$$

where $a, b, c \in \mathbb{Z}$ and we regard x and y as variables taking only *integer* values. (It is *linear* because the graph is a line; *Diophantine*, after the Greek mathematician Diophantus, refers to the restriction to integer values.)

We will now show how to determine whether a given linear Diophantine equation has solutions, and develop an algorithm for finding them whenever it does.

Theorem 2.4.1 *Suppose that $a, b, c \in \mathbb{Z}$ and that a and b are not both zero. Then the equation $ax+by = c$ has solutions $x, y \in \mathbb{Z}$ if and only if $\gcd(a, b)|c$.*

Proof. Suppose first that $ax + by = c$ has a solution with $x, y \in \mathbb{Z}$. Let $g = \gcd(a, b)$. Since $g|a$ and $g|b$, we have $g|(ax + by) = c$ by Prop. 2.1.1.

Conversely[3], suppose that $g|c$. This means that $c = gk$ for some $k \in \mathbb{Z}$. By Thm. 2.2.2, we know that $g = am + bn$ for some $m, n \in \mathbb{Z}$. Multiplying this by k gives

$$c = gk = (am + bn)k = a(mk) + (nk).$$

Since mk and nk are integers, we've shown that $ax+by = c$ for some $x, y \in \mathbb{Z}$ (namely $x = mk$ and $y = nk$). □

[3]Recall the "if and only if" means each assertion implies the other. We've just shown **P** ⇒ **Q**. Now we have to prove the converse, that **Q** ⇒ **P**, i.e., that if $g|c$, then the equation has integer solutions.

Example 2.4.2 Consider the equation $114x + 42y = 660$. We know that $\gcd(114, 42) = 6$ (see Example 2.2.4) and $6|660$, the equation has solutions. The Euclidean Algorithm gives $114 \cdot 3 + 42 \cdot (-8) = 1$ (see Example 2.2.5). Multiplying through by $660/6 = 110$ gives

$$114 \cdot 330 + 42 \cdot (-880) = 660,$$

so $x = 330$, $y = -880$ is a solution.

Example 2.4.3 The equation $21x+35y = 900$ has no integer solutions since 900 is not divisible by $\gcd(21, 35) = 7$.

Example 2.4.4 If $a = 0$ (but $b \neq 0$), then $ax + by = c$ becomes $by = c$, which has integer solutions if and only if $b|c$. This is consistent with the statement of the theorem since $\gcd(0, b) = b$.

Now we explain how to find all the solutions assuming we've found one.

Theorem 2.4.5 *Suppose that $a, b, c \in \mathbb{Z}$ and that $g = \gcd(a, b)$. If $x = x_0$, $y = y_0$ is an integer solution to $ax + by = c$, then all the integer solutions of $ax + by = c$ are given by*

$$x = x_0 + k \cdot \frac{b}{g}, \quad y = y_0 - k \cdot \frac{a}{g} \quad \textit{for } k \in \mathbb{Z}.$$

Proof. We first observe that if $x = x_0$, $y = y_0$ is an integer solution to $ax + by = c$, then so is $x = x_0 + k \cdot \frac{b}{g}$, $y = y_0 - k \cdot \frac{a}{g}$ since

$$a\left(x_0 + k \cdot \frac{b}{g}\right) + b\left(y_0 - k \cdot \frac{a}{g}\right) = ax_0 + \frac{kab}{g} + by_0 - \frac{kab}{g} = ax_0 + by_0 = c.$$

Next we show that if $x, y \in \mathbb{Z}$ is a solution of the equation $ax + by = c$, then it has the required form. Since $ax_0+by_0 = c$, we can rewrite the equation as $ax+by = ax_0+by_0$, which is equivalent to $a(x-x_0) = b(y_0-y)$. Dividing through by $g = \gcd(a, b)$ gives $\frac{a}{g}(x - x_0) = \frac{b}{g}(y_0 - y)$. Therefore $\frac{b}{g}(y_0 - y)$ is divisible by $\frac{a}{g}$. Since $g = \gcd(a, b)$, we know by Cor. 2.3.3, we know that $\frac{a}{g}$ and $\frac{b}{g}$ are relatively prime. So by Cor. 2.3.4, we know that $y_0 - y$ is divisible by $\frac{a}{g}$. This means that $y_0 - y = k \cdot \frac{a}{g}$ for some $k \in \mathbb{Z}$, which implies that $y = y_0 k \cdot \frac{a}{g}$. Substituting $y_0 - y = k \cdot \frac{a}{g}$ into the equation $\frac{a}{g}(x - x_0) = \frac{b}{g}(y_0 - y)$ gives

$$\frac{a}{g}(x - x_0) = \frac{b}{g} \cdot k \cdot \frac{a}{g},$$

which implies[4] that $x - x_0 = k \cdot \frac{b}{g}$. We have now shown that for x, y to be a solution, we must have

$$x = x_0 + k \cdot \frac{b}{g},\ y = y_0 - k \cdot \frac{a}{g} \quad \text{for } k \in \mathbb{Z}.$$

□

Example 2.4.6 Let's apply this to find *all* solutions of $114x+42y = 660$. We already found one solution in Example 2.4.2, namely $x_0 = 330$, $y_0 = -880$. Since $a/g = 19$, $b/g = 7$, Thm. 2.4.5 gives all the solutions as:

$$x = 330 + 7k, \quad y = -880 - 19k \qquad \text{for } k \in \mathbb{Z}.$$

We might also want to find solutions subject to some constraints. For example, we might be looking for solutions where x and y are positive, or non-negative. This translates into solving the corresponding inequalities to find suitable values of k (if there are any).

Example 2.4.7 Let's find all solutions of $114x + 42y = 660$ with $x, y \in \mathbb{N}$ (the set of positive integers). From Example 2.4.6, this translates into the inequalities

$$x = 330 + 7k > 0, \quad y = -880 - 19k > 0.$$

Solving the first inequality for k gives $7k > -330$, so $k > -330/7 \approx -47.1$. Solving the second gives $19k < -880$, so $k < -880/19 \approx -46.3$. The only integer satisfying these inequalities is $k = -47$. Substituting this into the formulas for x and y gives the solution:

$$x = 1, \quad y = 13.$$

We now sum up the method for solving $ax + by = c$ with $x, y \in \mathbb{Z}$.

1. Use the Euclidean algorithm to find $g = \gcd(a, b)$. If $g \nmid c$, then there are no solutions (Thm. 2.4.1).

2. If $g|c$, then "unwind" the equations from the Euclidean algorithm to get $g = am + bn$, so $x_0 = \frac{mc}{g}$, $y_0 = \frac{nc}{g}$ is a solution.

3. If $g|c$, then all solutions are given by (Thm. 2.4.1)

$$x = x_0 + k \cdot \frac{b}{g},\ y = y_0 - k \cdot \frac{a}{g} \quad \text{for } k \in \mathbb{Z}.$$

[4]This assume $a \neq 0$. The case $a = 0$ is left as an exercise.

4. If there are sconstraints on the solutions (e.g., x and y need to be positive), then solve the corresponding inequalities for k.

Example 2.4.8 You're in a shop where apples cost 27p and oranges cost 69p. What are all the possible ways of spending exactly £8.40 on apples and oranges? This translates into finding all solutions of

$$27x + 69y = 840$$

where x and y are non-negative integers[5].

1. We work through the Euclidean algorithm for $a = 27$, $b = 69$:

$$\begin{array}{rcrcl} 69 & = & 2 \cdot 27 & + & 15 \\ 27 & = & 1 \cdot 15 & + & 12 \\ 15 & = & 1 \cdot 12 & + & 3 \\ 12 & = & 4 \cdot 3 & + & 0. \end{array}$$

   The last non-zero remainder is 3, so $g = \gcd(27, 69) = 3$. Since $3|840$, the equation has integer solutions.

2. Solving for the remainder in each of the above equations gives:

$$\begin{aligned} 15 &= 69 - 2 \cdot 27 \\ 12 &= 27 - 15 \\ 3 &= 15 - 12. \end{aligned}$$

   Substituting each equation into the previous one gives:

$$\begin{array}{rcl} 3 & = & 15 - 12 \\ & = & 15 - (27 - 15) \quad = \quad 2 \cdot 15 - 27 \\ & & \qquad\qquad = \quad 2 \cdot (69 - 2 \cdot 27) - 27 = 2 \cdot 69 - 5 \cdot 27. \end{array}$$

   So we get $3 = \gcd(27, 69)$ in the form $27m + 69n$ by taking $m = -5$, $n = 2$. Multiplying through by $840/3 = 280$ gives the solution

$$x_0 = -1400, \quad y_0 = 560.$$

3. Since $a/g = 9$ and $b/g = 23$, Thm. 2.4.5 gives all integer solutions as:

$$x = -1400 + 23k, \quad y = 560 - 9k \qquad \text{for } k \in \mathbb{Z}.$$

[5]Assume you can buy *no* apples, but you can't buy a negative apple or just part of an apple.

4. The values of k giving non-negative solutions must satisfy

$$-1400 + 23k \geq 0, \quad 560 - 9k \geq 0,$$

The first inequality is equivalent to $23k \geq 1400$, or $k \geq 1400/23 \approx 60.9$; the second is equivalent to $9k \leq 560$, or $k \leq 560/9 \approx 62.2$. We can therefore have $k = 61$ or 62, giving the two solutions:

- 3 apples and 11 oranges;
- 26 apples and 2 oranges.

2.5 Prime factorization

Recall the following definition:

Definition 2.5.1 Suppose that n is an integer and $n > 1$. Then n is **prime** if its only positive divisors are 1 and n; otherwise n is **composite**.

A fact which should be familiar to you, though you may not have seen a proof, is that every integer bigger than 1 can be written as a product of prime numbers (allowing repetition). Recall how to do this in practice, at least for a reasonably small integer n. If n is not prime, then it factors as $n' \cdot n''$ where n' and n'' are positive integers less than n. Now similarly factor n' and n'', repeating this process, and stopping only after all the factors are prime. For example, $36 = 2 \cdot 18$, 2 is prime, but 18 is not, so we factor $18 = 2 \cdot 9$, and $9 = 3 \cdot 3$, finally giving the prime factorization $36 = 2 \cdot 2 \cdot 3 \cdot 3$. We could also write this is as $36 = 2^2 3^2$. Here are prime factorizations for the next few integers:

$$37 = 37 \text{ (a prime)} , \ 38 = 2 \cdot 19, \ 39 = 3 \cdot 13, \ 40 = 2^3 5.$$

Another important fact is that this factorization into primes is essentially unique; the only possible difference between two prime factorizations of the same integer n would be to change the order of the factors. The existence and uniqueness of such prime factorizations of integers is called the Fundamental Theorem of Arithmetic. We give the statement and proof below; first we prove a basic fact about divisibility by primes.

Proposition 2.5.2 *Suppose that $a, b \in \mathbb{Z}$ and p is a prime number. If $p|ab$ then $p|a$ or $p|b$.*

Proof. Suppose that $p|ab$ but $p \nmid a$. Then $\gcd(a, p) = 1$, so by Cor. 2.3.4, $p|b$. $\square$

Corollary 2.5.3 *Suppose that $a_1, a_2, \ldots, a_k \in \mathbb{Z}$ and p is a prime number. If $p|a_1a_2\cdots a_k$ then $p|a_i$ for some $i \in \{1, 2, \ldots, k\}$.*

Proof. We prove the corollary by induction on k. Note that the corollary is true if $k = 1$, in which case it just says that if $p|a_1$, then $p|a_1$,

Suppose now that $k > 1$ and the corollary is true with k replaced by $k - 1$. If $p|a_1a_2\cdots a_{k+1}$, then writing $a_1a_2\cdots a_k = (a_1a_2\cdots a_{k-1})a_k$ and applying Prop. 2.5.2, we see that $p|a_1a_2\cdots a_{k-1}$ or $p|a_k$. So if $p \nmid a_k$, then $p|a_1a_2\cdots a_{k-1}$, so the corollary for $k-1$ implies that $p|a_i$ for some i between 1 and $k - 1$. So in any case then $p|a_i$ for some i between 1 and k. □

Now we state and prove the **Fundamental Theorem of Arithmetic**:

Theorem 2.5.4 *Suppose that n is an integer greater than 1. Then there is a positive integer k and prime numbers $p_1, p_2, \ldots, p_k$ such that $n = p_1p_2\ldots p_k$. Moreover the factorization is unique, up to changing the order of the prime factors $p_1, p_2, \ldots, p_k$.*

We call such an expression for n a **prime factorization**. so the first part of the theorem says that every $n > 1$ has a prime factorization. Note that the primes p_i are allowed to repeat. Before proving the theorem, let's also clarify the meaning of the uniqueness assertion. It means that if we have two prime factorizations of n, say

$$n = p_1p_2\cdots p_k \quad \text{and} \quad n = q_1q_2\cdots q_\ell,$$

then in fact $k = \ell$ and after reordering the list of primes $q_1, q_2, \ldots, q_\ell$, we have $p_1 = q_1$, $p_2 = q_2$, $\ldots$, $p_k = q_k$. Another way to formulate this would be to require that $p_1 \le p_2 \le \cdots p_k$, and then there is exactly one such factorization.
Proof. We first note that if n is prime, then n has a prime factorization: take $k = 1$ and $p_1 = n$.

Now we prove by induction on n that if $n > 1$, then n has a prime factorization. We begin with $n = 2$, which is prime, so it has a prime factorization.

Now suppose that $n > 2$ and that the integers $2, 3, \ldots, n - 1$ all have prime factorizations. If n is prime, then it has a prime factorization, so we can assume n is not prime. If n is not prime, then n is divisible by some integer n' with $1 < n' < n$. So $n = n'n''$ for some integer $n'' = n/n'$, and $1 < n'' < n$. By the induction hypothesis, n' and n'' have prime factorizations, i.e., $n' = q_1q_2\cdots q_i$ for some $i \ge 1$ and primes $q_1, q_2, \ldots, q_i$, and $n'' = r_1r_2\cdots r_j$ for some $j \ge 1$ and primes $r_1, r_2, \ldots, r_j$. Therefore

$$n = n'n'' = q_1q_2\cdots q_ir_1r_2\cdots r_j,$$

so n has a prime factorization as well.

Now we prove the uniqueness assertion in the theorem. Again we first show that it is true if n is prime. We saw that if n is prime, then $n = p_1$ is a prime factorization of n, so we must show that it's the *only* prime factorization of n. So suppose $n = q_1 q_2 \cdots q_\ell$ with $q_1, q_2, \ldots, q_\ell$ prime. Since n is prime, $q_1 | n$ and $q_1 > 1$, we must have $q_1 = n$. This means that if $\ell > 1$, then $q_2 \cdots q_\ell = 1$, contradicting that $q_2 > 1$. So in fact $\ell = 1$ and $q_1 = n = p_1$.

Now we prove the uniqueness assertion for all $n > 1$ by induction on n. Since $n = 2$ is prime, the assertion is true for $n = 2$. Now suppose that $n > 2$ and that the prime factorizations of $2, 3, \ldots, n-1$ are unique (up to the order of the factors). We know uniqueness of the prime factorization if n is prime, so we can assume n is not prime. Suppose that

$$n = p_1 p_2 \cdots p_k = q_1 q_2 \cdots q_\ell$$

are prime factorizations of n. Since $p_1 | n$, we have $p_1 | q_1 q_2 \cdots q_\ell$. By Cor. 2.5.3 we must have $p_1 | q_i$ for some $i \in \{1, 2 \ldots, \ell\}$. reordering the q_i, we can assume $i = 1$, so $p_1 | q_1$. Since p_1 and q_1 are both prime, we must have (as above) that $p_1 = q_1$. Dividing both sides of the equation

$$n = p_1 p_2 \cdots p_k = q_1 q_2 \cdots q_\ell$$

by $p_1 = q_1$ gives $p_2 \cdots p_k = q_2 \cdots q_\ell$ as prime factorizations of n/p_1. Since n is not prime, we know that $1 < p_1 < n$, so $1 < n/p_1 < n$ and the induction hypothesis says that the prime factorization of n/p_1 is unique. This means that $k - 1 = \ell - 1$, and that after reordering $q_2, q_3, \ldots, q_k$, we have $p_i = q_i$ for $i = 2, \ldots, k$. It follows that $k = \ell$ and $p_i = q_i$ for $i = 1, 2, \ldots, k$, i.e., that the prime factorization of n is unique, after reordering the q_i if necessary. □

Note the following corollary (which uses only the existence of prime factorizations, not the uniqueness):

Corollary 2.5.5 *Suppose that m and n are integers, not both 0. Then m and n are relatively prime if and only if they have no common prime divisors.*

Proof. Recall that m and n are relatively prime if their greatest common divisor is 1. So if m and n have a common prime divisor, say p, they cannot be relatively prime. This shows that if m and n are relatively prime, then they have no common prime divisor.

Now suppose that m and n have no common prime divisor, and let $g = \gcd(m, n)$. We must show that $g = 1$. We will assume that $g > 1$ and arrive at a contradiction. If $g > 1$ by Thm. 2.5.4, g has a prime divisor p (take $p = p_1$ for example). Since $p | g$, and $g | m$ and $g | n$, we conclude that $p | m$

and $p|n$, contradicting our assumption that m and n have no common prime divisor. Therefore $g = 1$. □

For example, we can see easily that 867 and 3500 are relatively prime, without applying the Euclidean algorithm, or even finding all the prime factors of 867. We can just notice that $3500 = 10^2 \cdot 35 = 2^2 5^3 7$, and check that 867 is not divisible by these primes: 2, 5 and 7. To see that 867 is not divisible by 2 or 5, we can just look at the last digit. If 867 were divisible by 7, then so would be 860, and therefore so would be 86 (since 10 is not divisible by 7), but $86 = 12 \cdot 7 + 2$.

We can also give a criterion for one positive integer to be divisible by another in terms of their prime factorizations. Suppose that m and n are integers greater than 1. First write the prime factorization of m in the form

$$m = p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k},$$

where p_1, p_2, ..., p_k are *distinct* prime numbers (no two are the same), and r_1, r_2, ..., r_k are positive integers (for example, $3500 = 2^2 5^3 7$). Note that the primes $p_1, p_2, \ldots, p_k$ are precisely the prime divisors of m. It is clear that each p_i is a prime divisor of m; conversely if $p|m$ and p is prime, then by Cor. 2.5.3, $p|p_i$ for some i. Since p_i is prime, and $p > 1$, it follows that $p = p_i$. Similarly we can write

$$n = q_1^{s_1} q_2^{s_2} \cdots q_\ell^{s_\ell},$$

where q_1, q_2, ..., q_ℓ are the distinct prime divisors of n and s_1, s_2, ..., s_ℓ are positive integers. Now in the list $q_1, q_2, \ldots, q_\ell$ of primes dividing n, some might already be in the list $p_1, p_2, \ldots, p_k$ of primes dividing m, while others might not. Rather than distinguish between the possibilities and try to keep *extend* the list of primes $p_1, p_2, \ldots, p_k$ so that it includes all the prime divisors of n, and we would still have an expression for m of the form:

$$m = p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k},$$

where p_1, p_2, ... p_k are distinct primes, and now the r_i are allowed to be 0 (or positive). The advantage is that now we also have an expression

$$n = p_1^{s_1} p_2^{s_2} \cdots p_k^{s_k}$$

with the same primes p_1, p_2, ..., p_k and some non-negative integers s_1, s_2, ..., s_k. For example, if $m = 3500 = 2^2 5^3 7$ and $n = 504 = 2^3 3^2 7$, then combining the list of primes dividing m with those dividing n gives $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, $p_4 = 7$. We can then write both m and n in terms of these primes:

$$3500 = 2^2 3^0 5^3 7^1, \qquad 504 = 2^3 3^2 5^0 7^1.$$

In the following statement, $\min(x, y)$ denotes the minimum of x and y, i.e., the smaller of the two numbers x and y (naturally using either if $x = y$).

Corollary 2.5.6 *Suppose that $p_1, p_2, \ldots, p_k$ are distinct prime numbers and $r_1, r_2, \ldots, r_k, s_1, s_2, \ldots, s_k$ are non-negative integers. Let $m = p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k}$ and $n = p_1^{s_1} p_2^{s_2} \cdots p_k^{s_k}$.*

1. *Then $n|m$ if and only if $s_1 \leq r_1$, $s_2 \leq r_2$, $\ldots$, $s_{k-1} \leq r_{k-1}$ and $s_k \leq r_k$.*
2. *$\gcd(m, n) = p_1^{t_1} p_2^{t_2} \cdots p_k^{t_k}$ where $t_1 = \min(r_1, s_1)$, $t_2 = \min(r_2, s_2)$, $\ldots$, $t_{k-1} = \min(r_{k-1}, s_{k-1})$ and $t_k = \min(r_k, s_k)$.*

Proof. 1. Suppose that $n|m$. This means that $m = nx$ for some $x \in \mathbb{Z}$. If p is a prime dividing x, then $p|m$, so $p = p_i$ for some i. So the prime factorization of x provided by Thm. 2.5.4 has the form $x = p_1^{u_1} p_2^{u_2} \cdots p_k^{u_k}$ where $u_1, u_2, \ldots, u_k$ are non-negative integers. Now we have

$$p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k} = m = nx = p_1^{s_1+u_1} p_2^{s_2+u_2} \cdots p_k^{s_k+u_k}.$$

By the *uniqueness* assertion in Thm. 2.5.4, it follows that

$$r_1 = s_1 + u_1, \quad r_2 = s_2 + u_2, \quad \ldots, \quad r_k = s_k + u_k.$$

Since for each i we have $u_i \geq 0$, it follows that $s_i \leq r_i$ for $i = 1, 2, \ldots, k$.

Conversely if $s_i \leq r_i$ for $i = 1, 2, \ldots, k$, then letting $u_i = r_i - s_i$ gives $m = nx$, where x is the integer $p_1^{u_1} p_2^{u_2} \cdots p_k^{u_k}$. Therefore $n|m$.

2. Suppose that d is a common divisor of m and n. If p is a prime divisor of d, then p divides m (and n), so $p = p_i$ for some i. Therefore the prime factorization of d has the form

$$d = p_1^{v_1} p_2^{v_2} \cdots p_k^{v_k}$$

where $v_1, v_2, \ldots, v_k$ are non-negative integers. Since $d|m$, part 1 of the corollary says that $v_i \leq r_i$ for each $i = 1, 2, \ldots, k$. Similarly, since $d|n$, we have $v_i \leq s_i$ for each $i = 1, 2, \ldots, k$. Therefore $v_i \leq t_i = \min(r_i, s_i)$ for each i. Setting $g = p_1^{t_1} p_2^{t_2} \cdots p_k^{t_k}$, we see that $g|m$ and $g|n$, and that $d|g$ every common divisor d of m and n. Therefore $g = \gcd(m, n)$. □

Note that the corollary applies to any positive integers m and n; just choose the set of primes $p_1, p_2, \ldots, p_k$ so it includes all the prime divisors of m and all the prime divisors of n. For example, to apply the theorem to the integers $m = 3500$ and $n = 504$, we let:

- $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, $p_4 = 7$,

- $r_1 = 2$, $r_2 = 0$, $r_3 = 3$, $r_4 = 1$,
- $s_1 = 3$, $s_2 = 2$, $s_3 = 0$, $s_4 = 1$.

Since $s_1 > r_1$ (or $s_2 > r_2$), we see by part 1 of the corollary that 3500 is not divisible by 504. To compute $\gcd(m, n)$ using part 2 of the theorem, we find $t_1 = \min(2, 3) = 2$, $t_2 = \min(0, 2) = 0$, $t_3 = \min(3, 0) = 0$ and $t_4 = \min(1, 1) = 1$, so $\gcd(m, n) = 2^2 3^0 5^0 7^1 = 28$.

Chapter 3

Binary operations

3.1 Binary operations

Recall that in the very first lecture, I mentioned the notion of a group, which is a set with a binary operation having certain nice properties. I gave the example of the set of symmetries of a triangle, the operation being composition of symmetries. Another example is the set of integers with the operation of addition. (It turns out that the set of integers with the operation of multiplication is *not* a group. We'll see why later.)

Before giving the precise definition of a group by listing the "group axioms" that need to be satisfied, here is some background on binary operations.

Definition 3.1.1 Suppose that S is a set. A **binary operation** $*$ on S is a rule that assigns an element $a * b$ to each ordered pair of elements $a, b \in S$.

Here are some examples:

- The operation $+$ on $\mathbb{Z}$ (or on $\mathbb{N}$, $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$).
- The operation $\cdot$ (multiplication) on $\mathbb{Z}$ (or $\mathbb{N}$, etc.). Of course, we often omit the symbol $\cdot$ and just write ab for the product of a and b.
- The operation $-$ on $\mathbb{Z}$, but not on $\mathbb{N}$. (Why not?)
- The division operation $\div$ on the set of non-zero real numbers, but not on $\mathbb{R}$. (why not?)
- The operation of matrix addition on the set of 2×2 real matrices, denoted $M_2(\mathbb{R})$.
- The operation of matrix multiplication on $M_2(\mathbb{R})$. (Note there's no explicit symbol for the operation; we just write AB for the product.)

- The operation $*$ on $\mathbb{R}$ defined by $a * b = (a + b)/2$.
- The operation $\star$ on $\mathbb{Z}$ defined by $a \star b = a$.

- Let D_3 denote the set of symmetries of an equilateral triangle. There are 6 of these (draw the pictures...):
  - The identity; call this e.
  - Two 120° rotations; call these ρ_1 (clockwise) and ρ_2 (counterclockwise).
  - Three reflections (one axis through each vertex); call these σ_A, σ_B and σ_C (ordering the vertices clockwise).

  If α and β are symmetries, we let $\alpha \circ \beta$ denote the symmetry defined by "do β then α." Thinking of a symmetry as a "shape-preserving mapping" from the triangle back onto itself, it's clear that this is again a symmetry. This operation, called "composition," is described explicitly in the following table where the entry in the α-row and β-column is $\alpha\circ\beta$.

	e	ρ_1	ρ_2	σ_A	σ_B	σ_C
e	e	ρ_1	ρ_2	σ_A	σ_B	σ_C
ρ_1	ρ_1	ρ_2	e	σ_C	σ_A	σ_B
ρ_2	ρ_2	e	ρ_1	σ_B	σ_C	σ_A
σ_A	σ_A	σ_B	σ_C	e	ρ_2	ρ_1
σ_B	σ_B	σ_C	σ_A	ρ_1	e	ρ_2
σ_C	σ_C	σ_A	σ_B	ρ_2	ρ_1	e

- We might sometimes define a binary operation on a set S by listing its values in the form of a table like the one above. For example, let $S = \{a, b, c\}$ and define an operation $\diamond$ on S by the table:

	a	b	c
a	a	b	c
b	b	a	c
c	c	a	b,

  so for example $b \diamond c = c$.

- Let $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$ and define a "clock addition" operation as follows: $a \oplus b$ is the remainder of $a + b$ on division by 12. So for example, $7 \oplus 9 = 4$. We could similarly define a clock multiplication operation: $a \otimes b$ is the remainder of ab on division by 12. So for example $7 \otimes 9 = 3$.

- Let $\mathcal{F}$ be the set of functions from $\mathbb{R}$ to $\mathbb{R}$. If f and g are such functions, then their *composite* is the function $f \circ g$ defined by:

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

  i.e., the value of $f \circ g$ applied to $x \in \mathbb{R}$ is gotten by applying g, and then f. For example, if $f(x) = x^2$ and $g(x) = 2x$, then $(f \circ g)(x) = f(g(x)) = f(2x) = 4x^2$. Note that $(g \circ f)(x) = 2x^2$, so $f \circ g$ and $g \circ f$ are different functions.

Definition 3.1.2 Let $*$ be a binary operation on a set S.

1. We say $*$ is **commutative** if $a * b = b * a$ for all $a, b \in S$.
2. We say $*$ is **associative** if $a * (b * c) = (a * b) * c$ for all $a, b, c \in S$.

Which of the above examples are commutative?

- Addition and multiplication on $\mathbb{Z}$ (or $\mathbb{N}$, $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$) are commutative. Subtraction on $\mathbb{Z}$ is not; nor is division on $\{\, x \in \mathbb{R} \mid x \neq 0 \,\}$.

- Matrix addition on $M_2(\mathbb{R})$ is commutative; matrix multiplication is not. For example,

$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \text{ but } \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

  Thus there are elements $A, B \in M_2(\mathbb{R})$ with the property that $AB \neq BA$. (There are also pairs of elements with the property that $AB = BA$, but since the equation fails *for some* $A, B \in M_2(\mathbb{R})$, the operation is not commutative.)

- The operation $a * b = (a + b)/2$ on $\mathbb{R}$ is commutative; the operation $a \star b = a$ on $\mathbb{Z}$ is not.

- Composition of symmetries in D_3 is not commutative; for example $\rho_1\sigma_A = \sigma_C$, but $\sigma_A\rho_1 = \sigma_B$. Similarly the operation $\diamond$ on $S = \{a, b, c\}$ defined by the table above is not commutative.

- The clock addition and multiplication ($\oplus$ and $\otimes$) operations on the set $\{0, 1, 2, 3, \ldots, 11\}$ are commutative. To see this for addition, recall that $a \oplus b$ is defined as the remainder of $a + b$ on division by 12, and this is the same as the remainder of $b + a$ on division by 12. The same works for multiplication.

- The example given for composition of functions in $\mathcal{F}$ shows this operation is not commutative.

Which of the examples are associative?

- You know that addition and multiplication on $\mathbb{Z}$ (or $\mathbb{N}$, etc.) are associative, subtraction on $\mathbb{Z}$ and division on the set of non-zero real numbers are not.

- Matrix addition is associative, and you've probably done the tedious calculation that shows matrix multiplication on $M_2(\mathbb{R})$ is associative. For an *associative* operation $*$, we can omit the parentheses indicating the order for applying the operation to a sequence of elements, i.e., we can just write $a*b*c$ instead of $(a*b)*c$ or $a*(b*c)$ (since these define the same element). Note however that the order of a, b and c matters if the operation is not commutative. So for matrix multiplication, we could just write ABC instead of $(AB)C$; this is the same as $A(BC)$, but might not be the same as $BAC = (BA)C$.

- The operation $a * b = (a + b)/2$ on $\mathbb{R}$ is not associative since

$$(a * b) * c = \frac{a+b}{2} * c = \frac{(a+b)/2 + c}{2} = \frac{a}{4} + \frac{b}{4} + \frac{c}{2}$$

  is not always the same as

$$a * (b * c) = a * \frac{b+c}{2} = \frac{a + (b+c)/2}{2} = \frac{a}{2} + \frac{b}{4} + \frac{c}{4}.$$

  On the other hand, the operation $a \star b = a$ on $\mathbb{Z}$ is associative since

$$(a \star b) \star c = a \star c = a \quad \text{and} \quad a \star (b \star c) = a \star b = a.$$

- Composition of symmetries on D_3 is associative. This is because composition of functions is associative. I'll come back to this below.

- The operation $\diamond$ on $\{a, b, c\}$ defined in the table is *not* associative. You can check that $(c \diamond c) \diamond c$ is not the same as $c \diamond (c \diamond c)$!

- The operations $\oplus$ and $\otimes$ on $\{0, 1, 2, \ldots, 11\}$ are associative, but this is less obvious than their commutativity. Let's check an example for $\oplus$, say $a = 3$, $b = 7$, and $c = 9$. Then

$$a \oplus (b \oplus c) = 3 \oplus (7 \oplus 9) = 3 \oplus 4 = 7.$$

  On the other hand,

$$(a \oplus b) \oplus c = (3 \oplus 7) \oplus 9 = 10 \oplus 9 = 7.$$

  So the associativity formula holds in this particular example. You can check more examples for yourself and keep finding that it holds, for $\otimes$ as for $\oplus$. This won't *prove* that the operations are associative unless all possible combinations of values of a, b and c (of which there 12^3 in all!), but computing examples might reveal patterns and lead you to a proof that the formulas hold in general. The idea would be to show that $a \oplus (b \oplus c)$ and $(a \oplus b) \oplus c$ are both the same as the remainder of $a+b+c$ on division by 12, and similarly for multiplication. Rather than prove associativity this way though, we'll do it later using the notion of *congruences*. This will change the set-up, making the definitions more general and abstract, but in many ways easier to work with.

3.2 Composition of functions

Let's turn to the last example: the composition operation on the set $\mathcal{F}$ of functions from $\mathbb{R}$ to $\mathbb{R}$. This operation is associative; here's the proof: Suppose f, g and h are in $\mathcal{F}$. We have to show that for all real numbers x, applying the function $f \circ (g \circ h)$ to x gives the same number as applying $(f \circ g) \circ h$ to x. Indeed unravelling the definitions shows:

$$\begin{aligned} (f \circ (g \circ h))(x) &= f((g \circ h)(x)) = f(g(h(x))), \\ \text{and} \quad ((f \circ g) \circ h(x)) &= (f \circ g)(h(x)) = f(g(h(x))). \end{aligned}$$

There is nothing special about our having considered functions on $\mathbb{R}$ here. For any set A, we could consider the set S of functions from A to itself. Then composition of functions defines an associative binary operation on S. For

example, we could take A to be the set of points on an equilateral triangle. Then the set D_3 of symmetries is a *subset* of S with the property that if $f, g \in D_3$, then $f \circ g \in D_3$ (i.e., D_3 is *closed* under the operation). The formula $(f \circ g) \circ h = f \circ (g \circ h)$ still holds, so the operation is associative.

The notion of composition of functions makes sense in more generality. Instead of considering functions from a set A to itself, we can define composition as long as the domain and codomain are compatible in the following sense: if f is a function from A to B and g is a function from B to C (i.e., $f : A \to B$ and $g : B \to C$), their **composite** is the function

$$g \circ f : A \to C, \qquad \text{defined by } (g \circ f)(a) = g(f(a)) \text{ for } a \in A.$$

Note that since $f(a) \in B$, we can evaluate g at $f(a)$ to get an element $g(f(a)) \in C$. For example, if $f : \mathbb{N} \to \mathbb{R}$ is the function defined by $f(n) = \sqrt{n} - \pi$, and $g : \mathbb{R} \to \{-1, 0, 1\}$ is the function defined by

$$g(x) = \begin{cases} -1, & \text{if } x < 0, \\ 0, & \text{if } x = 0, \\ 1, & \text{if } x > 0, \end{cases}$$

then $g \circ f$ is the function defined by $(g \circ f)(n) = -1$ if $1 \leq n \leq 9$, and $(g \circ f)(n) = 1$ if $n \geq 10$.

We can describe $g \circ f$ in terms of input/output as follows: we start with input $a \in A$, apply the rule f to get an output $f(a) \in B$, and use this as the input for g to get a final output $g(f(a))$. In terms of mappings, we can think of $g \circ f$ as the map f from A to B followed by the map g from B to C, and represent it with arrows by:

$$A \xrightarrow{f} B \xrightarrow{g} C.$$

In any case, note the order in which we write the functions in the notation for composites: $g \circ f$ means "apply f, then apply g." The reason for the order is to maintain consistency with the order the functions appear in its definition as $g(f(a))$ (read "g of f of a"). While we should perhaps refer to $g \circ f$ as "f composed with g," there's a tendency to read it from left to right and call it "g composed with f." So while there may be some ambiguity in the terminology "this composed with that," there's no ambiguity about the meaning of $g \circ f$. Note that for the composite $g \circ f$ to be defined, we need the *codomain* of f to be the *same* set as the *domain* of g.

Let's clarify what it means for two functions to be the *same*. Suppose now we are given functions $f : A \to B$ and $g : C \to D$. We say f and g are **equal** and write $f = g$ if all the following hold:

- $A = C$ (the domains of f and g are the same);
- $B = D$ (the codomains of f and g are the same);
- and $f(a) = g(a)$ for all $a \in A = C$.

Now composition of functions is associative in the following sense. Suppose that $f : A \to B$, $g : B \to C$ and $h : C \to D$ are functions. We can then form the composite $g \circ f : A \to C$, and *its* composite with h is a function $h \circ (g \circ f) : A \to D$. On the other hand, we could first form the composite $h \circ g : B \to C$, and *then* the composite $(h \circ g) \circ f$, which is also a function from A to D. Unsurprisingly,

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

The proof consists simply of unravelling the definitions of all the composites. We already noted the two functions in question both have domain A and codomain D, and for all $a \in A$, we have

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

So we can omit the parentheses and just write $h \circ g \circ f$ for this function, which we can think of schematically as "combining" the three arrows in the diagram

$$A \xrightarrow{f} B \xrightarrow{g} C \xrightarrow{h} D.$$

3.3 Arithmetic modulo n

Modular arithmetic, or arithmetic mod n, provides some interesting examples of binary operations. I already gave the example of "clock arithmetic," or arithmetic mod 12 on the set $\{0, 1, 2, \ldots, 11\}$ (or equivalently $\{1, 2, 3, \ldots, 12\}$), always replacing the output of an operation with its remainder on division by 12. There's nothing special about 12 (other than a little familiarity with the operation in this context); we could in fact do arithmetic mod n with the set $\{0, 1, 2, \ldots, n-1\}$ for any positive integer n. It turns out neater though if we set things up a little more abstractly using the notion of *congruence classes* (also called *residue classes*).

Definition 3.3.1 Suppose that a and b are integers. We say that a is **congruent** to b **modulo** n if $a - b$ is divisible by n. The notation for this is $a \equiv b \pmod{n}$.

For example 43 is congruent to -5 modulo 12, or $43 \equiv -5 \bmod 12$, because the difference 48 is divisible by 12. Some more examples:

$$748374 \equiv 9833255574 \bmod 100, \qquad 7 \not\equiv 4 \bmod 2,$$

where of course $a \not\equiv b \bmod n$ means a is not congruent to b modulo n.

Note that $a \equiv b \bmod n$ if and only if $b \equiv a \bmod n$, so we also sometimes just say "a and b are congruent modulo n" (instaed of "a congruent to b modulo n" or "b congruent to a modulo n").

Proposition 3.3.2 *Suppose that a, b and n are integers and $n > 0$. Then the following are equivalent:*

(a) $a \equiv b \bmod n$;

(b) $a = b + kn$ *for some* $k \in \mathbb{Z}$;

(c) a *and* b *have the same remainder on division by* n.

Before giving the proof, let's make sure to understand the statement of the proposition. The statement is that the three assertions ((a), (b) and (c)) are equivalent. This means we have to prove that each assertion implies the others. It will be enough to prove that (a) $\Rightarrow$ (b), (b) $\Rightarrow$ (c), and (c) $\Rightarrow$ (a). (To deduce for example that (c) $\Rightarrow$ (b), note that we will have shown (c) $\Rightarrow$ (a) $\Rightarrow$ (b).)

Proof. (a) $\Rightarrow$ (b): By definition, $a \equiv b \bmod n$ means that $a - b$ is divisible by n, which means that $a - b = kn$ for some $k \in \mathbb{Z}$, so $a = b + kn$ for some $k \in \mathbb{Z}$.

(b) $\Rightarrow$ (c): For b to have remainder r on division by n means that $b = nq+r$ for some $q, r \in \mathbb{Z}$ with $0 \leq r < n$. If $a = b+kn$, then $a = (nq+r)+kn = n(q+k)+r$, so a also has remainder r on division by n.

(c) $\Rightarrow$ (a): Suppose that a and b have the same remainder, say r, on division by n. This means that $a = nq + r$ for some $q \in \mathbb{Z}$ and $b = ns + r$ for some $s \in \mathbb{Z}$. Therefore $a - b = n(q - k)$ is divisible by n, which means that $a \equiv b \bmod n$. □

Note the following immediate consequence (which is also easy to prove directly from the definition):

Corollary 3.3.3 *If* $a \equiv b \pmod n$ *and* $b \equiv c \pmod n$, *then* $a \equiv c \pmod n$.

Proof. If a and b have the same remainder on division by n, and so do b and c, then so do b and c. □

Here is another useful property of congruences:

Proposition 3.3.4 *Suppose that a, b, c, d and n are integers and $n > 0$. If $a \equiv b \bmod n$ and $c \equiv d \bmod n$, then*

1. *$a + c \equiv b + d \bmod n$, and*
2. *$ac \equiv bd \bmod n$.*

Proof. If $a \equiv b \bmod n$, then by Prop. 3.3.2, $a = b + kn$ for some $k \in \mathbb{Z}$. Similarly if $c \equiv d \bmod n$, then $c = d + jn$ for some $j \in \mathbb{Z}$. Therefore

$$a + c = b + kn + d + jn = (b + d) + (k + j)n,$$

so $a + c \equiv b + d \bmod n$ by Prop. 3.3.2 again. Similarly the formula

$$ac = (b + kn)(d + jn) = bd + (bj + kd + jkn)n.$$

shows that $ac \equiv bd \bmod n$ as well. □

Definition 3.3.5 If $n \in \mathbb{N}$ and $a \in \mathbb{Z}$, then we call the set

$$\{\, b \in \mathbb{Z} \,|\, b \equiv a \pmod{n} \}$$

the **congruence** (or **residue**) **class** of a **modulo** n, and denote it by $[a]_n$.

So for example

$$[5]_{12} = \{\, b \in \mathbb{Z} \,|\, b \equiv 5 \bmod 12 \,\} = \{\ldots, -19, -7, 5, 17, \ldots\}.$$

Note that in general

$$[a]_n = \{\, a + kn \,|\, k \in \mathbb{Z}\}$$

by Prop. 3.3.2. Note also the following:

Proposition 3.3.6 *Suppose $a, b \in \mathbb{Z}$, $n \in \mathbb{N}$. Then*

$$a \equiv b \bmod n \Leftrightarrow [a]_n = [b]_n.$$

Proof. $\Rightarrow$: Suppose that $a \equiv b \bmod n$, If $c \in [a]_n$, then $c \equiv a \bmod n$, so Cor. 3.3.3 implies that $c \equiv b \bmod n$, which means that $c \in [b]_n$. Similarly, if $c \in [b]_n$, then $c \in [a]_n$. Therefore $[a]_n = [b]_n$.

$\Leftarrow$: Suppose that $[a]_n = [b]_n$. Then $a \in [a]_n$ (since $a \equiv a \bmod n$), so $a \in [b]_n$ (since $[a]_n = [b]_n$), so $a \equiv b \bmod n$ (by definition of $[b]_n$). □

According to the Division Algorithm (Thm. 2.1.2), we can always write $a = nq + r$ for some $q, r \in \mathbb{Z}$ with $0 \leq r \leq n - 1$. Then $a \equiv r \bmod n$, so by Prop. 3.3.6, $[a]_n = [r]_n$. Therefore every congruence class $[a]_n$ modulo n can

be written in the form $[r]_n$ for some r between 0 and $n-1$. For example, $[-1000]_7 = [1]_7$ since $-1000 \equiv 1 \bmod 7$. So even though there are infinitely many integers a, they define only finitely many congruence classes modulo n, namely

$$[0]_n, [1]_n, \ldots, [n-1]_n.$$

Furthermore, it's easy to see these n congruence classes are different since if r and s are between 0 and $n-1$, and $[r]_n = [s]_n$, then $r \equiv s \bmod n$, so $r - s$ is divisible by n, but $r - s$ is between $1 - n$ and $n - 1$, so the only way it can be divisible by n is if $r - s = 0$, i.e., $r = s$.

Now we let $\mathbb{Z}_n$ denote the *set* of congruence classes modulo n, so:

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

is a set with n elements; each *element* of $\mathbb{Z}_n$ is itself a a *set* of infinitely many integers. So for example

$$\mathbb{Z}_3 = \{\, [0]_3, [1]_3, [2]_3 \,\},$$

where

$$\begin{aligned} [0]_3 &= \{\ldots, -6, -3, 0, 3, 6, 9, 12, \ldots\}, \\ [1]_3 &= \{\ldots, -5, -2, 1, 4, 7, 10, 13, \ldots\}, \\ \text{and } [2]_3 &= \{\ldots, -4, -1, 2, 5, 8, 11, 14, \ldots\}. \end{aligned}$$

We are now going to define binary operations, namely an *addition* and a *multiplication*, on $\mathbb{Z}_n$. We begin with addition. We'd like to define the sum of two congruence classes, say $[a]_n$ and $[b]_n$, by the formula

$$[a]_n + [b]_n = [a + b]_n$$

(the + inside the brackets being the usual addition of *integers*), but there could be a problem with this. To define the binary operation + on the set $\mathbb{Z}_n$, we have to define the element $X + Y \in \mathbb{Z}_n$ for each $X, Y \in \mathbb{Z}_n$. But for any given X, there are infinitely many integers a such that $X = [a]_n$, and similarly for Y (using capital letters here to remind us that X and Y are in fact *sets*). For example, our proposed definition says that $[3]_{10} + [8]_{10} = [3+8] = [11]_{10}$, but $[3]_{10}$ is the *same* congruence class modulo 10 as $[-17]_{10}$, so it had better be true that $[3]_{10} + [8]_{10} = [-17]_{10} + [8]_{10}$. And indeed it is, since $[11]_{10} = [-9]_{10}$. This was just an example. We need to check that *whenever* $[a]_n = [a']_n$, our formula gives the same value for the sum $[a']_n + [b]_n$ as for the sum $[a]_n + [b]_n$. Similarly we could have chosen another integer in the congruence class of b to compute the sum of $[a]_n$ and $[b]_n$, and the result shouldn't depend on this choice. So to check that the sum is *well-defined*, we have to check that if

$[a]_n = [a']_n$ and $[b]_n = [b']_n$, then $[a+b]_n = [a'+b']_n$. Similarly, we'd like to define the product of $[a]_n$ and $[b]_n$ by the formula

$$[a]_n[b]_n = [ab]_n,$$

but for this to be well-defined, we need to check that if $[a]_n = [a']_n$ and $[b]_n = [b']_n$, then $[ab]_n = [a'b']_n$. Fortunately, we essentially proved this already in Prop. 3.3.4.

Proposition 3.3.7 *Suppose that $n \in \mathbb{N}$ and $a, b, a', b' \in \mathbb{Z}$. If $[a]_n = [a']_n$ and $[b]_n = [b']_n$, then $[a+b]_n = [a'+b']_n$ and $[ab]_n = [a'b']_n$.*

Proof. If $[a_n] = [a']_n$ and $[b]_n = [b']_n$, then $a \equiv a' \bmod n$ and $b \equiv b' \bmod n$ by Prop. 3.3.2. So by Prop. 3.3.4,

$$a + b \equiv a' + b' \bmod n \quad \text{and} \quad ab \equiv a'b' \bmod n,$$

and this implies that $[a+b]_n = [a'+b']_n$ and $[ab]_n = [a'b']_n$. □

We can now define our binary operations on $\mathbb{Z}/\langle n\rangle$. For $X, Y \in \mathbb{Z}/\langle n\rangle$, choose any $a, b \in \mathbb{Z}$ so that $X = [a]_n$ and $Y = [b]_n$. We then define

$$X + Y = [a+b]_n \quad \text{and} \quad XY = [ab]_n.$$

According to the proposition, the congruence classes we just defined are independent of the choices of a and b in the congruence classes X and Y.

Suppose for example that $n = 10$, $X = \{\ldots, -17, -7, 3, 13, 23, \ldots\}$ and $Y = \{\ldots, -12, -2, 8, 18, 28, \ldots, \ldots,\}$ and let's compute $X + Y$ and XY. Since $X = [3]_{10}$ and $Y = [8]_{10}$, we have

$$\begin{aligned} X + Y = [3]_{10} + [8]_{10} &= [11]_{10} = [1]_{10} \\ \text{and} \quad XY = [3]_{10}[8]_{10} &= [24]_{10} = [4]_{10}. \end{aligned}$$

As you can see, if we systematically choose to express the congruence classes in the form $[r]_n$ with $r \in \{0, 1, \ldots, n-1\}$, this works just like our "clock arithmetic" where we defined the binary operation on the set $\{0, 1, \ldots, n-1\}$ (in the case $n = 12$). That set is defined very concretely as a set of integers, and so easier to grasp conceptually. On the other hand, the set $\mathbb{Z}_n$ of congruence classes is defined more abstractly; each element is itself a set of integers. The advantage of the more abstract definition is that it is often easier to work with; for example the proof of associativity is much simpler.

Proposition 3.3.8 *The addition and multiplication operations on $\mathbb{Z}_n$ are commutative and associative.*

Proof. We just give the proof for addition since the proofs for multiplication work in exactly the same way.

For commutativity, we must show that for all $X, Y \in \mathbb{Z}_n$, we have $X+Y = Y+X$. Choose $a, b \in \mathbb{Z}$ so $X = [a]_n$ and $Y = [b]_n$. Then

$$X + Y = [a]_n + [b]_n = [a+b]_n = [b+a]_n = [b]_n + [a]_n = Y + X.$$

For associativity, we must show that for all $X, Y, Z \in \mathbb{Z}_n$, we have $(X + Y) + Z = X + (Y + Z)$. Choose $a, b, c \in \mathbb{Z}$ so that $X = [a]_n$, $Y = [b]_n$ and $Z = [c]_n$. Then by the definition of addition of $\mathbb{Z}_n$, we have

$$\begin{aligned} & (X+Y)+Z &&= ([a]_n + [b]_n) + [c]_n \\ & &&= [a+b]_n + [c]_n = [(a+b)+c]_n \\ \text{and} \quad & X+(Y+Z) &&= [a]_n + ([b]_n + [c]_n) \\ & &&= [a]_n + [b+c]_n = [a+(b+c)]_n, \end{aligned}$$

and these are the same since $a+(b+c) = (a+b)+c$ by the associativity law for the usual addition of integers. □

Chapter 4

Groups

4.1 Definition of a group

I've already mentioned some examples of groups; now I'll give the precise definition of a *group* by listing the properties that a set with a binary operation needs to satisfy in order to be called a group.

Definition 4.1.1 A **group** is a set G with a binary operation $*$ satisfying the following properties:

1. $*$ is associative;
2. there is an element $e \in G$ such that $e * g = g * e = g$ for all $g \in G$;
3. if $g \in G$, then there is an element $h \in G$ such that $g * h = h * g = e$.

Before considering examples, here are some remarks about the properties:

1. You might also have seen the definition of a *group* include a *closure axiom*, stating that if $g, h \in G$, then $g*h \in G$. This is already included in the definition I gave (Defn. 3.1.1) for $*$ to be a *binary operation* on G.
2. Recall that property 1), **associativity**, means that $g*(h*k) = (g*h)*k$ for all $g, h, k \in G$.
3. An element e as in 2) is called an **identity element** for $*$. It's easy to see that a set with a binary operation can have at most one identity element. Indeed if e and e' are both identity elements, then $e * e' = e'$ (since e is an identity element), and $e * e' = e'$ since e' is an identity element, so $e = e * e' = e'$.

4. An element h in 3) is called an **inverse** of g (with respect to $*$). Note that h depends on g, but we'll see that if G is a group, then each $g \in G$ has exactly one inverse.

4.2 Examples of groups

Example 4.2.1 The set $\mathbb{Z}$ with the operation $+$ is a group since 1) $+$ is associative, 2) 0 is an identity element since

$$0 + n = n + 0 = n \quad \forall n \in \mathbb{Z},$$

and 3) if $n \in \mathbb{Z}$, then $-n \in \mathbb{Z}$ is an inverse of n since

$$n + (-n) = (-n) + n = 0.$$

We'll usually just write $(G, *)$ instead of "G with the operation $*$." So the preceding example would be denoted $(\mathbb{Z}, +)$. We'll even often omit $*$ when it's clear from the context.

Example 4.2.2 $(\mathbb{Z}, \cdot)$ is not a group. 1) and 2) are satisfied (the identity element being 1), but not every element of $\mathbb{Z}$ has an inverse with respect to multiplication. For example, there is no integer n such that $2n = 1$. (In fact the only elements with inverses are ± 1.)

Example 4.2.3 Let's try to fix the last example by considering $(\mathbb{R}, \cdot)$ instead of $(\mathbb{Z}, \cdot)$. Now 2 has an inverse, namely 1/2. But no, it's still not a group; 0 has no inverse.

Example 4.2.4 Let's try again. Let $\mathbb{R}^\times = \mathbb{R} \setminus \{0\}$ be the set of non-zero real numbers[1]. Note that the product of two elements of $\mathbb{R}^\times$ is again in $\mathbb{R}^\times$, and now 1), 2) and 3) are satisfied, so $(\mathbb{R}^\times, \cdot)$ is a group.

Example 4.2.5 $(M_2(\mathbb{R}), +)$ is a group. Recall $M_2(\mathbb{R})$ denotes the set of 2×2 real matrices. 1) Matrix addition is associative, 2) the matrix $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ is an identity element, and 3) the (additive) inverse of $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is $-A = \begin{pmatrix} -a & -b \\ -c & -d \end{pmatrix}$.

[1] See Section 1.2 for the $\setminus$ notation

Example 4.2.6 $M_2(\mathbb{R})$ is *not* a group under matrix multiplication. The operation is associative and has the identity element $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, but there are matrices without (multiplicative) inverses, for example $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$.

Example 4.2.7 Let $\mathrm{GL}_2(\mathbb{R}) = \{ A \in M_2(\mathbb{R}) \mid \det A \neq 0 \}$. Recall that if $A \in M_2(\mathbb{R})$, then $\det A$ (the determinant of A) is defined by the formula $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$. The determinant has the following important properties:

- $\det(AA') = (\det A)(\det A')$ for all $A, A' \in M_2(\mathbb{R})$. (The proof, which you've probably seen before, is left as an exercise.)
- If $\det A \neq 0$, then A has an inverse matrix. Proof: If $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $ad - bc \neq 0$, then setting

$$B = (ad - bc)^{-1} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \begin{pmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{pmatrix}$$

  gives $AB = BA = I$.

The first property shows that $\mathrm{GL}_2(\mathbb{R})$ is *closed* under matrix multiplication, i.e., if $A, A' \in \mathrm{GL}_2(\mathbb{R})$, then $AA' \in \mathrm{GL}_2(\mathbb{R})$. Note also that if $A \in \mathrm{GL}_2(\mathbb{R})$ and B is its inverse matrix, then

$$(\det A)(\det B) = \det(AB) = \det I = 1,$$

so $\det B$ is also non-zero. So every matrix in $\mathrm{GL}_2(\mathbb{R})$ has a (multiplicative) inverse in $\mathrm{GL}_2(\mathbb{R})$. Thus $\mathrm{GL}_2(\mathbb{R})$ is a group.

Example 4.2.8 Let D_3 denote the set of symmetries of an equilateral triangle. There are 6 of these (draw the pictures...):

- The identity; call this e.
- Two 120° rotations; call these ρ_1 (clockwise) and ρ_2 (counterclockwise).
- Three reflections (one axis through each vertex); call these ϕ_1, ϕ_2 and ϕ_3 (ordering the vertices clockwise).

If α and β are symmetries, we let $\alpha \circ \beta$ denote the symmetry defined by "do β then α." Thinking of a symmetry as a "shape-preserving mapping" from the triangle back onto itself, it's clear that this is again a symmetry and that the resulting binary operation (composition) is associative. The operation is described explicitly in the following table where the entry in the α-row and β-column is $\alpha \circ \beta$.

	e	ρ_1	ρ_2	ϕ_1	ϕ_2	ϕ_3
e	e	ρ_1	ρ_2	ϕ_1	ϕ_2	ϕ_3
ρ_1	ρ_1	ρ_2	e	ϕ_3	ϕ_1	ϕ_2
ρ_2	ρ_2	e	ρ_1	ϕ_2	ϕ_3	ϕ_1
ϕ_1	ϕ_1	ϕ_2	ϕ_3	e	ρ_1	ρ_2
ϕ_2	ϕ_2	ϕ_3	ϕ_1	ρ_2	e	ρ_1
ϕ_3	ϕ_3	ϕ_1	ϕ_2	ρ_1	ρ_2	e

In particular, note that e is the identity element, and each element has an inverse since $\rho_1 \circ \rho_2 = \rho_2 \circ \rho_1 = e$ and $\phi \circ \phi = e$ for each reflection ϕ. Note also that each element of the group appears exactly once in each row, and once in each column. We'll see this is always the case in the "multiplication table" for a group.

We could just as well consider symmetry groups of other geometric objects; for example, the set of symmetries of a square (see the exercises), the binary operation always being composition. This group, denoted D_4 has 8 elements. More generally, the symmetry group of a regular n-sided polygon (or n-gon, $n \geq 3$) is denoted D_n; it has $2n$ elements, of which n are rotations and n are rotations. We can also consider symmetry groups of objects in 3 (or more) dimensions; for example, we'll see that the symmetry group of a cube has 48 elements. Note also that symmetry groups, such as that of a circle, can be infinite.

Example 4.2.9 $(\mathbb{Z}_n, +)$ is a group. Recall that

$$\mathbb{Z}_n = \{\, [a]_n \mid a \in \mathbb{Z} \,\} = \{\, [0]_n, [1]_n, \ldots, [n-1]_n \,\}$$

is the set of congruence classes modulo n, and addition modulo n is defined by $[a]_n + [b]_n = [a+b]_n$. We saw that the operation is associative, it's clear that $[0]_n$ is an identity element, and the inverse of $[a]_n$ is $[-a]_n$.

Example 4.2.10 $(\mathbb{Z}_n, \cdot)$ is not a group (unless $n = 1$). The operation is associative and 1 is an identity element, but 0 has no inverse (unless $n = 1$ in which case 0 is the only element, so it's the identity and its own inverse). We could try to fix this up, like we did for multiplication on $\mathbb{R}$ and $M_2(\mathbb{R})$, by just working with elements that have multiplicative inverses. Note that for $[a]_n$ to have a multiplicative inverse in $\mathbb{Z}_n$ means there's a congruence class $[b]_n \in \mathbb{Z}_n$ such that $[a]_n \cdot [b]_n = [1]_n$.

Proposition 4.2.11 *Suppose $a, n \in \mathbb{Z}$ with $n \geq 1$. Then $[a]_n$ has a multiplicative inverse in $\mathbb{Z}_n$ if and only if a and n are relatively prime.*

Proof. $\Rightarrow$: Suppose that $[a]_n$ has an inverse. Then

$$[ab]_n = [a]_n \cdot [b]_n = [1]_n$$

for some $b \in \mathbb{Z}$. This means that $ab \equiv 1 \bmod n$, which means that $ab-1 = nk$ for some $k \in \mathbb{Z}$. Therefore $ab+n(-k) = 1$, so $ax+ny = 1$ for some $x, y \in \mathbb{Z}$. Cor. 2.5.5 therefore implies that a and n are relatively prime.

$\Leftarrow$: Suppose that a and n are relatively prime. Then by Cor. 2.5.5, $ax + ny = 1$ for some $x, y \in \mathbb{Z}$. Therefore $ax \equiv 1 \bmod n$, so

$$[a]_n[x]_n = [x]_n[a]_n = [1]_n,$$

showing that $[a]_n$ has a multiplicative inverse modulo n. □

Also note the following:

Proposition 4.2.12 *Suppose that $a, b, n \in \mathbb{Z}$ with $n > 0$. If $a \equiv b \bmod n$, then $\gcd(a, n) = \gcd(b, n)$.*

Proof. If $a \equiv b \bmod n$, then $a = b + nk$ for some $k \in \mathbb{Z}$. If d is a common divisor of a and n, then also $d|b = a - nk$, so d is a common divisor of b and n. Similarly if d is a common divisor of b and n, then it is a common divisor of b and n. Since the common divisors of a and n are the same as the common divisors of b and n, it follows that $\gcd(a, n) = \gcd(b, n)$. □

We are now ready to define

$$\mathbb{Z}_n^\times = \{\, [a]_n \in \mathbb{Z}_n \mid \gcd(a, n) = 1 \,\},$$

so $\mathbb{Z}_n^\times$ is the set of congruence classes $[a]_n$ whose elements are relatively prime to n. (Note that we can replace a by any b congruent to a modulo n and this won't change the congruence class $[b]_n = [a]_n$. Either of the two preceding propositions shows that the condition $\gcd(a, n) = 1$ in the definition of $\mathbb{Z}_n^\times$ depends only on $[a]_n$, not on the choice of integer a in the congruence class. Some examples are

$$\mathbb{Z}_6^\times = \{1, 5\}, \quad \text{and} \quad \mathbb{Z}_7^\times = \{1, 2, 3, 4, 5, 6\}.$$

Note that if $n = p$ is prime, then $\mathbb{Z}_p^\times = \{[1]_p, [2]_p, \ldots, [p-1]_p\}$.

Proposition 4.2.13 *$(\mathbb{Z}_n^\times, \cdot)$ is a group.*

Proof. First we have to check that $\cdot$ defines a binary operation on $\mathbb{Z}_n^\times$, i.e., if $[a]_n, [b]_n \in \mathbb{Z}_n^\times$, then $[ab]_n = [a]_n[b]_n \in \mathbb{Z}_n^\times$ (not just in $\mathbb{Z}_n$). So we have to check that if $a, b \in \mathbb{Z}$ are relatively prime to n, then so is ab. (This was an exercise.)

We already saw that $\cdot$ is associative. Its identity element is $[1]_n$, which is in $\mathbb{Z}_n^\times$ since $\gcd(1, n) = 1$. Finally Prop. 4.2.11 shows that if $[a]_n \in \mathbb{Z}_n^\times$, then $[a]_n$ has a multiplicative inverse $[b]_n \in \mathbb{Z}_n$. To complete the proof that $\mathbb{Z}_n^\times$ is a group, we just need to check that $[b]_n \in \mathbb{Z}_n^\times$, but this is immediate from (the pther direction of) Prop. 4.2.11. Indeed $[b]_n$ has a multiplicative inverse $[a]_n$, so $[b]_n \in \mathbb{Z}_n^\times$. □

4.3 Permutation groups

Suppose that A is a set and f and g are functions from A to A. Then their composite $f \circ g$ is also a function from A to A. So if we define $\mathcal{F}_A$ as the set of functions from A to A, then composition defines a binary operation on $\mathcal{F}_A$. Is $\mathcal{F}_A$ a group? Well we know composition is associative (see end of Section 3.2). We also have an identity element:

Definition 4.3.1 For any set A, we define the **identity function** on A as the function

$$\mathrm{id}_A : A \to A, \quad \text{where } \mathrm{id}_A(a) = a \text{ for all } a \in A.$$

Thus identity functions have the following property: if f is any function from A to a set B, then $f \circ \mathrm{id}_A = f$, and $\mathrm{id}_B \circ f = f$. This is clear from the definitions; for example, $\mathrm{id}_B \circ f$ is also function from A to B, and $(\mathrm{id}_B \circ f)(a) = \mathrm{id}_B(f(a)) = f(a)$ for all $a \in A$. In particular, taking $A = B$, we see that id_A is an identity element for the composition operation on $\mathcal{F}_A$.

Now for $\mathcal{F}_A$ to be a group, we would just need every element $f \in \mathcal{F}_A$ to have an inverse with respect to composition. But not every element does. An inverse would have to be a function $g : A \to A$ such that $g \circ f = f \circ g = \mathrm{id}_A$; in other words, g would be an *inverse function* of f.

Definition 4.3.2 Suppose that f is a function from A to B. We say that a function $g : B \to A$ is an **inverse function** of f if

$$g \circ f = \mathrm{id}_A \quad \text{and} \quad f \circ g = \mathrm{id}_B.$$

Note that the condition $g \circ f = \mathrm{id}_A$ means that $g(f(a)) = a$ for all $a \in A$; and the condition $f \circ g = \mathrm{id}_B$ means that $f(g(b)) = b$ for all $b \in B$. For

example, if $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = 2x + 1$, then $g : \mathbb{R} \to \mathbb{R}$ defined by $g(y) = (y-1)/2$ is an inverse function of f since $g(f(x)) = x$ for all $x \in \mathbb{R}$ and and $f(g(y)) = y$ for all y in $\mathbb{R}$.

In the preceding example, A and B were the same set, but A and B can be different. For example, let $A = \mathbb{R}$ and let B be the set of positive real numbers. Then the function $f : A \to B$ defined by $f(x) = e^x$ (an element of B since $e^x > 0$) has an inverse function $g : B \to A$, namely $g(y) = \ln y$, since $g(f(x)) = \ln(e^x) = x$ for all real numbers x and $f(g(y)) = e^{\ln y} = y$ for all *positive* real numbers y.

We can think of f and g in terms of the diagram

$$A \underset{g}{\overset{f}{\rightleftarrows}} B,$$

where each of the two functions f and g reverses what the other does. So if $f(a) = b$, then $g(b) = g(f(a)) = a$, and if $g(b) = a$, then $f(g(b)) = \mathrm{id}_B(b) = b$.

We will characterize the functions that have inverses, but first recall the definitions:

Definition 4.3.3 Suppose that f is a function from A to B. We say f is **injective** if it has the following property:

$$a, a' \in A,\ f(a) = f(a') \quad \Rightarrow \quad a = a'.$$

We say f is **surjective** if it has the following property:

$$b \in B \quad \Rightarrow \quad b = f(a) \text{ for some } a \in A.$$

A function $f : A \to B$ is **bijective** if it is both injective and surjective.

So f is injective (or *one-to-one*) if no two distinct elements of A are assigned the same value in B. For f to be surjective (or *onto*) means that its range

$$f(A) = \{\, b \in B \,|\, b = f(a) \text{ for some } a \in A \,\} = \{\, f(a) \,|\, a \in A \,\}$$

is all of B. For example, the function $f : \mathbb{Z} \to \mathbb{Z}$ defined by $f(n) = 2n + 1$ is injective since if $m, n \in \mathbb{Z}$ and $2m + 1 = 2n + 1$, then $m = n$. But f is not surjective since its range is the set of odd integers. The function $g : \mathbb{Z} \to \mathbb{Z}$ defined by

$$g(n) = \begin{cases} n/2, \text{if } n \text{ is even,} \\ (n-1)/2, \text{if } n \text{ is odd,} \end{cases}$$

is surjective but not injective. Note that $g \circ f$ is the identity function on $\mathbb{Z}$, but f and g are not inverse functions since $f \circ g$ is not the identity (since for example, $f(g(0)) = 1$).

Composition of functions preserve injectivity and surjectivity:

Proposition 4.3.4 *Suppose that $f : A \to B$ and $g : B \to C$ are functions.*

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

Proof. 1) Assuming f and g are injective, we need to show that if $a, a' \in A$ and $(g \circ f)(a) = (g \circ f)(a')$, then $a = a'$. Now

$$\begin{array}{lcll} (g \circ f)(a) = (g \circ f)(a') & \Rightarrow & g(f(a)) = g(f(a')) & \text{(by definition of } g \circ f) \\ & \Rightarrow & f(a) = f(a') & \text{(since } g \text{ is injective)} \\ & \Rightarrow & a = a' & \text{(since } f \text{ is injective).} \end{array}$$

Therefore $g \circ f$ is injective.

2) Suppose next that f and g are surjective. We need to show that if $c \in C$, then $b = (g \circ f)(a)$ for some $a \in A$. Since g is surjective, we know that $c = g(b)$ for some $b \in B$. Since f is surjective, we know that this $b = f(a)$ for some $a \in A$. So $(g \circ f)(a) = g(f(a)) = g(b) = c$. Therefore $g \circ f$ is surjective.

3) This is immediate from parts 1) and 2). □

We are now ready to describe which functions have inverse functions.

Proposition 4.3.5 *Suppose that $f : A \to B$ is a function. Then f has an inverse function if and only if f is bijective.*

Proof. We must prove 1) if f has an inverse function, then f is bijective, and 2) if f is bijective, then f has an inverse function.

1) Suppose that f has an inverse function, say $g : B \to A$, so $g(f(a)) = a$ for all $a \in A$ and $f(g(b)) = b$ for all $b \in B$. We must show that f is bijective; i.e., that it is both injective and surjective.

First we prove that f is injective. For $a, a' \in A$, we have

$$f(a) = f(a') \quad \Rightarrow \quad g(f(a)) = g(f(a')) \quad \Rightarrow \quad a = a'.$$

Therefore f is injective.

Now we prove f is surjective. Suppose that $b \in B$. We must show that $b = f(a)$ for some $a \in A$. Let $a = g(b)$. Then

$$f(a) = f(g(b)) = b.$$

Therefore f is surjective, so we conclude that f is bijective.

2) Now suppose that f is bijective. Recall that this means for each $b \in B$, there is a unique $a \in A$ such that $f(a) = b$. We want to show that f has an inverse function $g : B \to A$. We define $g : B \to A$ as follows: For $b \in B$, we let $g(b)$ be the *unique element* $a \in A$ such that $f(a) = b$. This is a function from B to A, since for each $b \in B$, we have specified a single value $a = g(b) \in A$. We now show that g is an inverse function of f. We must again show two things: that $g(f(a)) = a$ for all $a \in A$, and that $f(g(b)) = b$ for all $b \in B$.

Suppose first that $b \in B$. Then $f(g(b)) = f(a)$ where $a = g(b)$ is, by the definition of the function g, the unique element of A such that $f(a) = b$. So $f(g(b)) = f(a) = b$.

Now suppose that $a \in A$. We must show that $g(f(a)) = a$. We just saw that $f(g(b)) = b$ for all $b \in B$. Applying this to $b = f(a)$ gives $f(g(f(a))) = f(a)$. Thus letting $a' = g(f(a))$, we have $f(a') = f(a)$, and since f is assumed to be injective, this implies that $a = a'$, i.e., $g(f(a)) = a$. Therefore g is an inverse function of f. □

We are now ready to define *symmetric groups*, or *permutation groups*. Let A be any set. We'll mainly be concerned with the case where A is a finite set, but we'll start out working in more generality. We define S_A to be the set of bijective functions from A to A.

Proposition 4.3.6 *If A is a set, then S_A is a group under $\circ$.*

Proof. First we have to check that if $f, g \in S_A$, then $f \circ g \in S_A$, so that $\circ$ is indeed a binary operation on S_A. By definition, the composite $f \circ g$ is again a function from A to A, and it is bijective by Part 3 of Prop. 4.3.4.

Now we need to check S_A with the binary operation $\circ$ satisfies the group axioms. We already know that $\circ$ is associative. we also already saw that the identity function id_A on A satisfies

$$\mathrm{id}_A \circ f = f \circ \mathrm{id}_A = f$$

for all functions $f : A \to A$, so in particular for all $f \in S_A$. Note also that id_A is bijective, so $\mathrm{id}_A \in S_A$ is an identity element for the operation $\circ$. Finally Prop. 4.3.5 shows that f has an inverse function $g : A \to A$. We have to check that $g \in S_A$, i.e., that g is bijective. One way to see that it is bijective is to note that it has an inverse function, namely f, and apply Prop. 4.3.5 again.

Thus $(S_A, \circ)$ satisfies the group axioms. □

The group S_A under $\circ$ is called the **symmetric group**, or **permutation group**, on A, and its elements are called **permutations** of A. Now let's suppose that A is finite. Assume even more specifically, that $A = \{1, 2, \ldots, n\}$

where n is a positive integer. Rather than write $S_{\{1,2,\ldots,n\}}$, we write simply S_n, and call S_n the n^{th} symmetric group. We write e, instead of id_A, for the identity element of S_n, and generally omit the $\circ$ when writing composites.

There are two standard ways of denoting elements of S_n. One of these is to write

$$\begin{pmatrix} 1 & 2 & \cdots & n \\ a_1 & a_2 & \cdots & a_n \end{pmatrix}$$

for the function (or permutation) $\sigma \in S_n$ such that $\sigma(1) = a_1$, $\sigma(2) = a_2$, $\ldots$, $\sigma(n) = a_n$. Note that $a_1, a_2, \ldots, a_n$ must be a reordering, or permutation, of the integers $1, 2, \ldots, n$. So for example

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 5 & 6 & 1 & 4 \end{pmatrix}$$

is the element of $\sigma \in S_6$ such that $\sigma(1) = 2$, $\sigma(2) = 3$, $\sigma(3) = 5$, $\sigma(4) = 6$, $\sigma(5) = 1$ and $\sigma(6) = 4$.

This notation makes it easy to count the elements of S_n. Note that there are n possibilities for a_1, and having chosen a_1, there remain $n-1$ possibilities for a_2 (as it can't equal a_1), $n-2$ possibilities for a_3, $\ldots$, 2 possibilities for a_{n-1} and only 1 remaining for a_n. Therefore there are

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

elements of S_n. For example, the $3! = 6$ elements of S_3 are

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix},$$
$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix},$$
$$\begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}.$$

Here is an example to illustrate how to compute composites with this notation:

Example 4.3.7 Suppose that σ is the element of S_6 described above, and

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 4 & 1 & 5 & 6 \end{pmatrix}.$$

To compute $\sigma\tau$, we write τ above

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 5 & 6 & 1 & 4 \end{pmatrix},$$

and chase through what happens to each integer under *first* τ, and *then* σ, giving

$$\sigma\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 5 & 6 & 2 & 1 & 4 \end{pmatrix}.$$

The other standard notation is *cycle notation*. If $a_1, a_2, \ldots, a_k$ are *distinct* elements of $\{1, 2, \ldots, n\}$ (so $k \leq n$), we write

$$(a_1 a_2 a \cdots a_k)$$

for the permutation σ such that $\sigma(a_1) = a_2$, $\sigma(a_2) = a_3$, $\ldots$, $\sigma(a_{k-1}) = a_k$ and $\sigma(a_k) = a_1$, and $\sigma(b) = b$ otherwise. In other words, σ cycles through the elements $a_1, a_2, \ldots, a_k$ (going from a_k back to the beginning a_1), and leaves the remaining elements of $\{1, 2, \ldots, n\}$ alone. So for example the element $(1234) \in S_6$ is the permutation above we denoted by τ. An element of S_n of the form $(a_1 a_2 \cdots a_k)$ is called a k-**cycle** (or just a **cycle**). For example, the permutation denoted τ in Example 4.3.7 is the 4-cycle (1234) in S_6, since

$$\tau(1) = 2,\ \tau(2) = 3,\ \tau(3) = 4,\ \tau(4) = 1,$$

and $\tau(b) = b$ otherwise (i.e., $b = 5$ or 6). Every element of S_3 is a cycle (the identity can be viewed as the 1-cycle (1) for example). Maintaining the order in which the elements are listed above, we have

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

If $n > 3$, then S_n has elements which are not cycles. A general fact is that every element can be written as a product of disjoint cycles, i.e., in the form

$$(a_1 a_2 \cdots a_{k_1})(a_{k_1+1} a_{k_1+2} \cdots a_{k_1+k_2}) \cdots (a_{k_1+k_2+\cdots+k_{r-1}+1} \cdots a_{k_1+k_2+\cdots+k_r}),$$

where $a_1, a_2, \ldots, a_{k_1+k_2+\cdots+k_r}$ are distinct. We won't prove this, but here's how it works in practice: Take an element $a \in \{1, 2, \ldots, n\}$ and list the values $a_1 = a, a_2 = \sigma(a_1), a_3 = \sigma(a_2), \ldots$ until we get back to a. If k is the least integer such that $\sigma(a_k) = a$, then one of the cycles in the expression for σ is $(a_1 a_2 \cdots a_k)$, whose effect is described by:

$$a_1 \to a_2 \to \cdots \to a_k \to a_1.$$

Once an element $b \in \{1, 2, \ldots, n\}$ appears in such a list, there's no need to compute the values with $b_1 = b$, $b_2 = \sigma(b_1)$, etc., since we have already found the cycle in which it appears.

Let's carry this out for the cycle σ of Example 4.3.7. Starting from $a = 1$ gives:

$$1 \to 2 \to 3 \to 5 \to 1,$$

yielding the 4-cycle (1235). We have now accounted for 1, 2, 3 and 5, and starting from 4 gives (46). So a cycle expression for σ is

$$(1235)(46).$$

Note that we can "cycle" through the entries in a cycle without changing it, so we could have written, say (3512) instead of (1235). It's also the case that we can write *disjoint* cycles in any order without altering their composite. In other words, disjoint cycles "commute." So we could just as well write the cycle expression for σ as (46)(1235). For the sake of being systematic though, I've been writing the cycles starting with the smallest integer whenever there's a choice.

As an example of computing a composite in cycle notation, let's redo the calculation of $\sigma\tau$ in Example 4.3.7. We've already seen that $\sigma = (1235)(46)$ and $\tau = (1234)$, so we want to compute

$$(1235)(46)(1234).$$

Beginning with 1, we find that (1234) sends 1 to 2. Then 2 is not affected by (46), and finally (1235) sends 2 to 3, or more visually:

$$1 \overset{(1234)}{\longrightarrow} 2 \overset{(46)}{\longrightarrow} 2 \overset{(1235)}{\longrightarrow} 3.$$

Note that we start with the *rightmost* cycle (1234) and "move" to the left since that is the convention for composition of functions, but *within each cycle*, the entries "move" from left to right (until we get to the last entry). Since we want to write our answer in cycle notation, we next compute the value at 3:

$$3 \overset{(1234)}{\longrightarrow} 4 \overset{(46)}{\longrightarrow} 6 \overset{(1235)}{\longrightarrow} 6,$$

and then 6:

$$6 \overset{(1234)}{\longrightarrow} 6 \overset{(46)}{\longrightarrow} 4 \overset{(1235)}{\longrightarrow} 4,$$

and then 4 and so on until we get back to 1. Eventually we find $\sigma\tau$ has the effect:

$$1 \to 3 \to 6 \to 4 \to 2 \to 5 \to 1.$$

Since this accounts for all the elements $\{1, 2, 3, 4, 5, 6\}$, we conclude that $\sigma\tau$ is the 6-cycle (136425), which is indeed consistent with the calculation in Example 4.3.7.

4.4 Basic properties of groups

Recall that a group is a set G with a binary operation $*$ satisfying:

1. $*$ is associative;
2. there's an identity element $e \in G$ for $*$;
3. every element $g \in G$ has an inverse under $*$.

We have now seen a lot of examples of groups, including:

- $(\mathbb{Z}, +)$,
- $(\mathbb{R}^\times, \cdot)$,
- $(M_2(\mathbb{R}), +)$,
- $\mathrm{GL}_2(\mathbb{R})$ under matrix multiplication,
- $(\mathbb{Z}_n, +)$,
- $(\mathbb{Z}_n^\times, \cdot)$,
- D_n (for $n \geq 3$) under composition,
- S_n under composition.

Now we'll establish some basic properties that all groups have. I already mentioned that the identity element is unique, i.e., there is only one element $e \in G$ with the property that $e*a = a*e = a$ for all $a \in G$. (If e' were another element with this property, we'd have $e = e * e' = e'$, a contradiction.) Now we'll show that the inverse of each element is unique.

Proposition 4.4.1 *Suppose $(G, *)$ is a group and $a \in G$. Then there is a unique $b \in G$ such that $a * b = b * a = e$.*

Proof. We already know from the definition of a group that $a*b = b*a = e$ for some $b \in G$; we have to show there is *exactly one* element $b \in G$ with this property. So suppose b and b' are two such elements, i.e.,

$$a * b = b * a = e \quad \text{and} \quad a * b' = b' * a = e.$$

Then the above formulas and associativity imply

$$b = b * e = b * (a * b') = (b * a) * b' = e * b' = b'.$$

□

We've already been referring to the element b as in the proposition as an *inverse* of a. Now we can call it *the inverse* of a.

Definition 4.4.2 Suppose that $(G, *)$ is a group and $a \in G$. Then the **inverse** of a is the unique element $b \in G$ such that $a * b = b * a = e$; we denote this element by a^{-1}.

Here are some more notational conventions: just as for multiplication of real numbers, we often omit the symbol for the binary operation in a group (especially for an "abstract" group) and simply write ab instead of $a * b$.

A binary operation is often denoted + When some sort of "addition" underlies the definition (for example, addition of matrices). In that case, we would denote the inverse of a by $-a$ instead of a^{-1}. The symbol + is also usually reserved for commutative binary operations. Groups for which the binary operation is commutative have a special name, after the mathematician Abel.

Definition 4.4.3 We say a group $(G, *)$ is an **abelian** group if the operation $*$ is commutative; i.e., $a * b = b * a$ for all $a, b \in G$.

Among the above examples, $(\mathbb{Z}, +)$, $(\mathbb{R}^\times, \cdot)$, $(M_2(\mathbb{R}), +)$, $(\mathbb{Z}_n, +)$ and $(\mathbb{Z}_n^\times, \cdot)$ are abelian groups, but $\mathrm{GL}_2(\mathbb{R})$, D_n and S_n (for $n \geq 3$) are non-abelian.

Here's an important property that applies to all groups; it's called the **Cancellation Law**:

Proposition 4.4.4 *Suppose that G is a group and $a, b, c \in G$. If $ab = ac$ or $ba = ca$, then $b = c$.*

Proof. If $ab = ac$, then

$$b = eb = (a^{-1}a)b = a^{-1}(ab) = a^{-1}(ac) = (a^{-1}a)c = ec = c.$$

The proof that $ba = ca \Rightarrow b = c$ is similar. □

Make sure you understand the justification at each step in the string of equalities in the proof, and can supply the "similar" part of the proof yourself. In particular, note that it was important to "multiply" by a^{-1} on the *left* in the above argument since the operation is not necessarily commutative, and that to prove the other implication, you will need to multiply by a^{-1} on the right.

Note that if G is a group and $a, b \in G$, then there's an element $x \in G$ such that $ax = b$, namely $x = a^{-1}b$. Moreover the cancellation law says that x is the *unique* such element. Similarly there's a unique $y \in G$ such that $ya = b$, namely $y = ba^{-1}$. (Note that y may or may not be the same as x.) Let's record this consequence of the cancellation law:

Corollary 4.4.5 *If G is a group and $a, b \in G$, then there is a unique $x \in G$ such that $ax = b$ and a unique $y \in G$ such that $ya = b$.*

You can think of this as a statement about the "multiplication table" for the group. Suppose G has only finitely many elements, say $g_1, g_2, \ldots, g_n$. The entries in the row of a are then

$$ag_1, ag_2, \ldots, ag_n.$$

The corollary says that each element $b \in G$ appears *exactly once* in this list. So b appears exactly once in each row of the table. Similarly, b appears exactly once in each column of the table.

Here are some more general properties of inverses:

Proposition 4.4.6 *Suppose G is a group and $g, h \in G$. Then*

1. *If $ab = e$, then $a = b^{-1}$ and $b = a^{-1}$.*
2. *$(ab)^{-1} = b^{-1}a^{-1}$.*
3. *$(a^{-1})^{-1} = a$.*

Proof. 1) If $ab = e$, then since $b^{-1}b = e$, Prop. 4.4.4 implies that $a = b^{-1}$. Similarly, since $aa^{-1} = e$, Prop. 4.4.4 implies that $b = a^{-1}$.
2) Note that

$$(ab)(b^{-1}a^{-1}) = a(b(b^{-1}a^{-1})) = a((bb^{-1})(a^{-1})) = a(ea^{-1}) = aa^{-1} = e.$$

So applying part 1) (with ab in place of a and $b^{-1}a^{-1}$ in place of b) shows that $b^{-1}a^{-1} = (ab)^{-1}$.
3) Since $aa^{-1} = e$, part 1) with a^{-1} in place of b shows that $a = (a^{-1})^{-1}$. □

Recall that $b = a^{-1}$ means that $ab = e$ *and* $ba = e$. The content of part 1) is that it's enough to know that *either* of these equalities hold; the other follows. Also, note in part 2) that the inverse of ab is given by $b^{-1}a^{-1}$ (the order *reverses*), and this is *not* necessarily the same as $a^{-1}b^{-1}$ unless G is abelian.

Finally one more remark on notation. So far I've been very careful to place parentheses and show how associativity is being applied. For example, the proof of part 2) of the above proposition begins with 3 applications of associativity to shift around the parentheses, replacing an expression of the form $a(bc)$ with $(ab)c$ or vice-versa. By now you should realize that we don't really need the parentheses. Since $a(bc) = (ab)c$, we just write abc instead (or $a * b * c$ if the operation is $*$), and more generally we write $a_1a_2 \cdots a_n$. Note that the order of placement of the elements matters (unless the group is abelian), but the order in which the operation is applied to adjacent elements doesn't matter. With this mind, the formula in the proof of part 2) above becomes:

$$(ab)(b^{-1}a^{-1}) = abb^{-1}a^{-1} = aea^{-1} = aa^{-1} = e.$$

4.5 Powers of group elements

If $(G, *)$ is a group and $g \in G$, then we define the n^{th} *power* of g for positive integers n by

$$g^n = \underbrace{g * g * \cdot * g}_{n \text{ times}}.$$

We extend this to a define g^n for all $n \in \mathbb{Z}$ by setting $g^0 = e$ and $g^n = (g^{-n})^{-1}$ for $n < 0$ (note that we've already defined g^{-n} in this case since $-n > 0$, so we're defining g^n as the inverse of g^{-n}. Notice there's no inconsistency in the definition of g^{-1}. Taking $n = -1$ just gives the inverse of $g^1 = g$, which is what we've already been calling g^{-1}. Here are some examples:

- Suppose $G = D_3$ and $g = \rho_1$. Then $\rho_1^2 = \rho_2$, $\rho_1^3 = e$, $\rho_1^4 = \rho_1$. Of course $\rho_1^0 = e$ and $\rho_1^1 = \rho_1$ by definition. Some negative powers are $\rho_1^{-1} = \rho_2$ and $\rho_1^{-2} = (\rho_1^2)^{-1} = \rho_2^{-1} = \rho_1$.

- For the group of non-zero real numbers under multiplication, x^n has its usual meaning for $x \in \mathbb{R}^\times$, $n \in \mathbb{Z}$.

- Suppose $G = S_6$ and σ is the 6-cycle (123456). Then $\sigma^0 = e$, $\sigma^1 = \sigma$, and $\sigma^2 = \sigma\sigma = (123456)(123456)$ has the effect:

$$\begin{array}{llllll} 1 \xrightarrow{\sigma} 2 \xrightarrow{\sigma} 3, & & 2 \xrightarrow{\sigma} 3 \xrightarrow{\sigma} 4 \\ 3 \xrightarrow{\sigma} 4 \xrightarrow{\sigma} 5, & & 4 \xrightarrow{\sigma} 5 \xrightarrow{\sigma} 6 \\ 5 \xrightarrow{\sigma} 6 \xrightarrow{\sigma} 1, & & 6 \xrightarrow{\sigma} 1 \xrightarrow{\sigma} 2, \end{array}$$

  which in cycle notation is (135)(246). Similarly

$$\sigma^3 = (14)(25)(46), \quad \sigma^4 = (153)(264), \quad \sigma^5 = (165432), \quad \sigma^6 = e,$$

  and then the powers begin repeating, so $\sigma^7 = \sigma$, $\sigma^8 = \sigma^2$, etc. As for negative powers, σ^{-1} is the inverse functions of σ, which could be described as:

$$1 \to 6 \to 5 \to 4 \to 3 \to 2 \to 1,$$

  which is the same as σ^5, and we find $\sigma^{-2} = (\sigma^2)^{-1} = \sigma^4$, etc.

- Consider $\mathbb{Z}$ under addition. Whenever the operation is denoted $+$, then we'll write ng or $n \cdot g$ instead of g^n, and call these *multiples* instead of powers. Then nm has its usual meaning for $m, n \in \mathbb{Z}$.

- Consider the group $(\mathbb{Z}_{10}, +)$. Since the group operation is based on addition, we'll use the "additive" notation. Let's compute some multiples of $[4]$ in $\mathbb{Z}_{10}$: $-1 \cdot [4] = [-4] = [6]$, $0 \cdot [4] = [0]$, $1 \cdot [4] = [4]$, $2 \cdot [4] = [8]$, $3 \cdot [4] = [2]$, etc. It's easy to see that $n \cdot [a]$, the n^{th} multiple of $[a]$ in $\mathbb{Z}_m$, is given by $[na]$. If $n > 0$, this is clear from definitions since

$$n \cdot [a] \;=\; \underbrace{[a] + [a] + \cdots + [a]}_{n \text{ times}} \;=\; [\underbrace{a + a + \cdots + a}_{n \text{ times}}].$$

  It's also clear from the definition if $n = 0$. If $n < 0$, then $n \cdot [a]$ is defined as the (additive) inverse of $(-n) \cdot [a]$, and since $-n > 0$, we know that $(-n) \cdot [a] = [(-n)a]$. Therefore

$$n \cdot [a] = -[(-n)a] = [-(-n)a] = [na].$$

- Recall that $\mathbb{Z}_m^\times = \{\, [a]_m \mid \gcd(a, m) = 1 \,\}$ is a group under *multiplication* of congruence classes. For example,

$$\mathbb{Z}_{100}^\times = \{\, [1], [3], [7], [9], [11], [13], \ldots, [97], [99] \,\},$$

  the included residue classes being those with last digits $1, 3, 7, 9$. Computing a few powers of $[19]$, we have

$$[19]^2 = [19][19] = [361] = [61], \quad [19]^3 = [19]^2[19] = [61][19] = [59].$$

  Just as with addition of congruence classes, it's easy to see that $[a]_m^n = [a^n]_m$ is n is *positive* (where a^n is the "usual" n^{th} power of a). Note though that this formula makes no sense if n is *negative* since a^n is not an integer. To see how to compute $[19]^{-1}$ in $\mathbb{Z}_{100}^\times$ for example, let's recall what this means (see the proof of Prop. 4.2.13). We need to find a congruence class $[x] \in \mathbb{Z}_{100}^\times$ so that $[19x] = [19][x] = [1]$. In other words, we want $19x \equiv 1 \bmod 100$, or $19x - 100y = 1$ for some $y \in \mathbb{Z}$. We can solve this using the Euclidean Algorithm:

$$100 = 5 \cdot 19 + 5, \qquad 19 = 3 \cdot 5 + 4, \qquad 5 = 4 + 1,$$

  which gives

$$1 = 5 - 4 = 5 - (19 - 3 \cdot 5) = 4 \cdot 5 - 19 = 4(100 - 5 \cdot 19) - 19 = 4(100) - 21(19).$$

  We can therefore take $x = -21$ and conclude

$$[19]^{-1} = [-21] = [79].$$

What about $[19]^{-2}$? By definition, this is the inverse of $[19]^2 = [61]$, so we apply the Euclidean Algorithm again. Alternatively, we can use the fact that $[19]^{-2} = ([19]^{-1})^2$ by the Laws of Exponents we'll prove in a moment, and conclude

$$[19]^{-2} = [-21]^2 = [441] = [41].$$

Now let's return to the general situation and show that the powers of an element satisfy the laws of exponents.

Proposition 4.5.1 *Suppose G is a group, $g \in G$ and $m, n \in \mathbb{Z}$. Then $g^m g^n = g^{m+n}$ and $(g^m)^n = g^{mn}$.*

Proof. We'll just prove the first formula and leave the second as an exercise.

If m and n are both positive, then it is clear that $g^m g^n = g^{m+n}$ since

$$\underbrace{gg\cdots g}_{m \text{ times}} \ \underbrace{gg\cdots g}_{n \text{ times}} = \underbrace{gg\cdots g}_{(m+n) \text{ times}}.$$

(More formally, we would have defined $g^0 = e$ and g^n for $n \geq 1$ inductively by $g^n = g^{n-1}g$, and then proved $g^m g^n = g^{m+n}$ in this case by induction on n.)

The formula is also clear if $m = 0$ or $n = 0$.

So now suppose m and n are both negative. In that case $g^m = (g^{-m})^{-1}$, $g^n = (g^{-n})^{-1}$ and $g^{m+n} = (g^{-(m+n)})^{-1}$ by definition. But since $-m$, $-n$ are positive, we've already proved that $g^{-n}g^{-m} = g^{-n-m} = g^{-(m+n)}$. Applying Prop. 4.4.6, part 2) then gives

$$g^m g^n = (g^{-m})^{-1}(g^{-n})^{-1} = (g^{-n}g^{-m})^{-1} = (g^{-(m+n)})^{-1} = g^{m+n}.$$

Now suppose $m < 0$ and $n > 0$. If $m + n \geq 0$, then since $-m < 0$, we already know that $g^{-m}g^{m+n} = g^{-m+m+n} = g^n$. But since $g^m = (g^{-m})^{-1}$, multiplying by g^m on the left gives

$$g^m g^n = g^m g^{-m} g^{m+n} = e g^{m+n} = g^{m+n}.$$

If on the other hand, $m + n < 0$, then we already know that $g^{m+n}g^{-n} = g^m$ (all the exponents being negative), and multiplying on the right by g^n gives $g^m g^n = g^{m+n}$.

Finally if $m > 0$ and $n < 0$, the preceding cases show $g^{-m}g^{m+n} = g^n$ (regardless of the sign of $m + n$), and it follows as above that $g^m g^n = g^{m+n}$.

□

Note that when using additive notation, the laws of "exponents" become:

$$(m \cdot a) + (n \cdot a) = (m + n) \cdot a, \qquad m \cdot (n \cdot a) = (mn) \cdot a$$

for all $m, n \in \mathbb{Z}$ and $a \in G$ (a group with operation $+$).

Finally a word a caution. One of the usual laws of exponents for real numbers is that $x^n y^n = (xy)^n$ for $x, y \neq 0$, $n \in \mathbb{Z}$. It's easy to see that $g^n h^n = (gh)^n$ for g, h in an *abelian* group G, but this rule will *not* apply in a *non-abelian* group. Taking $n = 2$ for example, $g^2h^2 = gghh$, but $(gh)^2 = ghgh$, and these might not be the same.

4.6 Orders of group elements

Recall that if G is a group, then for $g \in G$, $n \in \mathbb{Z}$, we defined g^n (the n^{th} power of g) by setting

- $g^n = gg \cdots g$ (n times) for $n > 0$;
- $g^0 = e$;
- $g^n = (g^{-n})^{-1}$ for $n < 0$.

For example, the powers of $g = \rho_1 \in D_3$ are given by

n	-4	-3	-2	-1	0	1	2	3	4	5	6	7	$\cdots$
g^n	ρ_2	e	ρ_1	ρ_2	e	ρ_1	ρ_2	e	ρ_1	ρ_2	e	ρ_1	$\cdots$

Notice the pattern, which can be expressed as:

$$\rho_1^n = \begin{cases} e & \text{if } n \equiv 0 \pmod 3, \\ \rho_1 & \text{if } n \equiv 1 \pmod 3, \\ \rho_2 & \text{if } n \equiv 2 \pmod 3. \end{cases}$$

Similarly, for $G = \mathbb{Z}_{10}$ under $+$ taking powers of $[4]$ (or "multiples" since the operation is additive) gives

n	-4	-3	-2	-1	0	1	2	3	4	5	6	7	$\cdots$
$n \cdot [4]$	$[4]$	$[8]$	$[2]$	$[6]$	$[0]$	$[4]$	$[8]$	$[2]$	$[6]$	$[0]$	$[4]$	$[8]$	$\cdots$

On the other hand, if we take powers of 2 in $\mathbb{R}^\times$, there is no repetition.

These phenomena can be explained in general in terms of the *order* of the element, defined as follows:

Definition 4.6.1 Let g be an element of a group G. We say that g has **finite order** (in G) if $g^n = e$ for some $n \in \mathbb{N}$. In that case the least $n \in \mathbb{N}$ such that $g^n = e$ is called the **order** of g. If no such positive integer n exists, we say that g has **infinite order**.

For example, the element $\rho_1 \in D_3$ has order 3, and the element $[4] \in \mathbb{Z}_{10}$ has order 5. The element $2 \in \mathbb{R}^\times$ has infinite order; in fact the only elements of $\mathbb{R}^\times$ with finite order are ± 1. Note that the identity element e in any group has order 1.

Theorem 4.6.2 *Suppose that g is an element of a group G.*

1. *If g has infinite order, then $g^n = e \Leftrightarrow n = 0$.*
2. *If g has finite order d, then $g^n = e \Leftrightarrow d|n$.*

Proof. 1) Suppose that g has infinite order. Then by the definition of *infinite order*, there is no positive integer n such that $g^n = e$. It is also true by definition that $g^0 = e$, so we just have to prove that there is no *negative* integer n such that $g^n = e$. Suppose then that $n < 0$ and $g^n = e$. Then by definition of g^n in this case, we have $(g^{-n})^{-1} = e$. But then it follows that $g^{-n} = e$, and since $-n > 0$, this would contradict g having infinite order.

2) Suppose now that g has order d, i.e., d is the least positive integer such that $g^d = e$.

If $d|n$, then $n = dm$ for some $m \in \mathbb{Z}$. Since $g^d = e$, we find that

$$g^n = g^{dm} = (g^d)^m = e^m = e$$

(where the second equality is by the laws of exponents, Prop. 4.5.1).

Suppose conversely that $g^n = e$. By the Division Algorithm Thm. 2.1.2, we can write $n = dq + r$ for some $q, r \in \mathbb{Z}$ with $0 \leq r < d$. Applying Prop. 4.5.1 again gives

$$e = g^n = g^{dq+r} = g^{dq}g^r = (g^d)^q g^r = e^q g^r = g^r.$$

But since $r < d$ and d is the *least* positive integer such that $g^d = e$, it follows that r cannot be a positive integer. The only possibility then is that $r = 0$, so $n = dq$ is divisible by d. □

Corollary 4.6.3 *Suppose that g is an element of a group G.*

1. *If g has infinite order, then the powers of g are distinct; i.e., $g^m = g^n \Leftrightarrow m = n$.*
2. *If g has finite order d, then $g^m = g^n \Leftrightarrow m \equiv n \pmod{d}$.*

Proof. 1) Suppose that g has infinite order. Clearly if $m = n$, then $g^m = g^n$. Suppose conversely that $g^m = g^n$. Then $g^{m-n} = g^m g^{-n} = g^n g^{-n} = e$ (where the first equality is by Prop. 4.5.1), so part 1) of Thm. 4.6.2 implies that $m - n = 0$, so $m = n$.

2) Suppose now that g has order d. Then

$$\begin{array}{rcll} g^m = g^n & \Rightarrow & g^{m-n} = e & \text{(as in part 1)} \\ & \Rightarrow & d|(m-n) & \text{(by Thm. 4.6.2)} \\ & \Rightarrow & m \equiv n \pmod{d} & \text{by definition of congruence.} \end{array}$$

Note also that we can "reverse" the argument; i.e., each $\Rightarrow$ can be replaced by $\Leftrightarrow$, so we see in fact that $g^m = g^n \Leftrightarrow m \equiv n \pmod{d}$. □

Note that part 2) of the corollary describes what we saw in the examples of $\rho_1 \in D_3$ and $[4] \in \mathbb{Z}_{10}$; part 1) describes what we saw for $2 \in \mathbb{R}^\times$.

We've been considering the *order* of an *element* of a group (Defn. 4.6.1). There is also the notion of the *order* of a *group*, which is just its size.

Definition 4.6.4 Suppose that G is a group. If G has infinitely many elements, we say G has **infinite order**. Otherwise we say G has **finite order**, and we define the **order** of G to be the number of elements in G.

So for example, the *group* D_3 has order 6; its *element* ρ_1 has order 3. We'll see later how the two notions are related. For now let's just note the following:

Corollary 4.6.5 *If a group G has finite order, then so does every element of G.*

Proof. If $g \in G$ has infinite order, then Cor. 4.6.3 shows that its powers g^n would give infinitely many distinct elements of G. □

Let's consider the order of some permutations. The computation of the powers of $\sigma = (123456)$ in S_6 shows that σ has order 6. We saw also that its inverse is $(165432) = (654321)$. In fact, in general:

Proposition 4.6.6 *Suppose that $a_1, a_2, \ldots, a_k$ are distinct elements of the set $\{1, 2, \ldots, n\}$, and let $\sigma = (a_1\ a_2\ \cdots\ a_k)$. Then*

1. *$\sigma^{-1} = (a_k\ a_{k-1}\ \cdots\ a_1)$;*
2. *σ has order k.*

Proof. 1) Let $\tau = (a_k\ a_{k-1}\ \cdots\ a_1)$. We have to show that $\tau = \sigma^{-1}$. Since S_n is a group, we only have to show that $\tau\sigma = e$, which means that $\tau(\sigma(a)) = a$ for each $a \in \{1, 2, \ldots, n\}$.

- We find that $\tau(\sigma(a_1)) = \tau(a_2) = a_1$, $\tau(\sigma(a_2)) = \tau(a_3) = a_2$, $\ldots$, $\tau(\sigma(a_{k-1})) = \tau(a_k) = a_{k-1}$, and $\tau(\sigma(a_k)) = \tau(a_1) = a_k$.

- If $b \neq a_i$ for any $i = 1, 2, \ldots, k$, then $\tau(\sigma(b)) = b$.

So $\tau\sigma = e$, and therefore $\tau = \sigma^{-1}$.

2) We have to show that $\sigma^k = e$, and that $\sigma^i \neq e$ for $i = 1, 2, \ldots k-1$. We first compute $\sigma^i(a_1)$ for each i. Recall that $\sigma^i = \sigma\sigma\cdots\sigma$, repeated i times. So by the definition of σ, we have $\sigma(a_1) = a_2$, $\sigma^2(a_1) = \sigma(\sigma(a_1)) = \sigma(a_2) = a_3$, and by induction on i, we find that

$$\sigma^i(a_1) = \sigma(a_{i+1}) \qquad \text{if } 1 \leq i \leq k-1.$$

(The induction step is $\sigma^i(a_1) = \sigma(\sigma^{i-1}(a_1)) = \sigma(a_i) = a_{i+1}$.) Since $\sigma^i(a_1) = a_{i+1} \neq a_1$, this already shows that σ^i is not the identity for $1 \leq i \leq k-1$.

We must still show that σ^k is the identity. To see this, first note that

$$\sigma^k(a_1) = \sigma(\sigma^{k-1}(a_1)) = \sigma(a_k) = a_1.$$

Furthermore for $i = 1, 2, \ldots, k-1$, we have

$$\sigma^k(a_{i+1}) = \sigma^k(\sigma^i(a_1)) = \sigma^{i+k}(a_1) = \sigma^i(\sigma^k(a_1)) = \sigma^i(a_1) = a_{i+1}.$$

We have now shown that $\sigma^k(a_1) = a_1$, $\sigma(a_2) = a_2$, $\ldots$, $\sigma^k(a_k) = a_k$. Finally if $b \neq a_i$ for any $i = 1, 2, \ldots, k$, then $\sigma(b) = b$, so by induction on i, we get $\sigma^i(b) = b$ for all $i \geq 1$, and in particular, $\sigma^k(b) = b$. We have now shown that $\sigma^k = e$. □

4.7 Subgroups

Definition 4.7.1 Suppose that $(G, *)$ is a group. A subset $H \subseteq G$ is called a **subgroup** of G if H, with the operation $*$, is a group.

Here are some examples:

- The subset $H = \{\text{ even numbers }\}$ of $\mathbb{Z}$ (under addition) is a subgroup. Note first that $+$ defines a binary operation on H since the sum of two even numbers is even; the operation is associative; there is an identity element $0 \in H$; and if $n \in H$, then it has an inverse $-n \in H$.

- The subset $\mathbb{N}$ of $\mathbb{Z}$ is *not* a subgroup. Though $+$ defines a binary operation on $\mathbb{N}$, there is no identity element (or inverses).

- The subset $H = \{ e, \rho_1, \rho_2 \}$ of D_3 is a subgroup.

We will now establish a list of criteria which determine whether a given subset H of a group G is in fact a subgroup.

Note first of all that for H to be a subgroup of G, we need $*$ to define a binary operation on H. This means that if h and h' are elements of H, then the output $h * h'$ of the binary operation on G, is in fact in the subset H. We therefore require H to be *closed* under the operation $*$; i.e.,

1) $h, h' \in H \Rightarrow h * h' \in H$.

If 1) is satisfied, then we have a set H with a binary operation $*$. It is a group if it satisfies the three properties in the definition of a group (Definition 4.1.1). Note first that since G is a group, the operation $*$ is associative on G, so it must be associative on the subset H as well.

The second property is that there be an identity element for $*$ in H. We know already that G has an identity element e for $*$. Denote by e' the identity element for $*$ in H, so $e' \in H$ satisfies

$$e' * h = h * e' = h \quad \text{for all } h \in H.$$

In particular $e' * e' = e'$. But since e is the identity element for $*$ on G, and $e' \in G$, we also have $e * e' = e'$. Therefore $e * e' = e' * e'$, and the Cancellation Law (Prop. 4.4.4) implies that $e = e'$. So if there is an identity element for $*$ in H, that element must be e. So the second property we need for H to be a subgroup amounts to the requirement:

2) $e \in H$.

The last condition is that every $h \in H$ have an inverse in H with respect to $*$. This means that if $h \in H$, then there is an $h' \in H$ such that $h * h' = h' * h = e$. But we know already that h has an inverse in G, which we've denoted h^{-1}. Since $h * h' = h * h^{-1} = e$, Prop. 4.4.6 implies that $h' = h^{-1}$. So the third property amounts to the condition:

3) $h \in H \Rightarrow h^{-1} \in H$.

We have now proved the following:

Proposition 4.7.2 *Suppose that $(G, *)$ is a group and $H \subseteq G$. Then H is a subgroup of G if and only if the following conditions are all satisfied:*

1. *$h, h' \in H \Rightarrow h * h' \in H$;*

2. *$e \in H$;*

3. *$h \in H \Rightarrow h^{-1} \in H$.*

Example 4.7.3 Let $G = \mathbb{Z}$ under addition and suppose $m \in \mathbb{Z}$. Consider the subset

$$H = \{ \text{integer multiples of } m \} = \{ n \in \mathbb{Z} \mid n \text{ is divisible by } m \}.$$

We will check that H is a subgroup of $\mathbb{Z}$ by verifying that it satisfies the three conditions of Prop. 4.7.2.

1. Suppose $n, n' \in H$. Then $n = km$, $n' = k'm$ for some $k, k' \in \mathbb{Z}$, so $n + n' = km + k'm = (k + k')m$ is also divisible by m, so $n + n' \in H$. (I.e., the sum of two multiples of m is again a multiple of m; we knew this already as part of Prop. 2.1.1.)

2. Since $0 = 0 \cdot m$ is a multiple of m, we have $0 \in H$.

3. Suppose $n \in H$. Then $n = km$ for some $k \in \mathbb{Z}$, so its inverse $-n = -km$ is a multiple of k, so $-n \in H$.

Note that the set of even integers is the special case where $m = 2$. Taking $m = 1$ instead gives all of $\mathbb{Z}$; taking $m = 0$ would give the subgroup $\{0\}$. The set of integer multiples of m is usually denoted $m\mathbb{Z}$. We shall see that every subgroup of $\mathbb{Z}$ is of this form.

Remark 4.7.4 In the statement of the proposition, we really need all three conditions to guarantee that H is a subgroup. You might wonder if two of the conditions imply the third and are therefore sufficient to imply that H is a subgroup, but this isn't the case. For example, taking $G = \mathbb{Z}$:

1. The subset $\{0, 1, 2, 3, \ldots\}$ satisfies 1) and 2) but not 3), so it is not a subgroup.

2. The subset $\{-1, 0, 1\}$ satisfies 2) and 3) but not 1) (since 1 is in the subset, but $1 + 1$ is not), so it is not a subgroup.

3. The empty set $\emptyset$ satisfies 1) and 3) but not 2), so it is not a subgroup. (Note that 3) for example means that $-n$ must be in the subset whenever n is, but n is *never* in this subset, so the condition is automatically satisfied.)

Example 4.7.5 Let $G = \mathrm{GL}_2(\mathbb{R})$ and let

$$\mathrm{SL}_2(\mathbb{R}) = \{ A \in \mathrm{GL}_2(\mathbb{R}) \mid \det A = 1 \}.$$

(The GL is for *general linear*, and SL for *special linear*.)

1. Suppose $A, B \in \mathrm{SL}_2(\mathbb{R})$. Then $\det A = \det B = 1$, so $\det(AB) = (\det A)(\det B) = 1$ (an exercise) and $AB \in \mathrm{SL}_2(\mathbb{R})$.

2. We have $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R})$ since $\det I = 1$.

3. Suppose $A \in \mathrm{SL}_2(\mathbb{R})$. Then $\det A = 1$, so $\det(A^{-1}) = (\det A)^{-1} = 1$, so $A^{-1} \in \mathrm{SL}_2(\mathbb{R})$.

Since $\mathrm{SL}_2(\mathbb{R})$ satisfies 1), 2) and 3) of Prop. 4.7.2, it is a subgroup of $\mathrm{GL}_2(\mathbb{R})$.

4.8 Cyclic groups

We now turn our attention to a special type of subgroup:

Proposition 4.8.1 *Suppose that G is a group and $g \in G$. Then*

$$H = \{\, g^n \mid n \in \mathbb{Z} \,\}$$

is a subgroup of G.

Proof. As usual, we verify 1), 2) and 3) of Prop. 4.7.2.

1. Suppose $h, h' \in H$. Then $h = g^n, h' = g^{n'}$ for some $n, n' \in \mathbb{Z}$, so $hh' = g^n g^{n'} = g^{n+n'}$ by Prop. 4.5.1, so $hh' \in H$.

2. By definition $e = g^0$, so $e \in H$.

3. Suppose $h \in H$. Then $h = g^n$ for some $n \in \mathbb{Z}$, so $h^{-1} = (g^n)^{-1} = g^{-n}$ by Prop. 4.5.1, so $h^{-1} \in H$.

So H is a subgroup. □

Definition 4.8.2 If g is an element of a group G, then $\{\, g^n \mid n \in \mathbb{Z} \,\}$ is denoted $\langle g \rangle$ and called the **subgroup of G generated by g.**

Here are some examples:

- The subgroup of D_3 generated by ρ_1 is $\langle \rho_1 \rangle = \{e, \rho_1, \rho_2\}$ (since ρ_1 has order 3, Cor. 4.6.3 shows that $\rho_1^0, \rho_1^1, \rho_1^2$ already gives all the powers of ρ_1).

- The subgroup of $\mathbb{Z}$ (under $+$) generated by m is the set of multiples of m (Example 4.7.3), so for example $\langle 4 \rangle$ is the set of integer multiples of 4.

- The subgroup of $\mathbb{R}^\times$ generated by 4 is $\langle 4 \rangle = \{\, 4^n \mid n \in \mathbb{Z} \,\}$. (Note in particular that the meaning of "$\langle g \rangle$" depends on the group G in which we're working.)

- The subgroup of $\mathbb{Z}_{10}$ generated by $[4]$ consists of all *multiples* of $[4]$ (since the operation is $+$). Therefore $\langle [4] \rangle = \{[0], [4], [8], [2], [6]\}$ (this is all since $[4]$ has order 5 in $\mathbb{Z}_{10}$).

- The subgroup of $\mathbb{Z}_5$ generated by $[4]$ is $\{[0], [4], [3], [2], [1]\}$, which is all of $\mathbb{Z}_5$.

- The subgroup of $\mathbb{Z}_5^\times$ generated by $[4]$ is $\{[1], [4]\}$ since $[4]^2 = [1]$ in $\mathbb{Z}_5^\times$.

- The subgroup of S_6 generated by (123456) is

$$\{e, (123456), (135)(246), (14)(25)(36), (153)(264), (165432)\}.$$

- In any group G, we have $\langle e \rangle = \{e\}$ (where e is the identity element).

Recall we defined the *order* of an *element* of a group in Defn. 4.6.1. There is also the notion of the *order* of a *group* (Defn. 4.6.4) which is just its size. So for example, D_3 has order 6; the element ρ_1 of D_3 has order 3. Here is one way in which the two notions are related:

Proposition 4.8.3 *Suppose that g is an element of a group G.*

1. *If g has infinite order, then so does $\langle g \rangle$.*

2. *If g has order $d \in \mathbb{N}$, then so does $\langle g \rangle$.*

Proof. 1) Recall from Corollary 4.6.3 that if g has infinite order, then the group elements g^n for $n \in \mathbb{Z}$ are distinct. Since these are in $\langle g \rangle$, we see that $\langle g \rangle$ has infinitely many elements.
2) The second part of Corollary 4.6.3 says that if g has order d, then $g^m = g^n$ precisely when $m \equiv n \pmod{d}$, or equivalently, when m and n have the same remainder on division by d. Therefore the elements of $\langle g \rangle$ are precisely the elements g^r as r runs through the possible remainders $\{0, 1, \ldots, d-1\}$. So $\langle g \rangle = \{e, g, g^2, \ldots, g^{d-1}\}$ has precisely d elements. □

Going back to some of the examples:

- ρ_1 (in D_3) has order 3, and so does $\langle \rho_1 \rangle = \{e, \rho_1, \rho_2\}$.

- If $m \neq 0$, then $m \in \mathbb{Z}$ has infinite order, and so does $\langle m \rangle$ (the set of mutliples of m).

- $[4]$ in $\mathbb{Z}_{10}$ has order 5, and so does $\langle [4] \rangle = \{[0], [4], [8], [2], [6]\}$.
- $[4]$ in $\mathbb{Z}_5$ has order 5, and so does $\langle [4] \rangle = \{[0], [4], [3], [2], [1]\}$.

In the last example, the subgroup generated by g was the whole group G.

Definition 4.8.4 Suppose that G is a group. We say that G is a **cyclic** group if $G = \langle g \rangle$ for some $g \in G$. If $G = \langle g \rangle$, then we say that g is a **generator** (of G).

Here are some examples:

- $\mathbb{Z}$ is cyclic since the subgroup generated by 1 is the set of multiples of 1, which is all of $\mathbb{Z}$.
- $\mathbb{Z}_n$ is cyclic since the subgroup generated by $[1]$ is all of $\mathbb{Z}_n$. To see this note that if $[k] \in \mathbb{Z}_n$, then $[k] = k \cdot [1]$ is a multiple of $[1]$.
- $\mathbb{Z}_n^\times$ might or might not be cyclic, depending on n. It's easy to see that the group is cyclic for the first few values $n = 1, 2, \ldots, 7$; for example $\mathbb{Z}_7^\times = \langle [3] \rangle$ since the powers of $[3]$ in $\mathbb{Z}_7^\times$ are:

n	0	1	2	3	4	5
$[3]^n$	$[1]$	$[3]$	$[2]$	$[6]$	$[4]$	$[5]$.

  On the other hand $\mathbb{Z}_8^\times = \{[1], [3], [5], [7]\}$ is *not* cyclic since

$$\langle [1] \rangle = \{[1]\}, \quad \langle [3] \rangle = \{[1], [3]\}, \quad \langle [5] \rangle = \{[1], [5]\}$$

$$\text{and } \langle [7] \rangle = \{[1], [7]\},$$

  none of which is all of $\mathbb{Z}_8^\times$.
- For any $g \in G$, the subgroup $H = \langle g \rangle$ for any $g \in G$ is a cyclic group because the subgroup of H generated by g is the same as the subgroup of G generated by g (as can be seen directly from the definition), and this is all of $H = \langle g \rangle$.

Note that G is cyclic if *some* element of G is a generator. There may be several elements which are generators; for example, $\mathbb{Z}_5$ is generated by $[1]$, but we saw it was also generated by $[4]$. We'll see later how to tell exactly which elements of $\mathbb{Z}_n$ are generators. First here's a general criterion for an element of a finite cyclic group to be a generator.

Proposition 4.8.5 *Suppose that G is a finite group of order n.*

1. *If $g \in G$, then g has order at most n.*
2. *G is cyclic if and only if G has an element of order n.*
3. *If G is cyclic and $g \in G$, then g is a generator of G if and only if g has order n.*

Proof. Suppose that $g \in G$. Recall from Prop. 4.8.3 that the order of g is the number of elements of $\langle g \rangle$. Since $\langle g \rangle \subseteq G$, we see that the order of g is at most n. Moreover $\langle g \rangle = G$ if and only if g has order n. □

For example:

- D_3 has order 6, and the orders of its elements are 1 (the identity), 2 (the reflections) and 3 (the rotations). Therefore D_3 is not cyclic.
- $\mathbb{Z}_{10}$ has order 10, and the orders of its elements are:

element	[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
order	1	10	5	10	5	2	5	10	5	10.

  So the elements of $\mathbb{Z}_{10}$ which are generators are [1], [3], [7] and [9].

Recall that an *abelian* group is one in which the binary operation is commutative.

Proposition 4.8.6 *If G is a cyclic group, then G is abelian.*

Proof. If G is cyclic, then $G = \langle g \rangle$ for some $g \in G$. We must show that $hk = kh$ for all $h, k \in G$. Since $G = \langle g \rangle$, we know that $h = g^m$ and $k = g^n$ for some $m, n \in \mathbb{Z}$. Therefore

$$hk = g^m g^n = g^{m+n} = g^{n+m} = g^n g^m = kh,$$

by one of the laws of exponents (Prop. 4.5.1). □

This gives another way of seeing that D_3 is not cyclic (since it's not abelian). The proposition states that every cyclic group is abelian. On the other hand, there are abelian groups that aren't cyclic, for example $\mathbb{Z}_8^\times$. For an example of an infinite abelian which isn't cyclic, consider $\mathbb{R}^\times$, the group of non-zero real numbers under multiplication. To see that $\mathbb{R}^\times$ is not cyclic, we'll suppose that it is and arrive at a contradiction. Suppose that $\mathbb{R}^\times = \langle x \rangle$ for some $x \in \mathbb{R}^\times$. Since $-x \in \mathbb{R}^\times$, we must have $-x = x^n$ for some $n \in \mathbb{Z}$. It follows that $-1 = x^{n-1}$, which implies that $x = -1$ (and n is even). But then $\mathbb{R}^\times = \langle x \rangle = \langle -1 \rangle = \{1, -1\}$ is clearly a contradiction, so $\mathbb{R}^\times$ is not cyclic.

Recall that if g is an element of a group G, then

$$\langle g \rangle = \{ g^n \mid n \in \mathbb{Z} \}$$

is the *subgroup* of G *generated* by g. For example, the subgroup of D_3 generated by ρ_1 is $\langle \rho_1 \rangle = \{e, \rho_1, \rho_2\}$, and the subgroup of $\mathbb{Z}$ generate by m is the set of multiples of m. We say G is *cyclic* if $G = \langle g \rangle$ for some $g \in G$, and then we call g a *generator* of G. For example, $\mathbb{Z}_n$ is cyclic since it is generated by the element $[1]_n$. Note though that there may be other generators; for example $[4]_5$ is also a generator of $\mathbb{Z}_5$. We'll see in a moment how to tell exactly which elements of $\mathbb{Z}_n$ are generators. Note that the generators are the elements of order n. In fact, if we know the order of g, then a simple formula gives the order of any power of g (or multiple of g if the operation is $+$):

Theorem 4.8.7 *Suppose that G is a group, $g \in G$ has order d, and $a \in \mathbb{Z}$. Then g^a has order $d/\gcd(a,d)$.*

Proof. Recall that the order of g^a is the least positive integer n such that $g^{an} = (g^a)^n = e$. Since we are assuming g has order d, Thm. 4.6.2 shows that $g^{an} = e$ if and only if $d|an$. Now let $b = \gcd(a,d)$. Since b divides both a and d, we see that d/b, a/b and na/b are all integers. Moreover

$$d|na \quad \Leftrightarrow \quad (d/b)|(a/b)n.$$

(To see this, note that $na = kd$ for some $k \in \mathbb{Z}$ if and only if $na/b = k(d/b)$ for some $k \in \mathbb{Z}$.) By Cor. 2.3.3, a/b and d/b are relatively prime, so by Cor. 2.3.4, we see that if $(d/b)|n(a/b)$, then $(d/b)|n$. Also, if n is divisible by d/b, then so of course is $n(a/b)$, so

$$(d/b)|n(a/b) \quad \Leftrightarrow \quad (d/b)|n.$$

Therefore

$$(g^a)^n = e \quad \Leftrightarrow \quad (d/b)|n.$$

So the smallest positive integer n for which $(g^a)^n = e$ is the smallest positive integer divisible by d/b, which of course is d/b itself. Therefore the order of g^a is d/b. □

This gives a quick way to compute the order of any element of a cyclic group, once we have a generator. Consider for example the group $G = \mathbb{Z}_7^\times$. We saw this was cyclic, generated by $[3]_7$, so $[3]_7$ has order 6 and each element

can be written as a power of 3. Here then is a table with the order of each element of $\mathbb{Z}_7^\times$:

a	0	1	2	3	4	5
$[3]^a$	[1]	[3]	[2]	[6]	[4]	[5]
$\gcd(a,6)$	6	1	2	3	2	1
order of g^a	1	6	3	2	3	6.

As another example, consider the element $[4]_{10}$ in $\mathbb{Z}_{10}$. Since $[1]_{10}$ has order 10, the order of $[4]_{10} = 4 \cdot [1]_{10}$ is $10/\gcd(4,10) = 10/2 = 5$. In fact, we can now easily compute the order of any element of $\mathbb{Z}_n$:

Corollary 4.8.8 *Suppose that $a \in \mathbb{Z}$ and $N \in \mathbb{N}$. Then the element $[a]_n$ in $\mathbb{Z}_n$ has order $n/\gcd(a,n)$.*

Proof. The element $[1]_n$ has order n, so Theorem 4.8.7 shows that $[a]_n = a \cdot [1]_n$ has order $n/\gcd(a,n)$. Recall from Prop. 4.8.5 that $[a]_n$ generates $\mathbb{Z}_n$ if and only if $[a]_n$ has order n, which by Cor. 4.8.8 is equivalent to $\gcd(a,n) = 1$. So we have:

Corollary 4.8.9 *Suppose that $a \in \mathbb{Z}$ and $N \in \mathbb{N}$. Then $[a]_n$ generates $\mathbb{Z}_n$ if and only if a and n are relatively prime.*

Cyclic groups have the following convenient property:

Theorem 4.8.10 *Every subgroup of a cyclic group is cyclic.*

Proof. Suppose that G is a cyclic group and H is a subgroup of G. We must prove that H is cyclic.

Suppose first that $H = \{e\}$. Then $H = \langle e \rangle$, so H is cyclic.

Suppose now that $H \neq \{e\}$, so there is some element $h \in H$ such that $h \neq e$. Since G is cyclic, $G = \langle g \rangle = \{\, g^n \mid n \in \mathbb{Z} \,\}$ for some $g \in G$. Since h is in G, we must have $h = g^m$ for some $m \in \mathbb{Z}$. Since $h \neq e$, it follows that $m \neq 0$. If $m < 0$, then $h^{-1} = g^{-m}$ is also in H, and $-m > 0$. This shows that $g^n \in H$ for some positive integer n (taking either $n = m$ or $n = -m$).

Now let b be the *least* positive integer such that $g^b \in H$. Then

$$\langle g^b \rangle = \{\, (g^b)^k \mid k \in \mathbb{Z} \}$$

is the subgroup of H generated by g^b. We shall prove that in fact $H = \langle g^b \rangle$, and therefore H is cyclic.

Since $\langle g^b \rangle \subseteq H$, we just need to show that $H \subseteq \langle g^b \rangle$. So suppose $h' \in H$. We must show that h' is in $\langle g^b \rangle$. Since $h' \in G = \langle g \rangle$, we know that $h' = g^a$

for some $a \in \mathbb{Z}$. By the division algorithm, $a = bq + r$ for some $q, r \in \mathbb{Z}$ with $0 \leq r < b$. Substituting for a and applying Prop. 4.5.1 gives

$$g^a = g^{bq+r} = g^{bq}g^r = (g^b)^q g^r.$$

Therefore $g^r = (g^b)^{-q}g^a$. By assumption $g^a = h' \in H$. But also $(g^b)^{-q} \in \langle g^b \rangle \subseteq H$, and since H is a subgroup, it follows that $g^r \in H$. Since $r < b$ and b was assumed to be the least positive integer such that $g^r \in H$, it follows that r cannot be positive. This means that $r = 0$, so $a = bq$ and $h' = g^a = (g^b)^q$ is in $\langle g^b \rangle$. □

Corollary 4.8.11 *If H is a subgroup of $\mathbb{Z}$, then*

$$H = \langle m \rangle = \{\, km \mid k \in \mathbb{Z} \,\}$$

for some $m \in \mathbb{Z}$.

Proof. Since $\mathbb{Z}$ is cyclic, we can apply Thm. 4.8.10 to deduce that H is cyclic. Therefore it is generated by some $m \in H \subseteq \mathbb{Z}$. □

Example 4.8.12 Suppose that $a, b \in \mathbb{Z}$, and let

$$H = \{\, ax + by \mid x, y \in \mathbb{Z} \,\}$$

(the set of integer linear combinations of a and b). Then H is a subset of $\mathbb{Z}$, but we will check that H is in fact a subgroup of $\mathbb{Z}$. As usual, we check the three conditions in Prop. 4.7.2:

1. Suppose that $h, h' \in H$, so $h = ax + by$, $h' = ax' + by'$ for some $x, y, x', y' \in \mathbb{Z}$. Then

   $$h + h' = (ax + by) + (ax' + by') = a(x + x') + b(y + y')$$

   which has the form required to be in H (since $x + x'$ and $y + y'$ are integers).

2. $0 \in H$ since $0 = a \cdot 0 + b \cdot 0$.

3. If $h \in H$, then $h = ax + by$ for some $x, y \in \mathbb{Z}$, and its inverse in the group is $-h = -(ax + by) = a(-x) + b(-y)$, which is again in H sicne $-x, -y \in \mathbb{Z}$.

Since H is a subgroup of $\mathbb{Z}$, we must have that H is the set of multiples of some integer m. Furthermore if a or b is non-zero, then H contains non-zero elements, so we can take $m > 0$. But you already know this; the integer m is $\gcd(a, b)$. (Recall from Thm. 2.4.1 that an integer c is of the form $ax + by$ for some $x, y \in \mathbb{Z}$ if and only if c is a multiple of $\gcd(a, b)$.)

4.9 Cosets

Now we'll introduce the notion of the *cosets* of a subgroup. We'll use this to prove an important theorem about finite groups, called Lagrange's Theorem.

Definition 4.9.1 Suppose that $(G, *)$ is a group, H is a subgroup of G, and g is an element of G. The subset $g * H \subseteq G$ defined by

$$g * H = \{\, g * h \,|\, h \in H \,\}$$

is called a **left coset** of H in G.

If omitting the symbol for the binary operation, we would write gH instead of $g * H$.

Example 4.9.2 Let $G = D_3$, $H = \{e, \phi_1\}$ and $g = \rho_1$ (in the notation of Example 4.2.8). Then H is a subgroup of G, and

$$\rho_1 H = \{\, \rho_1 h \,|\, h \in H \,\} = \{\rho_1 e, \rho_1 \phi_1\} = \{\rho_1, \phi_3\}$$

is a left coset of H in G. For each $g \in G$, we have a left coset gH of H in G, but note that different values of G can give the same coset gH; for example,

$$eH = \{ee, e\phi_1\} = \{e, \phi_1\} \quad \text{and} \quad \phi_1 H = \{\phi_1 e, \phi_1 \phi_1\} = \{\phi_1, e\}$$

are the *same* subsets of G, so $eH = \phi_1 H$ is a *single* coset. (Note that for any G and H, we have $H = eH$, so the subgroup H is itself a left coset of H in G.) Let's compute *all* the left cosets of H in D_3. We've already computed gH for $g = e$, ϕ_1 and ρ_1; the remaining values of g give

$$\rho_2 H = \{\rho_2, \phi_2\}, \quad \phi_2 H = \{\phi_2, \rho_2\} \quad \text{and} \quad \phi_3 H = \{\phi_3, \rho_1\}.$$

So all together there are 3 left cosets of H in D_3, namely

$$\begin{array}{rrcccl} & eH & = & \phi_1 H & = & \{e, \phi_1\}, \\ & \rho_1 H & = & \phi_3 H & = & \{\rho_1, \phi_3\} \\ \text{and} & \rho_2 H & = & \phi_2 H & = & \{\rho_2, \phi_2\}. \end{array}$$

Note that each left coset of H in D_3 contains the same number of elements. Furthermore each element of D_3 appears in exactly one of the left cosets of H. We'll see this is what happens in general, but first let's consider another example.

Example 4.9.3 Let $G = \mathbb{Z}$, $n \in \mathbb{N}$ and $H = \langle n \rangle$. Recall that H is the set of integer multiples of n. For $a \in \mathbb{Z}$, we can form a left coset of $\langle n \rangle$ in $\mathbb{Z}$:

$$a + \langle n \rangle = \{\, a + kn \,|\, k \in \mathbb{Z} \,\}$$

(since $\langle n \rangle = \{\, kn \,|\, k \in \mathbb{Z} \,\}$). For example,

$$-7 + \langle 12 \rangle = \{\, \ldots, -19, -7, 5, 17, 29, 41, \ldots \,\}.$$

Recall from Prop. 3.3.2 that b is of the form $a+kn$ for some $k \in \mathbb{Z}$ if and only if $b \equiv a \bmod n$; i.e., if and only b is in the congrunce class $[a]_n$. So the left coset $a + \langle n \rangle$ is simply the congruence class $[a]_n$. In particular each integer is in exactly one of these left cosets; there is no overlap among them.

There is a completely analogous notion of *right cosets*. If H is a subgroup of a group G and g is an element of G, the set $Hg = \{\, hg \,|\, h \in H \,\}$ is called a **right coset** of H in G. For example, if $G = D_3$ and $H = \{e, \phi_1\}$, then $H\rho_1 = \{e\rho_1, \phi_1\rho_1\} = \{\rho_1, \phi_2\}$. Similarly computing all the right cosets, we find they are:

$$\begin{array}{rcccc} He & = & H\phi_1 & = & \{e, \phi_1\}, \\ H\rho_1 & = & H\phi_2 & = & \{\rho_1, \phi_2\} \\ \text{and } H\rho_2 & = & H\phi_3 & = & \{\rho_2, \phi_3\}. \end{array}$$

Note that there are 3 right cosets (the same number as there were left cosets, computed in Example 4.9.2), and that $H = He$ is a right coset (as well as left coset), but that the other two right cosets do *not* coincide with any left cosets.

For our purposes, it will suffice to work systematically with left cosets. We leave some properties and computations of right cosets as exercises. Of course if G is abelian, then $gH = Hg$, so there is no difference between left cosets and right cosets, and we could just call gH a *coset*.

The following proposition establishes a basic property of left cosets.

Proposition 4.9.4 *Suppose that G is a group, H is a subgroup of G, and g and g' are elements of G. Then the following are equivalent:*

1. $g'H = gH$;
2. $g' \in gH$;
3. $g^{-1}g' \in H$.

Proof. We first show that 1) $\Rightarrow$ 2): Note that $g' = g'e \in g'H$ (since $e \in H$), so if $g'H = gH$, then $g' \in gH$.

Now we show that 2) $\Rightarrow$ 3): If $g' \in gH$, then $g' = gh$ for some $h \in H$. Therefore $g^{-1}g' = g^{-1}gh = eh = h \in H$.

Finally we show that 3) $\Rightarrow$ 1). So we assume that $g^{-1}g' \in H$ and we will show that $g'H = gH$. Let $h = g^{-1}g'$. Note that this equation implies that $gh = g'$, and that $g = g'h^{-1}$.

To prove that $g'H = gH$, we will show that $g'H \subseteq gH$ *and* $gH \subseteq g'H$. Suppose that $x \in g'H$, so $x = g'h'$ for some $h' \in H$ (note we didn't say $g'h$ because we've already used h to denote a particular element of H, namely $g^{-1}g'$). Substituting $g' = gh$ into $x = g'h'$ gives $x = (gh)h'$. Since H is a subgroup of G and $h, h' \in H$, we have $hh' \in H$, so $x = (gh)h' = g(hh') \in gH$. We have now shown that $g'H \subseteq gH$.

Now suppose $y \in gH$, so $y = gh''$ for some $h'' \in H$. Substituting $g = g'h^{-1}$ gives $y = g'h^{-1}h''$. Since H is a subgroup and $h, h'' \in H$, we have $h^{-1}h'' \in H$, so $y = g'h^{-1}h'' \in g'H$. We have now also shown that $gH \subseteq g'H$, so now it follows that $gH = g'H$.

We have now shown that 1) $\Rightarrow$ 2) $\Rightarrow$ 3) $\Rightarrow$ 1). It follows that the 3 assertions are equivalent, since we can get from any one of them to any of the others by a sequence of proven implications. $\square$

In the context of Example 4.9.3, Prop. 4.9.4 says that

$$[b]_n = [a]_n \quad \Leftrightarrow \quad b \in [a]_n \quad \Leftrightarrow \quad b \equiv a \bmod n,$$

which we already knew.

Corollary 4.9.5 *Suppose that G is a group, H is a subgroup of G and g is an element of G. Then g is in exactly one left coset of H in G, namely gH.*

Proof. We have $g = ge \in gH$, so $g \in gH$. To see that this is the *only* left coset of H in G containing g, suppose that g is in the left coset $g'H$, where $g' \in G$. Prop. 4.9.4 (with the roles of g and g' reversed) shows that in fact $gH = g'H$. $\square$

Since we now know that gH is the *only* left coset of H in G containing g, we can call it *the* left coset of H in G containing g,

We saw the assertion of the corollary explicitly in Examples 4.9.2 and 4.9.3, but let's consider one more example:

Example 4.9.6 Consider

$$G = \mathbb{Z}_{13}^{\times} = \{[1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12]\}$$

and $H = \langle [3] \rangle = \{[1], [3], [9]\}$. (Since $3^3 \equiv 1 \bmod 13$, we get that $[3]$ has order 3 in G and there are 3 elements in $\langle [3] \rangle$.) We know that $H = eH$ is a

left coset of H in G. Another left coset is $[2]H = \{[2], [6], [5]\}$. We now know $[a]H$ for $a = [1], [2], [3], [5], [6], [9]$. We find also that $[4]H = \{[4], [12], [10]\}$ and $[7]H = \{[7], [8], [11]\}$. We've now accounted for all the elements of G in some left coset of H in G, so we have a complete list of the left cosets:

$$\{[1], [3], [9]\}, \quad \{[2], [5], [6]\}, \quad \{[4], [10], [12]\} \quad \text{and} \quad \{[7], [8], [11]\}.$$

4.10 Lagrange's Theorem

We now turn our attention to Lagrange's Theorem. Recall that the *order* of a finite group is simply the number of elements in the group. If G has order n, then every subset of G has at most n elements, so in particular any *subgroup* of G has order *at most* n. Lagrange's Theorem tells us something much stronger. Namely that if H is a subgroup of G, then the order of H is in fact a *divisor* of n.

The key idea in the proof of Lagrange's Theorem is to show that G can be divided into disjoint subsets each having the same as H. These subsets are precisely the left cosets of H in G. Recall that a left coset of H in G is a subset of G of the form

$$gH = \{\, gh \,|\, h \in H \,\}$$

for some $g \in G$. We showed in Cor. 4.9.5 that each element of G is in exactly one left coset of H. The other ingredient we need for the proof of Lagange's Theorem is that the left cosets of H in G all have the same size.

Lemma 4.10.1 *Suppose that H is subgroup of a group G, and that H has finite order d. Then every left coset of H in G has d elements.*

Proof. Suppose that gH is a left coset of H in G. We will show that the function $f : H \to gH$ defined by $f(h) = gh$ is bijective. It is surjective since the range of f is all of gH (by the definition of gH, each element has the form $gh = f(h)$ for some $h \in H$). The function is injective since $f(h) = f(h')$ means $gh = gh'$, which implies $h = h'$ by the Cancellation Law. Therefore f is bijective, so H and gH have the same number of elements. □

We are now ready to prove Lagrange's Theorem.

Theorem 4.10.2 *Suppose that G is a group of finite order. If H is a subgroup of G, then the order of G is divisible by the order of H.*

Proof. The proof is now a simple counting argument using Cor. 4.9.5 and Lemma 4.10.1.

Let n be the order of G and let d be the order of H. According to Cor. 4.9.5, each element of G is in exactly one left coset of H in G. So n, the number of elements of G, is gotten by adding up the numbers of elements in these left cosets. But Lemma 4.10.1 states that the number of elements in each coset is d. Therefore

$$n = \underbrace{d + d + \cdots + d}_{k \text{ times}},$$

where k is the number of left cosets of H in G. Therefore $n = kd$, so n is divisible by d. □

The idea of the proof is already visible in our computations of cosets in Examples 4.9.2 and 4.9.3. In Example 4.9.2 where $G = D_3$ and $H = \{e, \phi_1\}$, we saw that the 6 elements of D_3 were divided into the 3 left cosets of H in G, each of which had exactly 2 elements (2 being the order of H), so $6 = 3 \cdot 2$. Similarly for $G = \mathbb{Z}_{13}^{\times}$ and $H = \{1, 4, 9\}$, we found that the 12 elements of G were divided into the 4 cosets of H in G, each of which had 3 elements, so $12 = 4 \cdot 3$.

Definition 4.10.3 If H is a subgroup of a group G, then the number of left cosets of H in G is called the **index** of H in G, and denoted $[G : H]$.

If G is finite, then we see that $[G : H]$ is the number $k = n/d$ in the above proof of Lagrange's Theorem. For example, the index of $\{e, \phi_1\}$ in D_3 is 3. But if G is infinite we can still define the index of a subgroup H, and this index may be finite or infinite. For example, if $n > 0$, then the index of $\langle n \rangle$ in $\mathbb{Z}$ is the number of left cosets of $\langle n \rangle$ in $\mathbb{Z}$, i.e., the number of congruence classes modulo n, which is simply n. The subgroup $\{0\}$ has infinite index in $\mathbb{Z}$, but for a more interesting example with infinite index, consider the subgroup $\mathrm{SL}_2(\mathbb{R})$ of $\mathrm{GL}_2(\mathbb{R})$ If $A \in \mathrm{GL}_2(\mathbb{R})$, then the left coset

$$A\mathrm{SL}_2(\mathbb{R}) = \{\, AB \mid B \in SL_2(\mathbb{R}) \,\}$$

is the set of 2×2-matrices with the same determinant as A since

$$C \in A\mathrm{SL}_2(\mathbb{R}) \Leftrightarrow A^{-1}C \in \mathrm{SL}_2(\mathbb{R}) \Leftrightarrow \det(A^{-1}C) = 1 \Leftrightarrow \det A = \det C.$$

So there is one left coset for each possible determinant; i.e., for each non-zero real number.

Lagrange's Theorem is a statement about the order of any *subgroup* of G, but it also tells us something about the order of any *element* of G. Recall if $g \in G$, then the *order* of g is the smallest positive integer d such that $g^d = e$.

Corollary 4.10.4 *Suppose that G is a group of finite order n, and that $g \in G$. Then the order of g is a divisor of n.*

Proof. Recall (Prop. 4.8.3) that the order of g is the same as the order of $\langle g \rangle$, the subgroup of G generated by g. So apply Lagrange's Theorem to $H = \langle g \rangle$ to conclude that the order of g divides the order of G. □

Recall for example that the possible orders of elements of D_3 are 1 (the identity), 2 (the three rotations) and 3 (the two rotations). These are all divisors of 6 (the order of D_3), confirming what Cor. 4.10.4 says in this example. Note that Cor. 4.10.4 does *not* say that every positive divisor of n is the order of an element of G. The order of D_3 is 6, which has divisors 1, 2, 3 and 6. While D_3 has elements of orders 1, 2 and 3, it has no element of order of 6. (Indeed if it did, the group would have to be cyclic, but it is not even abelian.) Similarly, Lagrange Theorem (4.10.2) does *not* say that every divisor of n occurs as the order of a subgroup of G. It happens to be the case for $G = D_3$ that it has subgroups of orders 1, 2, 3 and 6, but we'll see examples later where not every divisor occurs.

Here's another consequence of Thm. 4.10.2:

Corollary 4.10.5 *Suppose that G is a group of order p, where p is a prime number. Then G is cyclic.*

Proof. Since $p > 1$, we can choose some element $g \in G$ with $g \neq e$. Consider the subgroup $\langle g \rangle$ of G, and let d be its order. Then $d|p$ by Thm. 4.10.2 and $d > 1$ since e and g are elements of $\langle g \rangle$. Therefore $d = p$, so $\langle g \rangle = G$ is cyclic. □

Here's an immediate consequence of Cor. 4.10.4:

Corollary 4.10.6 *Suppose that G is a group of order n, and g is an element of G. Then $g^n = e$.*

Proof. Let d denote the order of g. Cor. 4.10.4 tells us that $d|n$, i.e., that $n = dk$ for some $k \in \mathbb{Z}$. Therefore $g^n = g^{dk} = (g^d)^k = e^k = e$. □

The following consequence of Lagrange's Theorem is called **Fermat's Little Theorem**:

Corollary 4.10.7 *Suppose that p is a prime number and a is an integer.*

1. *$a^p \equiv a \bmod p$;*

2. *if a is not divisible by p, then $a^{p-1} \equiv 1 \bmod p$.*

Proof. We prove Part 2) first. If a is not divisible by p, then $\gcd(a,p) = 1$, so $[a]_p \in \mathbb{Z}_p^\times$. Since $\mathbb{Z}_p^\times$ has order $p-1$, Cor. 4.10.6 shows that

$$[a^{p-1}]_p = [a]_p^{p-1} = [1]_p,$$

which means that $a^{p-1} \equiv 1 \bmod p$.

Now let's deduce Part 1). If $p \not| a$, then Part 2) shows that $a^{p-1} \equiv 1 \bmod p$, from which it follows that $a^p \equiv a \bmod p$. On the other hand if $p|a$, then $p|a^p$, so $a^p \equiv a \equiv 0 \bmod p$. □

Fermat's Little Theorem lets you do some fun calculations of remainders. For example, let's compute the remainder 50^{50} on division by 13. We can first simplify the problem by noting that $50 \equiv -2 \bmod 13$, so $50^{50} \equiv (-2)^{50} \bmod 13$. Since -2 is not divisible by 13, Fermat's Little Theorem tells us that $(-2)^{12} \equiv 1 \bmod 13$, but better yet, $(-2)^{12k} \equiv 1^k \equiv 1 \bmod 13$ for every $k \in \mathbb{N}$. Therefore

$$(-2)^{50} = (-2)^{48}(-2)^2 \equiv (-2)^2 \equiv 4 \bmod 13,$$

so the remainder of 50^{50} on division by 13 is 4. Alternatively, in terms of residue classes modulo 13 the calculation becomes

$$\begin{aligned} [50^{50}] = [50]^{50} &= [-2]^{50} = [-2]^{48}[-2]^2 \\ &= \left([-2]^{12}\right)^4 [(-2)^2] = [1]^4[4] = [4]. \end{aligned}$$

For another example, let's compute 50^{100} mod 103. Now 103 is prime, and 50 is not divisible by 103, so by Fermat's Little Theorem, $50^{102} \equiv 1 \bmod 103$. At first glance that doesn't seem to help much, but we can think of this as saying that

$$50^{100} \cdot 50^2 \equiv 1 \pmod{103},$$

or that $[50^{100}]$ is the inverse of $[50^2]$ in $\mathbb{Z}_{103}^\times$. So we could compute 50^2, find the remainder on division by 13, and then use the Euclidean Algorithm to find the multiplicative inverse. (An alternative would be to first find the multiplicative inverse of 50 and then square; either works.) Since $2500 \equiv 28 \bmod 103$, we compute:

$$103 = 3 \cdot 28 + 19, \quad 28 = 19 + 9, \quad 19 = 2 \cdot 9 + 1,$$

giving

$$\begin{aligned} 1 = 19 - 2 \cdot 9 &= 19 - 2(28 - 19) = 3 \cdot 19 - 2 \cdot 28 \\ &= 3(103 - 3 \cdot 28) - 2 \cdot 28 = 3 \cdot 103 - 11 \cdot 28. \end{aligned}$$

Therefore $50^{100} \equiv -11 \equiv 92 \bmod 103$, so the remainder is 92. Again in terms of residue classes modulo 103

$$\begin{aligned}[50^{100}] &= [50]^{100} = [50]^{102}[50]^{-2} \\ &= [50]^{-2} = [50^2]^{-1} = [28]^{-1} = [-11] = [92],\end{aligned}$$

where $[a]^{-2}$ means $([a]^2)^{-1}$, and this is computed as above using the Euclidean algorithm.

4.11 Product groups

Before defining product groups, we need the notion of the *product* of two *sets*.

Definition 4.11.1 Suppose that A and B are sets. The **product** of A and B is defined to be the set

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

Thus an element of $A \times B$ is an *ordered pair* (a, b), where a is an element of A and b is an element of B. You should already be familiar with the example with $A = B = \mathbb{R}$; their product is the plane $\mathbb{R} \times \mathbb{R}$, often denoted $\mathbb{R}^2$. This construction makes sense for any sets A and B. If it happens that $A = B$, we might write A^2 instead of $A \times A$ (and more generally A^n for the *n-fold product* $A \times A \times \cdots \times A$), but let's consider an example where A and B are different: If $A = \{0, 1\}$ and $B = \{0, 1, 2\}$, then

$$A \times B = \{(0,0), (0,1), (0,2), (1,0), (1,1), (1,2)\}.$$

Now if G and H are *groups*, we will define a binary operation on the product

$$G \times H = \{\, (g, h) \mid g \in G,\ h \in H \,\}$$

making it into a group. The two groups G and H have binary operations we'll temporarily denote by $*_G$ and $*_H$. Now define a binary operation $*$ on $G \times H$ by

$$(g, h) * (g', h') = (g *_G g', h *_H h').$$

Note that in the first coordinate we applied the binary $*_G$ to the elements $g, g' \in G$ and got an element $g *_G g' \in G$, and in the second coordinate we applied $*_H$ to elements of H to get $h *_H h' \in H$, so that $*$ is indeed a binary operation on $G \times H$. Now as usual, we'll suppress the symbols for the binary operations when working with abstract groups.

Proposition 4.11.2 *If G and H are groups, then $G \times H$ is a group under the binary operation defined above.*

Proof. First we check that the binary operation on $G \times H$ is associative. So suppose $(g, h), (g', h'), (g'', h'') \in G \times H$. From the definition of the binary operation, we get

$$(g, h)((g', h')(g'', h'')) = (g, h)(g'g'', h'h'') = (g(g'g''), h(h'h'')),$$

and similarly

$$((g, h)(g', h'))(g'', h'') = (gg', hh')(g'', h'') = ((gg')g''), (hh')h'').$$

Since G and H are groups, the binary operations on G and H are associative, so $g(g'g'') = (gg')g''$ and $h(h'h'') = (hh')h''$. Therefore

$$(g(g'g''), h(h'h'')) = ((gg')g''), (hh')h''),$$

showing that the binary operation on $G \times H$ is indeed associative.

Next we must show that there is an identity element for the binary operation on $G \times H$. Since G and H are groups, there are identity elements $e_G \in G$ and $e_H \in H$ for their binary operations, so

$$e_G g = g = g e_G \qquad \text{and} \qquad e_H h = h = h e_H$$

for all $g \in G$ and $h \in H$. It follows that

$$(e_G, e_H)(g, h) = (e_G g, e_H h) = (g, h) = (g e_G, h e_H) = (g, h)(e_G, e_H)$$

for all $(g, h) \in G \times H$. This shows that the element $e = (e_G, e_H) \in G \times H$ is an identity element for the binary operation on $G \times H$.

Finally we have to show that every element of $G \times H$ has an inverse under the binary operation. So suppose that $(g, h) \in G \times H$. Since G is a group, g has an inverse g^{-1} in G, and similarly h has an inverse $h^{-1} \in H$. Now $(g^{-1}, h^{-1}) \in G \times H$ is an inverse of (g, h) since

$$(g, h)(g^{-1}, h^{-1}) = (gg^{-1}, hh^{-1}) = (e_G, e_H) = (g^{-1}g, h^{-1}h) = (g^{-1}, h^{-1})(g, h).$$

We have now shown that $G \times H$ is a group. □

Example 4.11.3 Suppose that m and n are positive integers, and consider the groups $\mathbb{Z}_m$ (under addition modulo m) and $\mathbb{Z}_n$ (under addition modulo n). We can then form the product group $\mathbb{Z}_m \times \mathbb{Z}_n$. Since $\mathbb{Z}_m$ has m elements and $\mathbb{Z}_n$ has n elements, it follows that $\mathbb{Z}_m \times \mathbb{Z}_n$ has mn elements.

For example, if $m = 2$ and $n = 3$, then $\mathbb{Z}_m$ and $\mathbb{Z}_n$ are precisely the sets A and B in the example above, so

$$\mathbb{Z}_2 \times \mathbb{Z}_3 = \{([0]_2, [0]_3), ([0]_2, [1]_3), ([0]_2, [2]_3), ([1]_2, [0]_3), ([1]_2, [1]_3), ([1]_2, [2]_3)\}.$$

The resulting binary operation is given in the table:

	$([0]_2, [0]_3)$	$([0]_2, [1]_3)$	$([0]_2, [2]_3)$	$([1]_2, [0]_3)$	$([1]_2, [1]_3)$	$([1]_2, [2]_3)$
$([0]_2, [0]_3)$	$([0]_2, [0]_3)$	$([0]_2, [1]_3)$	$([0]_2, [2]_3)$	$([1]_2, [0]_3)$	$([1]_2, [1]_3)$	$([1]_2, [2]_3)$
$([0]_2, [1]_3)$	$([0]_2, [1]_3)$	$([0]_2, [2]_3)$	$([0]_2, [0]_3)$	$([1]_2, [1]_3)$	$([1]_2, [2]_3)$	$([1]_2, [0]_3)$
$([0]_2, [2]_3)$	$([0]_2, [2]_3)$	$([0]_2, [0]_3)$	$([0]_2, [1]_3)$	$([1]_2, [2]_3)$	$([1]_2, [0]_3)$	$([1]_2, [1]_3)$
$([1]_2, [0]_3)$	$([1]_2, [0]_3)$	$([1]_2, [1]_3)$	$([1]_2, [2]_3)$	$([0]_2, [0]_3)$	$([0]_2, [1]_3)$	$([0]_2, [2]_3)$
$([1]_2, [1]_3)$	$([1]_2, [1]_3)$	$([1]_2, [2]_3)$	$([1]_2, [0]_3)$	$([0]_2, [1]_3)$	$([0]_2, [2]_3)$	$([0]_2, [0]_3)$
$([1]_2, [2]_3)$	$([1]_2, [2]_3)$	$([1]_2, [0]_3)$	$([1]_2, [1]_3)$	$([0]_2, [2]_3)$	$([0]_2, [0]_3)$	$([0]_2, [1]_3)$.

Note that the element $([1]_2, [1]_3)$ has order 6 since its positive multiples are $2([1]_2, [1]_3) = ([0]_2, [2]_3)$, $3([1]_2, [1]_3) = ([1]_2, [0]_3)$, $4([1]_2, [1]_3) = ([0]_2, [1]_3)$, $5([1]_2, [1]_3) = ([1]_2, [2]_3)$ and finally $6([1]_2, [1]_3) = ([0]_2, [0]_3)$ is the identity. Therefore $\mathbb{Z}_2 \times \mathbb{Z}_3$ is cyclic.

Consider on the other hand $\mathbb{Z}_2 \times \mathbb{Z}_4$. This group has 8 elements, but it's easy to see that $4(a, b) = (4 \cdot a, 4 \cdot b) = ([0]_2, [0]_4)$, so there are no elements of order 8, i.e., the group is not cyclic.

Example 4.11.4 For another example of a product group, consider $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ (so $G = H = \mathbb{R}$, as a group under addition). The binary operation on the product is then defined by

$$(x, y) + (x', y') = (x + x', y' + y'),$$

so this is just the usual vector addition on $\mathbb{R}^2$.

Consider also the group

$$\mathbb{Z} \times \mathbb{Z} = \{\, (a, b) \mid a, b \in \mathbb{Z} \,\}.$$

The operation is defined in the same way as for $\mathbb{R} \times \mathbb{R}$; in fact $\mathbb{Z} \times \mathbb{Z}$ is a subgroup of $\mathbb{R} \times \mathbb{R}$, and can be viewed as the set of points in the plane with *integer* coordinates. (In general if G and H are groups, and K is a subgroup of G and L is a subgroup of H, then $K \times L$ is a subgroup of $G \times H$.)

4.12 Homomorphisms

Roughly speaking, a *homomorphism* is a function from one group to another that is compatible with their algebraic structure. Since more than one group is involved in defining this notion, it's helpful at first to have symbols in place for their binary operations.

Definition 4.12.1 Let $(G, *_G)$ and $(H, *_H)$ be groups. A function $\phi : G \to H$ is a **homomorphism** (of **groups**) if

$$\phi(g *_G g') = \phi(g) *_H \phi(g') \qquad \text{for all } g, g' \in G.$$

Note that g and g' are elements of the group G (the domain of ϕ), so it makes sense to apply $*_G$ to them. This gives an element $g *_G g' \in G$, and as this is the domain of ϕ, we can apply ϕ to $g *_G g'$ to get an element $\phi(g *_G g')$ of the codomain H. On the other hand, it also makes sense to apply ϕ to g and g', giving elements $\phi(g), \phi(g') \in H$. We can then apply the binary operation $*_H$ on H to get an element $\phi(g) *_H \phi(g') \in H$. The definition says that ϕ is a homomorphism if for every pair of elements $g, g' \in G$, these two different procedures:

- apply $*_G$ and then ϕ,
- apply ϕ and then $*_H$,

always give the same element of H. So ϕ carries any "product" in G to the corresponding "product" in H (where here product is refering to the output of the relevant binary operation). When we drop the symbols for the binary operations, the criterion for ϕ to be a homomorphism is that

$$\phi(gg') = \phi(g)\phi(g').$$

Example 4.12.2 Let n be a positive integer, and define $\phi : \mathbb{Z} \to \mathbb{Z}_n$ by $\phi(a) = [a]_n$. To show that ϕ is a homomorphism, we need to check that

$$\phi(a + b) = \phi(a) + \phi(b) \qquad \text{for all } a, b \in \mathbb{Z}$$

(where $+$ on the left-hand-side is the usual addition on $\mathbb{Z}$, and $+$ on the right-hand-side is the addition operation on $\mathbb{Z}_n$). Indeed

$$\phi(a + b) = [a + b]_n = [a]_n + [b]_n = \phi(a) + \phi(b)$$

for all $a, b \in \mathbb{Z}$, where the middle equality is just the *definition* of the operation $+$ on $\mathbb{Z}_n$.

Example 4.12.3 If H is a subgroup of a group G, then the *inclusion* function $i : H \to G$ defined by $i(h) = h$ is a homomorphism, since $i(hh') = hh'$. (Recall the binary operation on a subgroup H is the *same* as the binary operation on G.)

Example 4.12.4 The determinant function $\det : \mathrm{GL}_2(\mathbb{R}) \to \mathbb{R}^\times$ is a homomorphism since $\det(AB) = \det(A)\det(B)$ for all $A, B \in \mathrm{GL}_2(\mathbb{R})$.

For a *non-example* of a homomorphism, consider the determinant as a function from $M_2(\mathbb{R})$ to $\mathbb{R}$. For this to be a homomorphism, we would need $\det(A+B)$ to be the same as $\det(A) + \det(B)$ for all $A, B \in M_2(\mathbb{R})$, but this isn't the case. For instance if $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $B = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$, then $\det(A+B) = \det(I) = 1$, but $\det(A) = \det(B) = 0$, so $\det A + \det B = 0 + 0 = 0$.

Example 4.12.5 Let $\exp : \mathbb{R} \to \mathbb{R}^\times$ be the exponential function $\exp(x) = e^x$. Now the binary operation on the domain $\mathbb{R}$ is addition, and the binary operation on the codomian $\mathbb{R}^\times$ is multiplication. Since

$$\exp(x+y) = e^{x+y} = e^x e^y = \exp(x)\exp(y) \quad \text{for all } x, y \in \mathbb{R},$$

we see that exp is a homomorphism.

Example 4.12.6 We will now define a homomorphism from D_n to S_n by considering the effect of a symmetry on the set of vertices of the regular n-gon. If $\sigma \in D_n$, then for each vertex P_i (where $1 \le i \le n$), $\sigma(P_i)$ is also a vertex. So restricting σ to the set of vertices $V = \{P_1, P_2, \ldots, P_n\}$ gives a function from V to V, which is still injective, hence is bijective (since V is finite). Therefore σ defines a permutation of V. To get an element $f(\sigma) \in S_n$ (the set of permutations of $\{1, 2, \ldots, n\}$), we replace the vertex P_i by i; in other words, we let $f(\sigma) = \tau$, where $\tau \in S_n$ is the permutation such that $\sigma(P_i) = P_{\tau(i)}$ for $i = 1, 2, \ldots, n$. We have now defined a function $f : D_n \to S_n$. For example if $n = 6$, then the function $f : D_6 \to S_6$ is described explicitly in the table, where ρ is a 60° clockwise rotation and ϕ is the reflection in the axis through P_1 (and P_4):

σ	$f(\sigma)$
e	e
ρ	(123456)
ρ^2	(135)(246)
ρ^3	(14)(25)(36)
ρ^4	(153)(264)
ρ^5	(165432)

σ	$f(\sigma)$
ϕ	(26)(35)
$\phi\rho$	(16)(25)(34)
$\phi\rho^2$	(15)(24)
$\phi\rho^3$	(14)(23)(56)
$\phi\rho^4$	(13)(46)
$\phi\rho^5$	(12)(36)(45).

To see that f is a homomorphism, suppose that $\sigma, \sigma' \in D_n$. Let $\tau = f(\sigma)$ and $\tau' = f(\sigma')$. To compute $f(\sigma\sigma')$, we compute $\sigma\sigma'(P_i)$, which is

$$\sigma(\sigma'(P_i)) = \sigma(P_{\tau'(i)}) = P_{\tau(\tau'(i))} = P_{\tau\tau'(i)}.$$

Therefore $f(\sigma\sigma') = f(\sigma)f(\sigma')$ for all $\sigma, \sigma' \in D_n$, and f is a homomorphism.

Example 4.12.7 Now fix an integer n and consider the function $\phi : \mathbb{Z} \to \mathbb{Z}$ defined by $\phi(a) = na$, i.e., ϕ is *multiplication by* n. Then ϕ is a homomorphism since

$$\phi(a+b) = n(a+b) = na + nb = \phi(a) + \phi(b) \quad \text{for all } a, b \in \mathbb{Z}.$$

Example 4.12.8 Now let G be any group and g an element of G. We define $\phi : \mathbb{Z} \to G$ by $\phi(a) = g^a$. Then ϕ is a homomorphism since

$$\phi(a+b) = g^{a+b} = g^a g^b = \phi(a)\phi(b) \quad \text{for all } a, b \in \mathbb{Z}.$$

(the middle equality being a consequence of Prop. 4.5.1, the law of exponents in groups). Note that the preceding example is a special case of this one, with $G = \mathbb{Z}$ and $g = n$.

Proposition 4.12.9 *If $\phi : G \to H$ is a homomorphism of groups, then*

1. $\phi(e_G) = e_H$;
2. $\phi(g^{-1}) = (\phi(g))^{-1}$.

Proof. 1) Since ϕ is a homomorphism and e_G and e_H are the identity elements in the respective groups, we have

$$\phi(e_G)\phi(e_G) = \phi(e_G e_G) = \phi(e_G) = e_H \phi(e_G).$$

Therefore the Cancellation Law (Prop. 4.4.4) implies that $\phi(e_G) = e_H$.

2) Since ϕ is a homomorphism, we have

$$\phi(g)\phi(g^{-1}) = \phi(gg^{-1}) = \phi(e_G) = e_H$$

(the last equality by Part 1). Therefore Prop. 4.4.6 implies that $\phi(g^{-1}) = (\phi(g))^{-1}$. (Note that the inverse on the left-hand-side is with respect to the operation on G, and on the right it is with respect to the operation on H.) □

The parts of the proposition can be viewed as special cases ($n = 0$ and $n = -1$) of the following general property of homomorphisms, for which the proof is left as an exercise.

Proposition 4.12.10 *Suppose $\phi : G \to H$ is a homomorphism of groups. If $g \in G$ and $n \in \mathbb{Z}$, then $\phi(g)^n = \phi(g^n)$.*

We also note another general property of homomorphisms whose proof is left as an exercise: the composite of two homomorphisms is a homomorphisms.

Proposition 4.12.11 *Suppose that G, H and K are groups, and that $\phi : G \to H$ and $\psi : H \to K$ are homomorphisms. Then $\psi \circ \phi : G \to K$ is a homomorphism.*

We now turn to a special class of homomorphisms:

Definition 4.12.12 If G and H are groups, then a function $\phi : G \to H$ is an **isomorphism** if it is a bijective homomorphism.

Example 4.12.13 Recall that in Example 4.12.6, we defined a homomorphism $f : D_n \to S_n$ for all $n \geq 3$. If $\sigma \in D_n$, then $f(\sigma)$ is given by the correponding permutation of the vertices of the n-gon. This homomorphism is always injective since σ is determined by its effect on the vertices, i.e., by the values of $\sigma(P_i)$ for $i = 1, 2, \ldots, n$. So if $f(\sigma) = f(\sigma')$, then $\sigma(P_i) = \sigma'(P_i)$ for all i, and this implies that $\sigma = \sigma'$.

In the case $n = 3$, the groups D_3 and S_3 each have 6 elements, so f must be bijective. Therefore the function $f : D_3 \to S_3$ is an isomorphism. With our usual notation for elements of D_3, we have:

σ	$f(\sigma)$
e	e
ρ	(123)
ρ^2	(132)

σ	$f(\sigma)$
ϕ	(23)
$\phi\rho$	(13)
$\phi\rho^2$	(12).

For $n > 3$, the group S_n has more elements than D_n, so f cannot be an isomorphism; however f does define an isomorphism to a *subgroup* of D_n.

Example 4.12.14 Define $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{C}$ by $\phi((x, y)) = x + iy$ for $x, y \in \mathbb{R}$. Then ϕ is bijective because, by definition, every complex number z has a unique expression in the form $z = x + iy$ with $x, y \in \mathbb{R}$, so $z = \phi((x, y))$ for a unique $(x, y) \in \mathbb{R} \times \mathbb{R}$. Furthermore ϕ is a homomorphism since

$$\begin{aligned} \phi((u, v) + (x, y)) &= \phi((u + x, v + y)) &&= (u + x) + i(v + y) \\ &= (u + iv) + (x + iy) &&= \phi(u, v) + \phi(x, y). \end{aligned}$$

Therefore ϕ is an isomorphism.

On the other hand, consider the function $\phi : \mathbb{R}^\times \times \mathbb{R}^\times \to \mathbb{C}^\times$ defined by $\phi((x, y)) = x + iy$. This function is *not* a homomorphism since $\phi((u, v)(x, y)) = \phi((ux, vy)) = ux + ivy$ is not in general the same as $\phi((u, v))\phi((x, y)) = (u + iv)(x + iy) = (ux - vy) + i(uy + vx)$. (The function also fails to be surjective since $1 = 1 + i \cdot 0$ is not the value of $\phi((x, y))$ for any $x, y \in \mathbb{R}^\times$.)

Example 4.12.15 Consider the groups $\mathbb{Z}_8^\times = \{[1]_8, [3]_8, [5]_8, [7]_8\}$ and $\mathbb{Z}_2 \times \mathbb{Z}_2 = \{([0]_2, [0]_2), ([0]_2, [1]_2), ([1]_2, [0]_2), ([1]_2, [1]_2)\}$. Define $\phi : \mathbb{Z}_8^\times \to \mathbb{Z}_2 \times \mathbb{Z}_2$ by $\phi([1]_8) = ([0]_2, [0]_2)$, $\phi([3]_8) = ([0]_2, [1]_2)$, $\phi([5]_8) = ([1]_2, [0]_2)$ and $\phi([7]_8) = ([1]_2, [1]_2)$. This is clearly a bijection, and comparing the tables:

	$[1]_8$	$[3]_8$	$[5]_8$	$[7]_8$
$[1]_8$	$[1]_8$	$[3]_8$	$[5]_8$	$[7]_8$
$[3]_8$	$[3]_8$	$[1]_8$	$[7]_8$	$[5]_8$
$[5]_8$	$[5]_8$	$[7]_8$	$[1]_8$	$[3]_8$
$[7]_8$	$[7]_8$	$[5]_8$	$[3]_8$	$[1]_8$

and

	$([0]_2, [0]_2)$	$([0]_2, [1]_2)$	$([1]_2, [0]_2)$	$([1]_2, [1]_2)$
$([0]_2, [0]_2)$	$([0]_2, [0]_2)$	$([0]_2, [1]_2)$	$([1]_2, [0]_2)$	$([1]_2, [1]_2)$
$([0]_2, [1]_2)$	$([0]_2, [1]_2)$	$([0]_2, [0]_2)$	$([1]_2, [1]_2)$	$([1]_2, [0]_2)$
$([1]_2, [0]_2)$	$([1]_2, [0]_2)$	$([1]_2, [1]_2)$	$([0]_2, [0]_2)$	$([0]_2, [1]_2)$
$([1]_2, [1]_2)$	$([1]_2, [1]_2)$	$([1]_2, [0]_2)$	$([0]_2, [1]_1)$	$([0]_2, [0]_2)$,

shows that ϕ is in fact a homomorphism. (The table for H is gotten from the table for G by replacing each g by $\phi(g)$.) Therefore ϕ is an isomorphism.

We now turn to some general properties of isomorphisms.

Proposition 4.12.16 *Suppose that G, H and K are groups, and that $\phi : G \to H$ and $\psi : H \to K$ are isomorphisms. Then $\psi \circ \phi : G \to K$ is an isomorphism.*

Proof. Recall (Prop. 4.3.4) that the composite of two bijective functions is bijective, and (Prop. 4.12.11) that the composite of two homomorphisms is a homomorphism. Therefore the composite of two isomorphisms is an isomorphism. □

Proposition 4.12.17 *Suppose that G and H are groups and $\phi : G \to H$ is an isomorphism. Let $\psi = \phi^{-1} : H \to G$ be the inverse function of ϕ. Then ψ is an isomorphism.*

Proof. First note that since ϕ is an isomorphism, it is bijective, hence has an inverse function by Prop. 4.3.5. Moreover the inverse function ψ is bijective (since it too has an inverse function, namely ϕ, and Prop. 4.3.5 states that a function is bijective *if and only if* it has an inverse function). So we only need to show that ψ is a homomorphism, i.e., that $\psi(hh') = \psi(h)\psi(h')$ for all $h, h' \in H$. Since $\phi \circ \psi = \mathrm{id}_H$, we see that

$$\phi(\psi(hh')) = hh' = \phi(\psi(h))\phi(\psi(h')).$$

Since ϕ is a homomorphism, $\phi(gg') = \phi(g)\phi(g')$ for all $g, g' \in G$. Applying this to $g = \psi(h)$, $g' = \psi(h')$ gives

$$\phi(\psi(h)\psi(h')) = \phi(\psi(h))\phi(\psi(h')).$$

we have now shown that

$$\phi(\psi(hh')) = \phi(\psi(h)\psi(h')).$$

Since ϕ is bijective, ϕ is in particular injective, so it follows from the preceding equation that $\psi(hh') = \psi(h)\psi(h')$. Therefore ψ is a bijective homomorphism, i.e., an isomorphism. □

Recall that a *homomorphism* from a group G to group H is a function $\phi : G \to H$ such that

$$\phi(gg') = \phi(g)\phi(g') \quad \text{for all } g, g' \in G.$$

For example, the functions

- $\phi : \mathbb{Z} \to \mathbb{Z}_n$ defined by $\phi(a) = [a]_n$ (Example 4.12.2),
- $\phi : \mathrm{GL}_2(\mathbb{R}) \to \mathbb{R}^\times$ defined by $\phi(A) = \det A$ (Example 4.12.4),
- $\phi : \mathbb{R} \to \mathbb{R}^\times$ defined by $\phi(x) = e^x$ (Example 4.12.5),

are all homomorphisms.

An *isomorphism* is a bijective homomorphisms; for example the functions $f : D_3 \to S_3$ defined in Example 4.12.13 is an isomorphism. So we say that D_3 is *isomorphic* to S_3.

Definition 4.12.18 If G and H are groups, then we say G is **isomorphic** to H if there is an isomorphism $\phi : G \to H$.

Note that a group is isomorphic to itself (by the identity function). According to Prop. 4.12.17 says that G is isomorphic to H if and only if H is isomorphic to G. Since the order doesn't matter, we'll often simply say instead that G and H are isomorphic. Prop. 4.12.16 says that if G and H are isomorphic and H and K are isomorphic, then G and K are isomorphic.

Proposition 4.12.19 *Suppose that G is a cyclic group.*

1. *If G has infinite order, then G is isomorphic to $\mathbb{Z}$.*
2. *If G has order n, then G is isomorphic to $\mathbb{Z}_n$.*

Proof. Let g be a generator of G, so $G = \langle g \rangle$. Then g has the same order as G (Prop. 4.8.3).

1) If G has infinite order, then we define $\phi : \mathbb{Z} \to G$ by $\phi(a) = g^a$ as in Example 4.12.8. Then ϕ is a homomorphism (by the Laws of exponents). The range of ϕ is

$$\{ g^a \mid a \in \mathbb{Z} \} = \langle g \rangle = G,$$

so ϕ is surjective. If $\phi(a) = \phi(b)$, then $g^a = g^b$, so $a = b$ by Cor. 4.6.3. Therefore ϕ is injective. Therefore ϕ is an isomorphism.

2) If G has order n, then we would like to define $\phi : \mathbb{Z}_n \to G$ by $\phi([a]_n) = g^a$, but we first have to check that this makes sense. We have to check that if a and b are integers in the same residue class modulo n (so $[a]_n = [b]_n$), then the formula for the output of ϕ gives the same element of G (i.e., $g^a = g^b$). This is part of Cor 4.6.3, which says in fact that

$$[a]_n = [b]_n \quad \Leftrightarrow g^a = g^b.$$

So not only is ϕ well-defined, but it is injective. We see also that ϕ is surjective as in part 1. Finally, since

$$\phi([a]_n)\phi([b]_n) = g^a g^b = g^{a+b} = \phi([a+b]_n) = \phi([a]_n + [b]_n)$$

for all $[a]_n, [b]_n \in \mathbb{Z}_n$, we see that ϕ is a homomorphism, and therefore an isomorphism. □

We now also have the following corollary of Lagrange's Theorem:

Corollary 4.12.20 *If G is a group of prime order p, then G is isomorphic to $\mathbb{Z}_p$.*

Proof. By Cor. 4.10.5, G is cyclic. Therefore by Prop. 4.12.19, G is isomorphic to $\mathbb{Z}_p$. □

A key point about isomorphisms is that isomorphic groups are essentially interchangeable; they have exactly the same properties (at least as groups). We'll see in a moment how this works in practice for specific properties.

Now we make precise how various properties are shared by isomorphic groups.

Proposition 4.12.21 *Suppose that G and H are isomorphic groups. Then*

1. *G is abelian if and only if H is abelian;*

2. *G is cyclic if and only if H is cyclic.*

Proof. 1) Let $\phi : G \to H$ be an isomorphism. Suppose that G is abelian. We wish to show that H is abelian, i.e., $hh' = h'h$ for all $h, h' \in H$. Since ϕ is surjective, we have $h = \phi(g)$ and $h' = \phi(g')$ for some $g, g' \in G$. Since g is abelian,

$$\begin{aligned} gg' = g'g &\Rightarrow \phi(gg') = \phi(g'g) \\ &\Rightarrow \phi(g)\phi(g') = \phi(g')\phi(g) \quad \text{since } \phi \text{ is a homomorphism)} \\ &\Rightarrow hh' = h'h. \end{aligned}$$

Therefore H is abelian.

Conversely, if H is abelian, then we use the existence of an isomorphism $\psi : H \to G$ to deduce that G is abelian.

2) Let $\phi : G \to H$ be an isomorphism and suppose that $G = \langle g \rangle$ is cyclic. Then every element of G is of the form g^n for some $n \in \mathbb{Z}$. Since ϕ is surjective, every element of H is of the form $\phi(g^n)$ for some $n \in \mathbb{Z}$. By Prop. 4.12.10, $\phi(g^n) = \phi(g)^n$, so every element of H is in $\langle \phi(g) \rangle$. Therefore $H = \langle \phi(g) \rangle$ is cyclic.

As in Part 1), we see that if H is cyclic, then (using the inverse isomorphism ψ) so is G. □

Example 4.12.22 It follows from the proposition that if G is abelian and H is not, then G is *not* isomorphic to H. For example, the groups $\mathbb{Z}_6$ and D_3 both have order 6, but $\mathbb{Z}_6$ is abelian and D_3 is not, so the groups are not isomorphic.

Example 4.12.23 Similarly we see that if G is cyclic and H is not, then G and H are not isomorphic. For example, $\mathbb{Z}_4$ and $\mathbb{Z}_2 \times \mathbb{Z}_2$ are both abelian groups of order 4, but $\mathbb{Z}_4$ is cyclic and $\mathbb{Z}_2 \times \mathbb{Z}_2$ is not. Therefore these groups are *not* isomorphic.

Proposition 4.12.24 *Suppose that G and H are groups and $\phi : G \to H$ is an isomorphism. Then for each $g \in G$, the order of g in G is the same as the order of $\phi(g)$ in H.*

Proof. Suppose that g has finite order d, and let $h = \phi(g)$. Then $g^d = e_G$, so

$$h^d = \phi(g)^d = \phi(g^d) = \phi(e_G) = e_H$$

by Prop. 4.12.10. Therefore d is a multiple of the order of h. On the other hand, $\psi = \phi^{-1} : H \to G$ is an isomorphism and $\psi(h) = g$, so the order of g is a multiple of the order of h. Therefore g and h have the same order.

Since we've shown that if either of g or h has finite order, then so does the other, it also follows that if either has infinite order, then so does the other. □

Example 4.12.25 Consider the two groups S_4 and D_{12}. Both are non-abelian groups of order 24, but D_{12} has an element of order 12 (a 30° rotation) and S_4 does not. Therefore the two groups are not isomorphic.

We now return to general homomorphisms (not necessarily bijective). Given any homomorphism $\phi : G \to H$, we shall associate to it a subgroup of G called the *kernel* of ϕ, and a subgroup of H called the *image* of ϕ. We've already defined the image of ϕ; it's just the range of ϕ, or

$$\phi(G) = \{\, \phi(g) \,|\, g \in G \,\},$$

also denoted image(ϕ). This is by definition a subset of H; now we show it is in fact a *subgroup* of H.

Proposition 4.12.26 *Suppose that G and H are groups and $\phi : G \to H$ is a homomorphism. Then $\phi(G)$ is a subgroup of H.*

Proof. We verify the usual criteria to check that $\phi(G)$ is a subgroup:

1) The identity element is in $\phi(G)$ since $e_H = \phi(e_G) \in \phi(G)$.

2) Suppose that $h, h' \in \phi(G)$. Then $h = \phi(g)$, $h' = \phi(g')$ for some $g, g' \in G$, so

$$hh' = \phi(g)\phi(g') = \phi(gg') \in \phi(G).$$

3) Suppose that $h \in \phi(G)$. Then $h = \phi(g)$ for some $g \in G$. Therefore

$$h^{-1} = \phi(g)^{-1} = \phi(g^{-1}) \in \phi(G).$$

It follows that $\phi(G)$ is a subgroup of H. □

Before giving examples, we define the *kernel* of a homomorphism.

Definition 4.12.27 Suppose that $\phi : G \to H$ is a homomorphism of groups. The **kernel** of ϕ is the following subset of G:

$$\ker(\phi) = \{\, g \in G \,|\, \phi(g) = e_H \,\}.$$

Proposition 4.12.28 *Suppose that G and H are groups and $\phi : G \to H$ is a homomorphism. Then* $\ker(\phi)$ *is a subgroup of G.*

Proof. Again we verify the usual criteria:

1) The identity element e_G is in $\ker(\phi)$ since $\phi(e_G) = e_H$.

2) Suppose that $g, g' \in \ker(\phi)$. Then $\phi(g) = \phi(g') = e_H$, so

$$\phi(gg') = \phi(g)\phi(g') = e_H e_H = e_H.$$

Therefore $gg' \in \ker(\phi)$.

3) Suppose that $g \in \ker(\phi)$. Then $\phi(g) = e_H$, so

$$\phi(g^{-1}) = \phi(g)^{-1} = e_H^{-1} = e_H.$$

Therefore $g^{-1} \in \ker(\phi)$.

It follows that $\ker(\phi)$ is a subgroup of G. □

Example 4.12.29 Recall the homomorphism $\phi : \mathbb{Z} \to \mathbb{Z}_n$ defined by $\phi(a) = [a]_n$ from Example 4.12.2. Then ϕ is surjective since every element of $\mathbb{Z}_n$ is of the form $[a]_n$ for some $a \in \mathbb{Z}$. Therefore the image of ϕ is $\phi(\mathbb{Z}) = \mathbb{Z}_n$.

Now we compute the kernel of ϕ. The criterion for an integer a to be in the kernel is that $\phi(a) = [0]_n$ (the identity element of $\mathbb{Z}_n$). Since $\phi(a) = [a]_n = [0]_n$ if and only if a is divisible by n, the kernel of ϕ is the set of integers divisible by n; i.e.,

$$\ker(\phi) = \langle n \rangle \subseteq \mathbb{Z}.$$

Example 4.12.30 If $\phi : G \to H$ is an *isomorphism*, then ϕ is surjective, so $\phi(G) = H$. And ϕ is injective, so

$$g \in \ker(\phi) \quad \Leftrightarrow \quad \phi(g) = e_H = \phi(e_G) \quad \Leftrightarrow \quad g = e_G.$$

Therefore $\ker(\phi) = \{e_G\}$.

Example 4.12.31 If H is a subgroup of G, then the *inclusion* function $i : H \to G$ defined by $i(h) = h$ is a homomorphism (Example 4.12.3), with $\phi(H) = H$ and $\ker(\phi) = \{e\}$.

Example 4.12.32 Recall that the determinant function $\det : \mathrm{GL}_2(\mathbb{R}) \to \mathbb{R}^\times$ is a homomorphism (Example 4.12.4). The function is surjective since for any $x \in \mathbb{R}^\times$, we have $\begin{pmatrix} x & 0 \\ 0 & 1 \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R})$ and $\det \begin{pmatrix} x & 0 \\ 0 & 1 \end{pmatrix} = x$. Therefore the image of det is $\mathbb{R}^\times$. The kernel of det is

$$\mathrm{SL}_2(\mathbb{R}) = \{\, A \in \mathrm{GL}_2(\mathbb{R}) \mid \det A = 1 \,\}.$$

Example 4.12.33 Let $n \in \mathbb{N}$ and consider the homomorphism $\phi : \mathbb{Z} \to \mathbb{Z}$ defined by $\phi(a) = na$ (Example 4.12.7). The image of ϕ consists of the integer multiples of n, so $\phi(\mathbb{Z}) = \langle n \rangle$. On the other hand, the kernel of n consists of the integers a such that $na = 0$. Since $n \neq 0$, this implies that $a = 0$, so $\ker(\phi) = \{0\}$.

Example 4.12.34 As a final example, suppose that G is a group and $g \in G$, and define $\phi : \mathbb{Z} \to G$ by $\phi(a) = g^a$ (Example 4.12.8). Then the image of ϕ is

$$\phi(\mathbb{Z}) = \{\, g^a \mid a \in \mathbb{Z} \,\} = \langle g \rangle.$$

The kernel of ϕ consists of the integers a such that $g^a = \phi(a) = e$. According to Thm. 4.6.2, we therefore have

$$\ker(\phi) = \langle d \rangle$$

if g has finite order d, and $\ker(\phi) = \{0\}$ if g has infinite order.

4.13 Conjugacy classes

Definition 4.13.1 Let g and g' be elements of a group G. We say that g' is *conjugate* to g (in G) if $g' = hgh^{-1}$ for some $h \in G$.

We give some examples:

- Every $g \in G$ is conjugate to itself in G since $g = ege^{-1}$ where e is the identity element of G.
- If $A, A' \in \mathrm{GL}_n(\mathbb{R})$, then A' is conjugate to A if there exists $B \in \mathrm{GL}_n(\mathbb{R})$ such that $A' = BAB^{-1}$, so the notion becomes the familiar one of similarity of matrices.
- If G is abelian, then $hgh^{-1} = g$ for all $g, h \in G$, so the only element to which g is conjugate is itself.
- In the group S_3, the 2-cycle (12) is conjugate to the 2-cycle (13) since $(13) = (23)(12)(23)^{-1}$.

Before pursuing the notion further, we recall the notion of an equivalence relation, Recall first that a *binary relation* (or simply a *relation*) on a set S is a rule $\sim$ which is satisfied by a set of ordered pairs of elements of S; we write $x \sim y$ if the rule is satisfied by the pair $(x, y) \in S \times S$. (Thus specifying a relation $\sim$ is the same as specifying the subset of $S \times S$ consisting of the pairs (x, y) satisfying $x \sim y$.) Some familiar relations are $=$ on any set S, the usual inequality relation $<, \leq, >, \geq$ on $\mathbb{R}$, and recall that for each positive integer n we defined relation on $\mathbb{Z}$ by stipulating that $x \equiv y \bmod n$ if $n|(x - y)$.

Definition 4.13.2 A relation $\sim$ on a set S is *equivalence relation* if it satisfies the following properties:

1. (reflexivity) if $x \in S$, then $x \sim x$;
2. (symmetry) if $x, y \in S$ and $x \sim y$, then $y \sim x$;
3. (transitivity) if $x, y, z \in S$, $x \sim y$ and $y \sim z$, then $x \sim z$.

Returning to the examples:

- The equality relation $=$ is an equivalence relation on any set S.
- The inequality relations on $\mathbb{R}$ are *not* equivalence relations since they are not symmetric (and $<$ and $>$ are not even reflexive).
- The congruence mod n relation on $\mathbb{Z}$ is an equivalence relation since it is evidently reflexive and symmetric, and Cor. 3.3.3 states that it is transitive.

A useful interpretation of an equivalence relation on a set S is that it partitions S into disjoint subsets.

Definition 4.13.3 If $\sim$ is an equivalence relation on a set S, and $x \in S$, we define the *equivalence class* of x (*under* $\sim$) to be

$$[x] = \{\, y \in S \,|\, y \sim x \,\}.$$

For example, the equivalence class of x under equality is $[x] = \{x\}$. The equivalence of the integer a under congruence modulo n is its congruence class $[a]_n$.

Proposition 4.13.4 *If $\sim$ is an equivalence relation on S, then each element of S belongs to exactly one equivalence class.*

Proof. If $x \in S$, then $x \in [x]$ since $x \sim x$ (by reflexivity). We must prove that if $[y]$ is any equivalence class to which x belongs, then in fact $[x] = [y]$, so suppose that $x \in [y]$, i.e. that $x \sim y$, so that by symmetry we also have $y \sim x$. Then $[y] \subset [x]$ since if $z \in [y]$, then $z \sim y$, and since $y \sim x$, transitivity implies that $z \sim x$ so $z \in [x]$. Similarly $[x] \subset [y]$ since if $z \in [x]$, then transitivity implies that $s \in [y]$. □

. In fact, given any partition of a set S into disjoint subsets, one can define an equivalence relation on S by stipulating that $x \sim y$ whenever x and y belong to the same subset. The equivalence classes then become the subsets forming the partition. Recall that we previously made important use of such a partition on a group G, namely the set of left cosets of a subgroup H. We proved (Cor. 4.9.5) that each element of G belongs to exactly one

left coset of H in G. Recall that g and g' belong to the same left coset of H if and only if $g^{-1}g' \in H$ (Prop. 4.9.4), so the corresponding equivalence relation can be defined by saying that $g \sim g'$ if $g^{-1}g' \in H$. Note that for the subgroup $H = \langle n \rangle$ of $G = \mathbb{Z}$, this gives $a \sim b$ if and only if $b - a$ is divisible by n, i.e., $b \equiv a \bmod n$, recovering the notion of congruence modulo n as an example.

We now return to the notion of conjugacy, which gives another important example of an equivalence relation on a group. For the rest of the section we use $\sim$ to denote the relation defined by conjugacy, i.e., for $g, g' \in G$, we write $g \sim g'$ to mean that g is conjugate to g' in G.

Proposition 4.13.5 *The conjugacy relation $\sim$ is an equivalence relation on G.*

Proof. We verify that $\sim$ satisfies the three properties in Definition 4.13.2:

1. If $g \in G$, then $g \sim g$ since $g = ege^{-1}$.

2. If $g, g' \in G$ and $g \sim g'$, then $g = hg'h^{-1}$ for some $h \in G$, so

$$g' = eg'e^{-1} = (h^{-1}h)g'(h^{-1}h) = h^{-1}(hg'h^{-1})h = h^{-1}g(h^{-1})^{-1}.$$

   Therefore $g' \sim g$ (since $h^{-1} \in G$).

3. Suppose that $g, g', g'' \in G$ are such that $g \sim g'$ and $g' \sim g''$. Then $g = hg'h^{-1}$ for some $h \in G$, and $g' = h'g''(h')^{-1}$ for some $h' \in G$. Therefore

$$g = hg'h^{-1} = h(h'g''(h')^{-1})h^{-1} = (hh')g''(hh')^{-1},$$

   so $g \sim g''$ (since $hh' \in G$).

□

Now we let $[g]$ denote the conjugacy class of g in G, i.e., the equivalence class of g under the relation $\sim$. (Note that in the case $G = \mathbb{Z}$, this should not be confused with a congruence class modulo n, which we initially denoted $[a]_n$ but usually abbreviated to $[a]$.)

Example 4.13.6 If G is abelian, then each $g \in G$ is only conjugate to itself, so $[g] = \{g\}$.

Example 4.13.7 We determine the conjugacy class of each element of $G = D_4$, where as usual we let ρ be a 90° clockwise rotation and ϕ a fixed reflection.

- e is only conjugate to itself, so $[e] = \{e\}$.
- Computing $h\rho h^{-1}$ for each $h \in D_4$, we find that $h\rho h^{-1} = \rho$ if h is a rotation and $h\rho h^{-1} = \rho^3$ if h is any reflection, so $[\rho] = \{\rho, \rho^3\}$. It follows also that $[\rho^3] = [\rho]$.
- Similarly we find that $h\rho^2 h^{-1} = \rho^2$ for all $h \in D_4$, so $[\rho^2] = \{\rho^2\}$.
- Computing $h\phi h^{-1}$ for each $h \in H$, we find that if $h = e$, ϕ, ρ^2 or $\phi\rho^2$, then $h\phi h^{-1} = \phi$, and otherwise $h\phi h^{-1} = \phi\rho^2$, so $[\phi] = \{\phi, \phi\rho^2\} = [\phi\rho^2]$.
- Similarly we find that $[\phi\rho] = \{\phi\rho, \phi\rho^3\} = [\phi\rho^3]$.

Example 4.13.8 In the permutation group $G = S_n$, one finds that the conjugate of a k-cycle is again a k-cycle; more precisely:

$$h(a_1\, a_2\, \cdots\, a_k)h^{-1} = (h(a_1)\, h(a_2)\, \cdots\, h(a_k)$$

(the verification is left as an exercise). For example in S_3, if $h = (23)$, then $h(1) = 1$, $h(2) = 3$ and $h(3) = 2$, so $h(12)h^{-1} = (h(1)\, h(2)) = (13)$. More generally one finds that conjugation preserves cycle structure in the following sense: Recall that each $g \in S_n$ is written as a composite of disjoint cycles of length $k_1, k_2, \ldots, k_r$:

$$(a_1 \cdots a_{k_1})(a_{k_1+1} \cdots a_{k_1+k_2}) \cdots (a_{k_1+k_2+\cdots+k_{r-1}+1} \cdots a_{k_1+k_2+\cdots+k_r}).$$

(Inserting the 1-cycle (a) for each a such that $g(a) = a$ we can ensure that each element of $\{1, 2, \ldots, n\}$ appears exactly once, so $k_1 + k_2 + \cdots + k_r = n$.) We then find that hgh^{-1} is the composite of disjoint cycles of the same lengths as for g:

$$(h(a_1) \cdots h(a_{k_1}))(h(a_{k_1+1}) \cdots h(a_{k_1+k_2})) \cdots (h(a_{k_1+k_2+\cdots+k_{r-1}+1}) \cdots h(a_{k_1+k_2+\cdots+k_r})).$$

Moreover any two elements of S_n with the same cycle structure (i.e., with the same cycle lengths when written as a composite of disjoint cycles) are conjugate, so there is one conjugacy class for each possible cycle structure, and the cycle structures are in bijection with the partitions of n, i.e., the ways of writing n as a sum of positive integers $n = k_1 + k_2 + \cdots + k_r$, which we may reorder so $k_1 \geq k_2 \geq \cdots \geq k_r$.

We illustrate all this for $n = 4$, listing the partitions of n and the corresponding conjugacy classes:

partition	structure	elements of conjugacy class
$1+1+1+1$	$(a)(b)(c)(d)$	e
$2+1+1$	$(ab)(c)(d)$	$(12), (13), (14), (23), (24), (34)$
$3+1$	$(abc)(d)$	$(123), (132), (124), (142), (134), (143), (234), (243)$
4	$(abcd)$	$(1234), (1243), (1324), (1342), (1423), (1432)$
$2+2$	$(ab)(cd)$	$(12)(34), (13)(24), (14)(23)$

Returning to the example of $G = D_4$, note that even though the group is not abelian, the element $g = \rho^2$ has the property that $[g] = \{g\}$, i.e., $hgh^{-1} = g$ for all $h \in G$, i.e., $hg = gh$ for all $h \in G$. In an abelian group all elements have this property, and in every group the identity has this property. More generally we consider the set of elements with this property:

Definition 4.13.9 Let G be a group. The *center* of G is the set

$$Z_G = \{\, g \in G \,|\, gh = hg \text{ for all } h \in G \,\} = \{\, g \in G \,|\, [g] = \{g\} \,\}.$$

If $g \in G$, the *centralizer* of g in G is the set

$$Z_G(g) = \{\, h \in G \,|\, gh = hg \,\} = \{\, h \in G \,|\, g = hgh^{-1} \,\}.$$

We leave the proof of the following proposition as an exercise:

Proposition 4.13.10 *If G is a group and $g \in G$, then Z_G and $Z_G(g)$ are subgroups of G.*

Let us determine the center of G in a few examples:

- If G is abelian, then $Z_G = G$.
- If $G = D_4$, then $Z_G = \{e, \rho^2\}$ (from Example 4.13.7).
- If $G = S_4$, then $Z_G = \{e\}$ (from Example 4.13.8).
- If $G = \mathrm{GL}_2(\mathbb{R})$, then $Z_G = \left\{ \begin{pmatrix} x & 0 \\ 0 & x \end{pmatrix} \,\middle|\, x \in \mathbb{R}^\times \right\}$ (by an exercise).

Turning to centralisers $Z_G(g)$, we have the following connection with conjugacy classes. Recall that if H is a subgroup of G, then the index of H in G, denoted $[G : H]$, is the number of left cosets of H in G; if G is a finite group, then the proof of Lagrange's Theorem shows that $[G : H]$ is the order of G divided by the order of H.

Proposition 4.13.11 *Let G be a finite group, and $g \in G$. Then the number of elements in $[g]$ is the index of $Z_G(g)$ in G. In particular the number of elements of $[g]$ divides the order of G.*

Proof. Let S denote the set of left cosets of $Z_G(g)$ in G. We will define a bijection between S and $[g]$. By definition, the index of $Z_G(g)$ in G is the number of elements of S, so this will prove the proposition.

Let $K = Z_G(g)$, so an element of S is a left coset hK for some $h \in G$. We will define $\alpha : S \to [g]$ by setting $\alpha(hK) = hgh^{-1}$. Then $hgh^{-1} \in [g]$,

but we must check the function α is well-defined, i.e., if $hK = h'K$ then $hgh^{-1} = h'g(h')^{-1}$. So suppose that $h'K = hK$, so $h' = hk$ for some $k \in K$. Since $K = Z_G(g)$, we have $g = kgk^{-1}$, so

$$h'g(h')^{-1} = hkg(hk)^{-1} = h(kgk^{-1})h^{-1} = hgh^{-1}$$

as required.

Now we check that α is bijective. First note that α is surjective since if $g' \in [g]$, then $g' = hgh^{-1} = \alpha(hK)$ is in the range of α. To prove that α is injective, we must prove that that if $\alpha(hK) = \alpha(h'K)$ for $h, h' \in G$, then in fact $hK = h'K$. If $\alpha(hK) = \alpha(h'K)$, then $hgh^{-1} = h'g(h')^{-1}$, so

$$(h^{-1}h')g(h^{-1}h')^{-1} = h^{-1}(h'g(h')^{-1})h = h^{-1}(hgh^{-1})h = g.$$

This implies that $h^{-1}h' \in Z_G(g) = K$, which in turn implies that $h' \in hK$, so $h'K = hK$ as required. □

Take for example the group $G = D_4$. The computations in Example 4.13.7 show:

- $Z_G(e) = Z_G(\rho^2) = G$ (in fact $Z_G(g) = G$ if and only if $g \in Z_G$).
- $Z_G(\rho) = Z_G(\rho^2) = \langle\rho\rangle$
- $Z_G(\phi) = Z_G(\phi\rho^2) = \{e, \rho^2, \phi, \phi\rho^2\}$
- $Z_G(\phi\rho) = Z_G(\phi\rho^3) = \{e, \rho^2, \phi\rho, \phi\rho^3\}$.

Finally let us return the example of $G = S_4$. Note that indeed the cardinality of each conjugacy class $[g]$ divides 24, the order of S_4. According to Prop. 4.13.11 the order of G divided by the cardinality of $[g]$ is the order of $Z_G(g)$, as recorded in the table:

g	e	(12)	(123)	(1234)	(12)(34)
cardinality of $[g]$	1	6	8	6	3
order of $Z_G(g)$	24	4	3	4	8

We can verify explicitly that $Z_G(g)$ has the indicated order for each g above:

- $Z_G(e) = G = S_4$ has order 24;
- $Z_G((12)) = \{e, (12), (34), (12)(34)\}$ has order 4;
- $Z_G((123)) = \langle(123)\rangle$ has order 3;
- $Z_G((1234)) = \langle(1234)\rangle$ has order 4;
- we leave it as an exercise to find the 8 elements of $Z_G((12)(34))$.

Chapter 5

Rings

5.1 Definition of a ring

Up until now, we've mainly been discussing *groups*. The remaining lectures will focus on another type of algebraic structure, called a *ring*. A ring is a set with *two* binary operations satisfying certain axioms. We've actually already worked quite a bit with a particular ring, namely the set of integers with its binary operations of addition *and* multiplication. The division algorithm, Thm. 2.1.2, for example, is really about the *ring* $\mathbb{Z}$, since it involves *both* of these binary operations. The set of integers is the "model" for an abstract ring. The axioms are based on the key properties satisfied by the binary operations of addition and multiplication on $\mathbb{Z}$.

Definition 5.1.1 A **ring** is a set R with binary operations $+$ and $*$ satisfying:

1. $(R, +)$ is an abelian group;
2. the operation $*$ is associative and has an identity element in R;
3. $x * (y + z) = (x * y) + (x * z)$ and $(y + z) * x = (y * x) + (z * x)$ for all $x, y, z \in R$.

Just as we used $(G, *)$ to denote G with its binary operation $*$, we will use $(R, +, *)$ to denote R with its binary operations $+$ and $*t$. Since the first operation plays a role analogous to addition in $\mathbb{Z}$, it is often called the *addition* operation for the ring R and denoted by $+$. Similarly the second operation $*$ is often called the *multiplication* on R, but as with the usual multiplication, the symbol for it is sometimes omitted. Since both operations are associative, we generally don't bother to write parentheses to keep track

of the order in which the operation $+$ or $*$ is performed in a single "sum" or "product." Furthermore, it's understood that multiplications are performed before additions unless parentheses indicate otherwise, so for example $xy+z$ means $(xy)+z$ rather than $x(y+z)$.

Note that we demand that $+$ be associative, commutative, have an identity element and that there be an additive inverse for each element (as $(R,+)$ has to be an abelian group), but we demand less of the $*$ operation, just that it be associative and have an identity element. Note also that the last condition *relates* the operations $+$ and $*$. This relation is called the *distributive law*, or rather *laws*, since there are two equations which need to be satisfied for each x, y and z in R. (We could call one of these laws *right distributive* and the other *left distributive*, but I'll refrain since it's not worth remembering which is which.) Of course the two equations are equivalent if $*$ is commutative, but we haven't assumed this.

5.2 Examples of rings

Of course, the set $\mathbb{Z}$ with the usual addition and multiplication operations is a ring. We've already seen that $(\mathbb{Z},+)$ is an abelian group, we know that multiplication is associative and has identity element 1, and the distributive law holds:

$$a(b+c) = ab + ac \qquad \text{for all } a,b,c \in \mathbb{Z}.$$

Similarly, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are all rings with their usual addition and multiplication operations.

Example 5.2.1 Another example is $\mathbb{Z}_n = \{[0]_n, [1]_n, \ldots, [n-1]_n\}$, the set of residue classes modulo n, with its addition and multiplication operations:

$$[a]_n + [b]_n = [a+b]_n, \qquad [a]_n[b]_n = [ab]_n.$$

We've already seen that $(\mathbb{Z}_n,+)$ is an abelian group, and that multiplication is associative (Prop. 3.3.8) with identity element $[1]_n$. In order to conclude that $\mathbb{Z}_n$ is a ring, we just have to check that $X(Y+Z) = XY + XZ$ for all $X,Y,Z \in \mathbb{Z}_n$, i.e., that

$$[a]_n([b]_n + [c]_n) = [a]_n[b]_n + [c]_n \qquad \text{for all } a,b,c \in \mathbb{Z}.$$

(Since multiplication on $\mathbb{Z}_n$ is commutative, the other distributive law $(Y+Z)X = YX + ZX$ is equivalent.) For $a,b,c \in \mathbb{Z}$, we have

$$[a]_n([b]_n + [c]_n) = [a]_n[b+c]_n = [a(b+c)]_n$$

by definition of the operations. On the other hand

$$[a]_n[b]_n + [a]_n[c]_n = [ab]_n + [ac]_n = [ab + ac]_n,$$

also by definition. The distributive law for $\mathbb{Z}$ states that $a(b + c) = ab + ac$, so it follows that

$$[a]_n([b]_n + [c]_n) = [a(b + c)]_n = [ab + ac]_n = [a]_n[b]_n + [a]_n[c]_n.$$

So $\mathbb{Z}_n$ is a ring.

Example 5.2.2 Another important example is that of a *polynomial ring*. For the moment, let's just consider polynomials with real coefficients, i.e., expressions of the form:

$$f(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0,$$

where m is a non-negative integer and $a_0, a_1, \ldots, a_m$ are real numbers. The set of such polynomials is denoted $\mathbb{R}[x]$. Let's recall the definition of addition and multiplication of polynomials. Suppose that $f(x)$ is as above and

$$g(x) = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0, \qquad \text{with } b_0, b_1, \ldots, b_n \in \mathbb{R}.$$

So using summation notation,

$$f(x) = \sum_{i=0}^{m} a_i x^i \quad \text{and} \quad g(x) = \sum_{j=0}^{n} b_j x^j.$$

To write down the formula for $f(x) + g(x)$, it's convenient to assume the leading terms have the same degree, which we can do by writing $f(x)$ as

$$0x^n + 0x^{n-1} + \cdots + 0x^{m+1} + a_m x^n m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0,$$

if $m < n$ (i.e., setting $a_{n+1} = \cdots = a_{m-1} = a_m = 0$), and similarly modifying the expression for $g(x)$ if $n < m$. Then we define

$$f(x)+g(x) = \sum_{i=0}^{n}(a_i+b_i)x^i = (a_n+b_n)x^n+(a_{n-1}+b_{n-1})x^{n-1}+(a_1+b_1)x+(a_0+b_0),$$

an element of $\mathbb{R}[x]$ since $a_i + b_i \in \mathbb{R}$ for $i = 0, \ldots, n$. To define the multiplication operation, we set

$$f(x)g(x) = \sum_{k=0}^{m+n} d_k x^k,$$

where d_k is the sum of terms of the form $a_i b_j$ for which $i + j = k$, so

$$d_k = \sum_{i=0}^{k} a_i b_{k-i}$$

(where we set $a_i = 0$ if $i > m$ and $b_j = 0$ if $j > n$). These are just formulas for the familiar algebraic operations on polynomials, written in a way that shows the output is again a polynomial. For example if $f(x) = x^3 + 2x + 1$ and $g(x) = x^2 + 1$, then

$$f(x) + g(x) = x^3 + x^2 + 2x + 2 \quad \text{and} \quad f(x)g(x) = x^5 + 3x^3 + x^2 + 2x + 1.$$

To see that $\mathbb{R}[x]$ with these operations is a ring, we first check that the operations are associative. So suppose that $f(x)$ and $g(x)$ are as above, and $h(x) = \sum_{k=0}^{\ell} c_k x^k$. Then

$$(f(x) + g(x)) + h(x) = \sum_{i=0}^{n} (a_i + b_i)x^i + \sum_{i=0}^{n} c_i x^i = \sum_{i=0}^{n} ((a_i + b_i) + c_i)x^i$$

(where we've assumed $m = n = \ell$ as above), and similarly

$$f(x) + (g(x) + h(x)) = \sum_{i=0}^{n} (a_i + (b_i + c_i))x^i,$$

These agree since $(a_i + b_i) + c_i = a_i + (b_i + c_i)$ for each i, so polynomial addition is associative. As for multiplication, we have

$$(f(x)g(x))h(x) = \left(\sum_{r=0}^{m+n} d_r x^r\right)\left(\sum_{k=0}^{\ell} c_k x^k\right) = \sum_{s=0}^{m+n+\ell} e_s x^s$$

where d_r is the sum of the $a_i b_j$ such that $i + j = r$, and e_s is the sum of the $d_r c_k$ such that $r + k = s$. Therefore e_s is the sum of the $(a_i b_j)c_k$ such that $(i + j) + k = s$. (Note that we have just used the distributive law for real numbers to rewrite each term in the sum for e_s as

$$d_r c_k = (a_0 b_r + a_i b_{r-1} + \cdots + a_r b_0)c_k = (a_0 b_r)c_k + (a_1 b_{r-1})c_k + \cdots + (a_r b_0)c_k.)$$

Similarly the coefficient of x^s in $f(x)(g(x)h(x))$ is the sum of the terms $a_i(b_j c_k)$ for which $i+(j+k) = s$. It follows that for $s = 0, 1, \ldots, m+n+\ell$, the coefficients of x^s are the same for the two polynomials, so $(f(x)g(x))h(x) = f(x)(g(x)h(x))$ and the multiplication operation is associative.

Note also that both operations are commutative and have identity elements. (The identity elements are simply the *constant* polynomials 0 and 1.) To conclude that $(\mathbb{R}[x], +)$ is an abelian group, we just have to note that each polynomial

$$f(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0$$

in $\mathbb{R}[x]$ has an additive inverse in $\mathbb{R}[x]$, namely

$$(-a_m)x^m + (-a_{m-1})x^{m-1} + \cdots + (-a_1)x + (-a_0).$$

So the only thing remaining to check in order to conclude that $(\mathbb{R}[x], +, \cdot)$ is a ring is the distributive law, which is left as an exercise.

Example 5.2.3 In all the preceding examples, the multiplication operation was commutative, but here is an example where it's not. Recall that on $M_2(\mathbb{R})$ we have the binary operations of matrix addition and matrix multiplication. We have already know that $(M_2(\mathbb{R}), +)$ is an abelian group, and that matrix multiplication is associative and has identity element $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Finally we need to check *both* distributive laws. Suppose then that

$$A = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}, \quad B = \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} c_1 & c_2 \\ c_3 & c_4 \end{pmatrix}$$

are elements of $M_2(\mathbb{R})$. Then

$$\begin{aligned}
A(B+C) &= \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}\begin{pmatrix} b_1+c_1 & b_2+c_2 \\ b_3+c_3 & b_4+c_4 \end{pmatrix} \\
&= \begin{pmatrix} a_1(b_1+c_1)+a_2(b_3+c_3) & a_1(b_2+c_2)+a_2(b_4+c_4) \\ a_3(b_1+c_1)+a_4(b_3+c_3) & a_3(b_2+c_2)+a_4(b_4+c_4) \end{pmatrix} \\
&= \begin{pmatrix} a_1b_1+a_1c_1+a_2b_3+a_2c_3 & a_1b_2+a_1c_2+a_2b_4+a_2c_4 \\ a_3b_1+a_3c_1+a_4b_3+a_4c_3 & a_3b_2+a_3c_2+a_4b_4+a_4c_4 \end{pmatrix} \\
&= \begin{pmatrix} a_1b_1+a_2b_3 & a_1b_2+a_2b_4 \\ a_3b_1+a_4b_3 & a_3b_2+a_4b_4 \end{pmatrix} + \begin{pmatrix} a_1c_1+a_2c_3 & a_1c_2+a_2c_4 \\ a_3c_1+a_4c_3 & a_3c_2+a_4c_4 \end{pmatrix} \\
&= AB + AC,
\end{aligned}$$

which proves one of the distributive laws. I'll leave you to check the proof of the other, which is very similar.

Now let's consider some *non-examples* of rings. Let E denote the set of *even* integers under the usual binary operations of addition and multiplication. Then E is a subgroup of $\mathbb{Z}$ under addition, hence is an abelian group.

The operation $\cdot$ is associative and satisfies the distributive law with respect to addition, but there is no identity element in E for multiplication, so E is *not* a ring.

Here is an example where only the distributive law fails. Consider the set $\mathbb{R}$ with the operations $+$ and $\bullet$, where $+$ is the usual addition operation and $\bullet$ is defined by

$$x \bullet y = x + y + xy = (x-1)(y-1) + 1, \qquad \text{for } x, y \in \mathbb{R}.$$

We already know that $(\mathbb{R}, +)$ is an abelian group. The operation $\bullet$ is associative since

$$\begin{aligned} (x \bullet y) \bullet z = (x + y + xy) \bullet z &= (x + y + xy) + z + (x + y + xy)z \\ &= x + y + z + xy + xz + yz + xyz, \end{aligned}$$

and

$$\begin{aligned} x \bullet (y \bullet z) = x \bullet (y + z + yz) &= x + (y + z + yz) + x(y + z + yz) \\ &= x + y + z + xy + xz + yz + xyz. \end{aligned}$$

The operation is also obviously commutative since $x \bullet y = y \bullet x$, and 0 is an identity element since $x \bullet 0 = 0 \bullet x = 0 + x + 0x = x$ for all $x \in \mathbb{R}$. However the distributive law fails since

$$x \bullet (y + z) = x + (y + z) + x(y + z) = x + y + z + xy + xz,$$

which coincides with

$$(x \bullet y) + (x \bullet z) = x + y + xy + x + z + xz$$

only if $x = 0$. So for example $1 \bullet (0+0) = 1 \bullet 0 = 1$, but $1 \bullet 0 + 1 \bullet 0 = 1 + 1 = 2$. Therefore $(\mathbb{R}, +, \bullet)$ is *not* a ring.

5.3 Basic properties of rings

We now establish a few basic properties which hold for all rings since they can be deduced from the axioms in the definition. When working with an abstract ring $(R, +, *)$, we will write 0_R for the additive identity element, 1_R for the multiplicative identity element, and $-x$ for the additive inverse of x.

Proposition 5.3.1 *Suppose that $(R, +, *)$ is a ring.*

1. $0_R * x = 0_R = x * 0_R$ *for all* $x \in R$.

2. $(-x) * y = -xy = x * (-y)$ *for all* $x, y \in R$.

Proof. 1) Since 0_R is the additive identity element, we know $0_R + 0_R = 0_R$. So by the distributive law,

$$0_R * x = (0_R + 0_R) * x = (0_R * x) + (0_R * x)$$

for all $x \in R$. Since $0_R + (0_R * x) = 0_R * x$, we get that

$$0_R + (0_R * x) = (0_R * x) + (0_R * x).$$

Now since $(R, +)$ is a group, we can apply the cancellation law (Prop. 4.4.4) to conclude that $0_R = 0_R * x$. Similarly,

$$0_R + (x * 0_R) = x * 0_R = x * (0_R + 0_R) = (x * 0_R) + (x * 0_R),$$

so $0_R = x * 0_R$ as well, completing the proof of 1).

2) For all $x, y \in R$, we have

$$\begin{array}{rcll} ((-x) * y) + (x * y) & = & ((-x) + x) * y & \text{(by distributivity)} \\ & = & 0_R * y & \text{(by definition of } -) \\ & = & 0_R & \text{(by Part 1).} \end{array}$$

Therefore $(-x) * y$ is the additive inverse of $x * y$, i.e., $(-x) * y = -xy$. Similarly

$$(x * (-y)) + (x * y) = x * ((-y) + y) = x * 0_R = 0_R$$

shows that $(-x) * y = -xy$. □

Recall that if G is a group, then for $g \in G$ and $n \in \mathbb{Z}$, we defined the element $g^n \in G$, called the n^{th} *power* of g in G. In particular, if $n > 0$, then

$$g^n = \underbrace{gg \cdots g}_{n \text{ times.}}$$

For groups where the binary operation is being denoted by $+$, we instead call it the n^{th} *multiple* of g and denote it $n \cdot g$.

In particular, if $(R, +, *)$ is a ring, then $(R, +)$ is a group, so we can speak of the multiples $n \cdot x$ of an element $x \in R$ for $n \in \mathbb{Z}$. We can also define its powers

$$x^n = \underbrace{x * x \cdots x}_{n \text{ times,}}$$

provided $n \geq 0$. (We can define $x^0 = 1$, but recall that defining x^n for $n < 0$ requires having multiplicative inverses.)

Example 5.3.2 Consider for example an element $[a]_n$ in the ring $\mathbb{Z}_n$. Then one finds that $m \cdot [a]_n = [ma]_n$ for $m \in \mathbb{Z}$, and $[a]_n^m = [a^m]_n$ for $m \in \mathbb{N}$.

Example 5.3.3 The n^{th} power of the element $f(x) = x$ in the polynomial ring $\mathbb{R}[x]$ is, of course, what we've been denoting as x^n all along.

We'll refer to the additive version of Prop. 4.5.1 as the *law of multiples*:

$$(m \cdot x) + (n \cdot x) = (m + n) \cdot x \quad \text{and} \quad m \cdot (n \cdot x) = (mn) \cdot x$$

for $m, n \in \mathbb{Z}$, $x \in R$. Moreover since $(R, +)$ is abelian, we have $(n \cdot x) + (n \cdot y) = n \cdot (x + y)$ for $n \in \mathbb{Z}$, $x, y \in R$. The proof of Prop. 4.5.1 applies also apply to the powers of x, provided $m, n \geq 0$, giving

$$x^m x^n = x^{m+n} \quad \text{and} \quad (x^m)^n = x^{mn}$$

(but note that we can't conclude anything about $x^n y^n$ without assuming that $\cdot$ is commutative).

We also have the following behavior of the multiples of elements with respect to multiplication in the ring:

Proposition 5.3.4 *If R is a ring, then*

$$(n \cdot x)y = n \cdot (xy) = x(n \cdot y)$$

for all $x, y \in R$ and $n \in \mathbb{Z}$.

The proof is left as an exercise, but let's look carefully at the meaning of the equation. Note that $(n \cdot x)y$ is the product of two elements of the ring R, namely $n \cdot x$ and y; the first of these elements $n \cdot x$ is defined as the n^{th} multiple of the element x. On the other hand, $n \cdot (xy)$ is the n^{th} multiple of the element xy, where xy is the product in R of the elements x and y. We can't simply apply associativity since $\cdot$ is not the multiplication operation in R, and indeed the integer n might not even be an element of R. Consider the case $n = 2$, which already gives the key idea for how to prove the proposition in general. The first equality is saying that $(x + x)y = (xy) + (xy)$, which is of course a consequence of the *distributive* law.

It is often useful to consider the multiples of the multiplicative identity element 1_R, and there is the following notation for them: If $m \in \mathbb{Z}$, then we let $m_R = m \cdot (1_R)$ denote the m^{th} multiple of 1_R. (Note that for $m = 1$, this gives $1_R = 1_R$, and for $m = 0$, we have that $0 \cdot (1_R)$ is *by definition* the additive identity element 0_R, so there is no conflict in notation.) So for example

- for $R = \mathbb{Z}$, we have $m_R = m$;
- for $R = \mathbb{Z}_n$, we have $m_R = m \cdot [1]_n = [m]_n$;
- for $R = M_2(\mathbb{R})$, we have $m_R = \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix}$.

Proposition 5.3.5 *Suppose that R is a ring. Then*

1. *$m_R x = m \cdot x = x m_R$ for all $m \in \mathbb{Z}$, $x \in R$;*
2. *$m_R + n_R = (m+n)_R$ and $m_R n_R = (mn)_R$ for all $m, n \in \mathbb{Z}$;*
3. *$(m_R)^n = (m^n)_R$ for all $m \in \mathbb{Z}$, $n \in \mathbb{N}$.*

Proof. 1) This is a corollary of Prop. 5.3.4; the proof is left as an exercise.

2) Since $m_R = m \cdot (1_R)$ and $n_R = n \cdot (1_R)$, the law of multiples shows that

$$m_R + n_R = m \cdot (1_R) + n \cdot (1_R) = (m+n) \cdot (1_R) = (m+n)_R.$$

According to Part 1, applied with $x = n_R$, we have

$$m_R \cdot n_R = m \cdot (n_R) = m \cdot (n \cdot (1_R)) = (mn) \cdot (1_R) = (mn)_R$$

by the law of multiples (and the definitions of n_R and $(mn)_R$).

3) We prove this by induction on n. For $n = 1$ there is nothing to prove. If $n \geq 1$ and $(m_R)^n = (m^n)_R$, it follows that

$$(m_R)^{n+1} = m_R \cdot (m_R)^n = m_R \cdot (m^n)_R = (m^{n+1})_R,$$

where the last equality is by Part 2) with m^n in place of n. □

5.4 Subrings

Recall that a *subgroup* of a group G is a subset of G which itself is a group with the same binary operation as on G. There's a similar notion of a *subring* of a ring.

Definition 5.4.1 Suppose that $(R, +, *)$ is a ring and S is a subset of R. Then S is a **subring** of R if S, with the operations $+$ and $*$, is a ring and $1_S = 1_R$.

Note the extra technical condition that $1_S = 1_R$ in the definition. We didn't need this in the definition of a subgroup since it was automatically the case that the identity elements had to agree (see the discussion before Prop. 4.7.2). For this reason we don't need to explicitly require that $0_S = 0_R$ (as it follows from S being a subgroup of R under $+$).

Example 5.4.2 We have the inclusions

$$\mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C},$$

where each subset is a subring of all the larger rings.

Note that the set of positive integers $\mathbb{N}$ is not a subring of these since $\mathbb{N}$ is not a group under $+$ (it lacks an identity element and inverses). The set of even integers is not a subring of these since it lacks a multiplicative identity element.

Another *non-example* to note is the subset $\{0\}$ of any of the above rings. Even though $\{0\}$ is a ring under $+$ and multiplication, it is not a *subring* of $\mathbb{Z}$ since its multiplicative identity element is 0, which is different from the multiplicative identity element 1 of the ring $\mathbb{Z}$. The extra condition that $1_R = 1_S$ rules this out from being considered a subring.

Example 5.4.3 Identifying the real number $a \in \mathbb{R}$ with the *constant* polynomial $f(x) = a$, we can view $\mathbb{R}$ as a subring of the ring $\mathbb{R}[x]$ of Example 5.2.2.

Proposition 5.4.4 *Suppose that $(R, +, *)$ is a ring and $S \subseteq R$. Then S is a subring of R if and only if all of the following hold:*

1. *$0_R \in S$ and $1_R \in S$;*
2. *if $x, y \in$, then $x + y \in S$ and $x * y \in S$;*
3. *if $x \in S$, then $-x \in S$.*

Proof. Suppose first that S is a subring of R. Then S is a group under $+$, so it is a subgroup of $(R, +)$. Therefore by Prop. 4.7.2, we have

- if $x, y \in S$, then $x + y \in S$;
- $0_R \in S$;
- if $x \in S$, then $-x \in S$.

The fact that $1_R \in S$ follows from the definition of a subring, which requires that $1_R = 1_S$. Finally if $x, y \in S$, then since $*$ is a binary operation on S, we must have $x * y \in S$. We have now shown that 1), 2) and 3) are all satisfied.

Now suppose that 1), 2) and 3) are satisfied by S, and we will show that S is a subring. From 2), we know that $+$ and $*$ define binary operations on S. Since $0_R \in S$ (by 1) and $-x \in S$ whenever $x \in S$ (by 3), we know that S is a subgroup of $(R, +)$, and so $(S, +)$ is an abelian group. Since R is a ring, the operation $\cdot$ is associative on R, hence associative on S. Since $1_R \in S$, we have that 1_R is the identity element for $*$ on S. Finally since the distributive laws hold for the operations $+$ and $*$ on R, they must hold for the operations on S as well. Therefore S is a ring under the operations $+$ and $*$. Since its multiplicative identity is 1_R, we have that $1_S = 1_R$, so S is a subring of R. □

Example 5.4.5 Consider the subset of diagonal matrices

$$S = \left\{ \left(\begin{array}{cc} a & 0 \\ 0 & d \end{array} \right) \middle| \, a, d \in \mathbb{R} \right\}.$$

in the ring $R = M_2(\mathbb{R})$ of Example 5.2.3: We verify the conditions 1), 2) and 3) of Prop. 5.4.4.

1) The matrices $0_R = \left(\begin{array}{cc} 0 & 0 \\ 0 & 0 \end{array} \right)$ and $1_R = \left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right)$ are both in S.

2) If $A = \left(\begin{array}{cc} a & 0 \\ 0 & d \end{array} \right)$ and $A' = \left(\begin{array}{cc} a' & 0 \\ 0 & d' \end{array} \right)$ are in S, then so is

$$A + A' = \left(\begin{array}{cc} a + a' & 0 \\ 0 & d + d' \end{array} \right).$$

3) If $A = \left(\begin{array}{cc} a & 0 \\ 0 & d \end{array} \right)$ and $A' = \left(\begin{array}{cc} a' & 0 \\ 0 & d' \end{array} \right)$ are in S, then so is

$$A + A' = \left(\begin{array}{cc} aa' & 0 \\ 0 & dd' \end{array} \right).$$

Let's take note though of a *non-example*. The subset

$$T = \left\{ \left(\begin{array}{cc} a & 0 \\ 0 & 0 \end{array} \right) \middle| \, a \in \mathbb{R} \right\}$$

of R satisfies conditions 2)and 3), but not 1) since $\left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right)$ is not in S.

5.5 Groups of units

Suppose that $(R, +, *)$ is a ring. It is part of the definition of a ring that $(R, +)$ needs to be a group, but we don't require that $(R, *)$ be a group. We require only that $*$ be associative and have an identity element in R, but not that every element have an inverse. In fact, since $0_R * x = 0_R$ for all $x \in R$, the additive identity 0_R will never have a multiplicative inverse unless $0_R = 1_R$, in which case $x = x * 1_R = x * 0_R = 0_R$ for *all* $x \in R$, so $R = \{0_R\}$ has only one element. So except in that very special case, $(R, *)$ will *not* be a group. However, we can still associate a multiplicative group to the ring, called the group of units of R.

Definition 5.5.1 Suppose that $(R, +, *)$ is a ring. An element $x \in R$ is called a **unit** in R if x has a multiplicative inverse in R, i.e.,if there exists an element $y \in R$ such that

$$x * y = 1_R = y * x.$$

Example 5.5.2 Let $R = \mathbb{Z}_n$ as in Example 5.2.1, and suppose that $[a]_n \in \mathbb{Z}_n$. Then $[a]_n$ is a unit in $\mathbb{Z}_n$ if there is a residue class $[b]_n \in \mathbb{Z}_n$ such that $[a]_n[b]_n = [1]_n$. Since $[a]_n[b]_n = [ab]_n$, we see that the condition for $[a]_n$ to be a unit in $\mathbb{Z}_n$ is that

$$ab \equiv 1 \pmod{n} \quad \text{for some } b \in \mathbb{Z}.$$

So by Prop. 4.2.11, $[a]_n$ is a unit in $\mathbb{Z}_n$ if and only if $\gcd(a, n) = 1$, i.e., if and only

$$[a]_n \in \mathbb{Z}_n^\times = \{\, [a]_n \in \mathbb{Z}_n \mid \gcd(a, n) = 1 \,\}.$$

More generally, for any ring R we let $R^\times$ denote the set of units in R. So for example, if $R = \mathbb{R}$, then the set of units is $\mathbb{R}^\times = \{\text{non-zero real numbers}\}$. In fact, whenever we used the superscript $^\times$ before, it was consistent with the definition just given.

We have seen already that $\mathbb{Z}_n^\times$ and $\mathbb{R}^\times$ were groups under multiplication (Prop. 4.2.13 for $\mathbb{Z}_n^\times$). These are special cases of the more general fact:

Theorem 5.5.3 *Suppose that $(R, +, *)$ is a ring. Then $(R^\times, *)$ is a group.*

Proof. We must first show that $*$ defines a binary operation on $R^\times$, i.e., if x and x' are elements of $R^\times$, then so is $x * x'$. So we must show that if x and x' have multiplicative inverses in R, then so does $x \cdot x'$. Let y be the multiplicative inverse of x and y' that of x'. So $x * y = y * x = 1_R$ and $x' * y' = y' * x' = 1_R$. Then $y' * y$ is the multiplicative inverse of $x * x'$ since

$$(x * x') * (y' * y) = x * (x' * y') * y = x * 1_R * y = x * y = 1_R,$$

and similarly $(y' * y) * (x * x') = 1_R$.

We must now show that R with the binary operation $*$ satisfies the definition of a group. First of all, since $*$ is associative on R, it is also associative on the subset $R^\times \subseteq R$. To see that there is an identity element for $*$ on $R^\times$, note that $1_R \in R^\times$ since 1_R has a multiplicative inverse in R, namely 1_R. Finally we have to show that if $x \in R^\times$, then x has an inverse element in $R^\times$ for $*$. By the definition of $R^\times$, we know that x has an inverse element $y \in R$. We need only show that $y \in R^\times$, i.e., that y has a multiplicative in R, and indeed it does, namely x. We have now shown that $(R^\times, *)$ is a group. □

Example 5.5.4 Let $R = M_2(\mathbb{R})$ as in Example 5.2.3. Recall that a matrix A has a multiplicative inverse in $M_2(\mathbb{R})$ if and only if $\det A \neq 0$. Indeed if $AB = I$ for some $B \in M_2(\mathbb{R})$, then

$$(\det A)(\det B) = \det(AB) = \det I = 1,$$

so $\det A \neq 0$. Conversely if $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $ad - bc \neq 0$, then A has

$$A^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

as its multiplicative inverse. Therefore $(M_2(\mathbb{R}))^\times = \mathrm{GL}_2(\mathbb{R})$.

Finally we remark that if $x \in R^\times$, then x^m is defined as an element of the group $R^\times$ for *all* integers m, not just for positive integers. So for example if $R = \mathbb{Z}_n$ for some positive integer n, and $a \in \mathbb{Z}$ is such that $\gcd(a, n) = 1$, then $[a]_n^m$ is defined for all $m \in \mathbb{Z}$, and in particular $[a]_n^{-1}$ is defined. Note though that $[a^{-1}]_n$ is not usually defined since a^{-1} is not an integer unless $a = \pm 1$. So for example, $[3]_7^{-1} = [5]_7$.

5.6 Types of rings

Recall that a *ring* is a set R with associative binary operations $+$ and $*$ such that:

- $(R, +)$ is an abelian group,
- $*$ has an identity element
- and the distributive laws hold.

We saw several examples, the most basic being $\mathbb{Z}$. Some others were $\mathbb{R}$, $\mathbb{Z}_n$, $\mathbb{R}[x]$ and $M_2(\mathbb{R})$, each with their addition and multiplication operations. We defined the notions of *subring* and *unit group* of a ring, and *multiples* and (non-negative) *powers* of elements.

Here is another example of a ring:

Example 5.6.1 Let $R = \mathbb{Z} \times \mathbb{Z} = \{ (m, n) \mid m, n \in \mathbb{Z} \}$ with addition and multiplication defined *componentwise*, so for $(m, n), (m', n') \in R$,

$$(m, n) + (m', n') = (m + m', n + n') \quad \text{and} \quad (m, n)(m', n') = (mm', nn').$$

This is an example of a *product ring*. More generally if R and S are rings, we can make $R \times S$ a ring by defining the binary operations on the product componentwise. This is left as an exercise.

We now define some special types of rings. Recall that in the ring axioms we require the addition operation to be commutative, but the multiplication operation need not be.

Definition 5.6.2 We say that a ring $(R, +, *)$ is **commutative** if the operation $*$ on R is commutative, i.e.,

$$x * y = y * x \quad \text{for all } x, y \in R.$$

Most of the examples we've considered have been commutative. In particular, the rings $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Z}_n$, $\mathbb{Z} \times \mathbb{Z}$ and $\mathbb{R}[x]$ are commutative. On the other hand, the matrix ring $M_2(\mathbb{R})$ is *not* commutative.

Definition 5.6.3 We say that a ring R is an **integral domain** (or simply a **domain**) if R is commutative, $0_R \neq 1_R$ and

$$x, y \in R,\ xy = 0_R \quad \Rightarrow \quad x = 0_R \text{ or } y = 0_R.$$

For example, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{R}[x]$ are integral domains. To prove that $\mathbb{R}[x]$ is a domain, suppose that

$$f(x) = a_m x^m + a_{m-1} x^{m-1} + \cdots + a_1 x + a_0,\ g(x) = b_n x^n + b_{n-1} x^{n-1} + \cdots + b_1 x + b_0$$

are non-zero elements of $\mathbb{R}[x]$. Since some coefficient a_i is non-zero, we can assume that $a_m \neq 0$ by removing any leading terms whose coefficient is 0. Similarly we can assume $b_n \neq 0$. Then the coefficient of x^{m+n} in $f(x)g(x)$ is $a_m b_n$, which is non-zero.

The ring $\mathbb{Z} \times \mathbb{Z}$ (Example 5.6.1) is an example of a ring which is *not* an integral domain since $(1, 0)(0, 1) = (0, 0)$.

As for $\mathbb{Z}_n$, we have:

Proposition 5.6.4 *Let n be a positive integer. Then $\mathbb{Z}_n$ is a domain if and only if n is prime.*

Proof. For $n = 1$, we have $[0]_n = [1]_n$, so $\mathbb{Z}_n$ is not an integral domain. Nor is 1 prime, so the proposition holds in this case.

If $n > 1$ and n is composite, then $n = ab$ for some integers a, b with $1 < a < n$ and $1 < b < n$. So $[a]_n \neq [0]_n$ and $[b]_n \neq [0]_n$, but $[a]_n[b]_n = [ab]_n = [n]_n = [0]_n$, so $\mathbb{Z}_n$ is not an integral domain, and the proposition holds in this case as well.

Finally suppose $n = p$ is prime. If $[a]_p, [b]_p \in \mathbb{Z}_p$ and $[a]_p[b]_p = [0]_p$, then $[ab]_p = [0]_p$, so $p|ab$. By Prop. 2.5.2, $p|a$ or $p|b$. Therefore either $[a]_p = [0]_p$ or $[b]_p = [0]_p$. Note also that $[1]_p \neq [0]_p$, so $\mathbb{Z}_p$ is an integral domain, and the proposition holds in all cases. □

The nice thing about integral domains is that a cancellation law holds:

Proposition 5.6.5 *Suppose that R is an integral domain, $x, y, z \in R$ and $x \neq 0_R$. If $xy = xz$, then $y = z$.*

Proof. If $xy = xz$, then

$$0_R = xz - xz = xy + x(-z) = x(y - z).$$

Since R is an integral domain and $x \neq 0_R$, it follows that $y - z = 0_R$. Adding z to both sides of the equation then gives $y = z$. □

A type of ring even nicer than an integral domain is called a *field*.

Definition 5.6.6 A ring R is called a **field** if R is an integral domain and every non-zero element of R is a unit (where *non-zero* means different from 0_R).

Proposition 5.6.7 *Suppose that R is a commutative ring. Then R is a field if and only if*

$$R^\times = R \setminus \{0_R\} = \{\, x \in R \,|\, x \neq 0_R \,\}.$$

Proof. Suppose R is a field (as in Defn. 5.6.6). Then R is an integral domain, so $0_R \neq 1_R$. Therefore $0_R r = 0_R$ never equals 1_R, so 0_R cannot be a unit. On the other hand every non-zero element of R is a unit, so $R^\times = R \setminus \{0_R\}$.

Conversely suppose that $R^\times = R \setminus \{0_R\}$. Then every non-zero element of R is a unit, so we just have to show that R is an integral domain in order to conclude that R is a field. Since $0_R \notin R^\times$ and $1 \in R^\times$, we know that $0_R \neq 1_R$. So suppose that $xy = 0_R$, but $x \neq 0_R$. Then $x \in R^\times$, so x has a multiplicative inverse $x^{-1} \in R$, and

$$y = 1_R y = x^{-1}xy = x^{-1}0_R = 0_R.$$

Recall we assumed R was commutative. Therefore R satsifies all the criteria in the Def. 5.6.3. □

The fields we've met so far are $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Z}_p$ for p prime.

5.7 Matrix rings

Let $M_2(\mathbb{Z})$ denote the set of matrices in $M_2(\mathbb{R})$ with integer entries, so

$$M_2(\mathbb{Z}) = \left\{ \left(\begin{array}{cc} a & b \\ c & d \end{array} \right) \middle| \, a, b, c, d \in \mathbb{Z} \right\}.$$

Then it's easy to see that $M_2(\mathbb{Z})$ satisfies the conditions of Prop. 5.4.4, so is therefore a subring of $M_2(\mathbb{R})$. In fact if R is *any* ring and $n \geq 1$, we let $M_n(R)$ denote the set of $n \times n$-matrices with entries in R. We can then define binary operations on $M_n(R)$ using the usual formulas for matrix addition and multiplication. Using the subscript ij to denote the entry in the i^{th} row and j^{th} column, this means that if

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix},$$

then

$$A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} + b_{n1} & a_{n2} + b_{n2} & \cdots & a_{nn} + b_{nn} \end{pmatrix}$$

and AB is the matrix

$$\begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + \cdots + a_{1n}b_{n1} & \cdots & a_{11}b_{1n} + a_{12}b_{2n} + \cdots + a_{1n}b_{nn} \\ a_{21}b_{11} + a_{22}b_{21} + \cdots + a_{2n}b_{n1} & \cdots & a_{21}b_{1n} + a_{22}b_{2n} + \cdots + a_{2n}b_{nn} \\ \vdots & & \vdots \\ a_{n1}b_{11} + a_{n2}b_{21} + \cdots + a_{nn}b_{n1} & \cdots & a_{n1}b_{1n} + a_{n2}b_{2n} + \cdots + a_{nn}b_{nn} \end{pmatrix}.$$

In other words, the ij entry of $A + B$ is $a_{ij} + b_{ij}$, and the ij-entry of AB is

$$\sum_{k=1}^{n} a_{ik}b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}.$$

(Note that since addition on R being associative, we don't need parentheses to specify the order in which to perform the addition operations.) We now check that these operations make $M_n(R)$ a ring, called the **ring** of $n \times n$-**matrices over** R.

Proposition 5.7.1 *If R is a ring, then $M_n(R)$ is a ring under matrix addition and multiplication.*

Proof. We first check that $M_n(R)$ is an abelian group under $+$. To see that matrix addition is associative, suppose $A, B, C \in M_n(R)$ with ij-entries are a_{ij}, b_{ij} and c_{ij}. Then the ij-entry of $(A + B) + C$ is $(a_{ij} + b_{ij}) + c_{ij}$, which is the same as $a_{ij} + (b_{ij} + c_{ij})$ since the operation $+$ is associative on the original ring R. Therefore the ij-entries of $A + (B + C)$ and $(A + B) + C$ are the same for all i, j (with $1 \leq i \leq n$, $1 \leq j \leq n$). Similarly we see that since $+$ is commutative on R, so is matrix addition on $M_n(R)$. There is an identity element, namely the matrix $\mathbf{0}$ all of whose entries are 0_R. Finally the additive inverse of A is $-A$, whose ij-entry is $-a_{ij}$ (the negative of a_{ij} in the ring R). Therefore $M_n(R)$ is an abelian group under matrix addition.

Next we check that matrix multiplication on $M_n(R)$ is associative and has an identity element. We compute the ij-entry of $(AB)C$ and compare to that of $A(BC)$. Using sumation notation, we have that the ij-entry of AB is $r_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$. Therefore the ij-entry of $(AB)C$ is

$$\sum_{\ell=1}^{n} r_{i\ell}c_{\ell j} = \sum_{\ell=1}^{n} \left(\sum_{k=1}^{n} a_{ik}b_{k\ell} \right) c_{\ell j} = \sum_{\ell=1}^{n} \left(\sum_{k=1}^{n} (a_{ik}b_{k\ell})c_{\ell j} \right)$$

by one of the *distributive* laws on the original ring R. Similarly, letting s_{ij} denote the ij-entry of BC, we see that the ij-entry of $A(BC)$ is

$$\sum_{k=1}^{n} a_{ik}e_{kj} = \sum_{k=1}^{n} a_{ik} \left(\sum_{\ell=1}^{n} b_{k\ell}c_{\ell j} \right) = \sum_{k=1}^{n} \left(\sum_{\ell=1}^{n} a_{ik}(b_{k\ell}c_{\ell j}) \right)$$

by the *other* distributive law on R. Since *addition* on R is *commutative*, the double sums don't depend on the order in which we arrange their n^2 terms (with k and ℓ each running from 1 to n). Since multiplication on R is *associative*, we have $(a_{ik}b_{k\ell})c_{\ell j} = a_{ik}(b_{k\ell}c_{\ell j})$, so we have exactly the same terms in the two double sums. This shows that the ij-entries of $(AB)C$ and $A(BC)$ coincide for all i, j, so $(AB)C = A(BC)$. The identity element is the matrix $\mathbf{1}$ whose entries are 1_R along the diagonal and 0_R otherwise. To check

this, denote the entries of $\mathbf{1}$ by δ_{ij}, so δ_{ij} is 1_R or 0_R according to whether or not $i = j$. Then the ij-entry of $\mathbf{1}A$ is

$$\sum_{k=1}^{n} \delta_{ij} a_{kj} = \delta_{i1} a_{1j} + \delta_{i2} a_{2j} + \cdots + \delta_{in} a_{nj}.$$

For $k \neq i$, we have $\delta_{ik} a_{kj} = 0_R a_{kj} = 0_R$, so the only term in the sum other than 0_R is $\delta_{ii} a_{ij} = 1_R a_{ij} = a_{ij}$. Therefore $\mathbf{1}A = A$. Similarly we see that $A\mathbf{1} = A$, so $\mathbf{1}$ is an identity element.

Finally we have to check the distributive laws. We compare the ij-entries of $A(B + C)$ and $AB + AC$. Let t_{ij} denote the ij-entry of $B + C$. Then the ij entry of $A(B + C)$ is

$$\begin{aligned}
&a_{i1}t_{1j} + a_{i2}t_{2j} + \cdots + a_{in}t_{nj} \\
&= a_{i1}(b_{1j} + c_{1j}) + a_{i2}(b_{2j} + c_{2j}) + \cdots + a_{in}(b_{nj} + c_{nj}) \\
&= a_{i1}b_{1j} + a_{i1}c_{1j} + a_{i2}b_{2j} + a_{i2}c_{2j}) + \cdots + a_{in}b_{nj} + a_{in}c_{nj} \\
&= (a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}) + (a_{i1}c_{1j} + a_{i2}c_{2j} + \cdots + a_{in}c_{nj}),
\end{aligned}$$

where the first equality is from the definition of matrix addition ($t_{kj} = b_{kj} + c_{kj}$), the second by a distributive law on R, and the third by commutativity of $+$ on R. Since $a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}$ is the ij-entry of AB and $a_{i1}c_{1j} + a_{i2}c_{2j} + \cdots + a_{in}c_{nj}$ is the ij-entry of AC, we conclude that $A(B+C) = AB + AC$. Similarly we find that $(A + B)C = AC + BC$ for all $A, B, C \in R$. □

Consider for example $M_2(\mathbb{Z}_3)$. This is a ring with $81 = 3^4$ elements. Let $A = \begin{pmatrix} [2] & [1] \\ [1] & [0] \end{pmatrix}$ and $B = \begin{pmatrix} [1] & [0] \\ [1] & [1] \end{pmatrix}$. Let's compute $A + B$, AB and BA:

$$\begin{aligned}
A + B &= \begin{pmatrix} [2] & [1] \\ [1] & [0] \end{pmatrix} + \begin{pmatrix} [1] & [0] \\ [1] & [1] \end{pmatrix} \\
&= \begin{pmatrix} [2]+[1] & [1]+[0] \\ [1]+[1] & [0]+[1] \end{pmatrix} \\
&= \begin{pmatrix} [0] & [1] \\ [2] & [1] \end{pmatrix},
\end{aligned}$$

$$\begin{aligned}
AB &= \begin{pmatrix} [2] & [1] \\ [1] & [0] \end{pmatrix} \begin{pmatrix} [1] & [0] \\ [1] & [1] \end{pmatrix} \\
&= \begin{pmatrix} [2][1]+[1][1] & [2][0]+[1][1] \\ [1][1]+[0][1] & [1][0]+[0][1] \end{pmatrix}, \\
&= \begin{pmatrix} [0] & [1] \\ [1] & [0] \end{pmatrix},
\end{aligned}$$

$$\begin{aligned} BA &= \begin{pmatrix} [1] & [0] \\ [1] & [1] \end{pmatrix}\begin{pmatrix} [2] & [1] \\ [1] & [0] \end{pmatrix} \\ &= \begin{pmatrix} [1][2]+[0][1] & [1][1]+[0][0] \\ [1][2]+[1][1] & [1][1]+[1][0] \end{pmatrix} \\ &= \begin{pmatrix} [2] & [1] \\ [0] & [1] \end{pmatrix}. \end{aligned}$$

It is an exercise to check that if R is a commutative ring, then the unit group of $M_2(R)$ is

$$\mathrm{GL}_2(R) = \{\, A \in M_2(R) \mid \det(A) \in R^\times \,\}.$$

This gives a way of constructing some interesting finite groups. For example, the above matrices A and B are elements of the group $\mathrm{GL}_2(\mathbb{Z}_3)$, which has order 48.

5.8 Ring homomorphisms

Recall that a *homomorphism of groups* is a function from one group to another that is compatible with their binary operations (see Definition 4.12.1). There is an analogous notion of a *homomorphism of rings*.

Definition 5.8.1 Suppose that $(R, +_R, *_R)$ and $(S, +_S, *_S)$ are rings. A function $\phi : R \to S$ is a **homomorphism** (of **rings**) if all of the following hold:

1. $\phi(x +_R y) = \phi(x) +_S \phi(y)$ for all $x, y \in R$;
2. $\phi(x *_R y) = \phi(x) *_S \phi(y)$ for all $x, y \in R$;
3. $\phi(1_R) = 1_S$.

Condition 1) just says that ϕ is a group homomorphism from the abelian group R (under $+_R$) to the abelian group S (under $+_S$). Condition 2) is the analogous one for the multiplication operations. Recall (Prop. **??**) that the condition in the definition of a group homomorphism guarantees that one identity element is sent to the other, so in particular 1) implies that $\phi(0_R) = 0_S$. On the other hand R and S are not groups under their multiplication operations, so condition 2) might not imply that $\phi(1_R) = 1_S$. We require it explicitly by imposing condition 3). The subscripts are included above to emphasize which binary operations are being applied, but they will not usually appear in practice (nor will the $*$).

Example 5.8.2 Define $\phi : \mathbb{Z} \to \mathbb{Z}_n$ by $\phi(a) = [a]_n$. We check that this function satisfies the three conditions in the definition of a ring homomorphism. If a and b are any integers, then

$$\phi(a+b) = [a+b]_n = [a]_n + [b]_n = \phi(a) + \phi(b),$$

where the middle equality is precisely the *definition* of the + operation on $\mathbb{Z}_n$. Similarly we see that

$$\phi(ab) = [ab]_n = [a]_n[b]_n = \phi(a)\phi(b).$$

Finally $\phi(1) = [1]_n$, which is the multiplicative identity element of $\mathbb{Z}_n$. Therefore ϕ is a ring homomorphism.

Example 5.8.3 If R is any ring, then we can define a homomorphism $\phi : \mathbb{Z} \to R$ by $\phi(n) = n_R = n \cdot 1_R$ (the n^{th} multiple of 1_R in R). Then

$$\phi(m+n) = (m+n) \cdot 1_R = m \cdot 1_R + n \cdot 1_R = \phi(m) + \phi(n),$$

where the middle equality is by the *laws of multiples* (i.e., the additive version of Prop. 4.5.1). We also have

$$\phi(mn) = (mn) \cdot 1_R = m \cdot (n \cdot 1_R),$$

again by the laws of multiples. Applying Prop. 5.3.4 with m in place of n, 1_R in place of x and $n_R = n \cdot 1_R$ in place of y gives

$$m \cdot (n \cdot 1_R) = (m \cdot 1_R)(n \cdot 1_R) = \phi(m)\phi(n).$$

Finally $\phi(1) = 1_R$ by definition. So ϕ is a homomorphism. Example 5.8.2 is just the special case of this example with $R = \mathbb{Z}_n$.

Example 5.8.4 If R is a subring of S, then the inclusion function $i : R \to S$ defined by $i(r) = r$ is a homomorphism.

Example 5.8.5 If m and n are positive integers and $n|m$, then we can define a function $\phi : \mathbb{Z}_m \to \mathbb{Z}_n$ by $\phi([a]_m) = [a]_n$. Note that we have to check that this function is well-defined. (Recall what this means. An element of $\mathbb{Z}_m$ is a residue class, say X. There are infinitely many ways to write X in the form $[a]_m$ where $a \in \mathbb{Z}$; indeed $X = [a]_m$ for any $a \in X$. We have to check that our formula for $\phi(X)$ doesn't depend on which a we choose.) So suppose $[a]_m = [a']_m$. Then $a \equiv a' \bmod m$; i.e., $m|(a - a')$. We are assuming $n|m$, which then implies that $n|(a-a')$, so $[a]_n = [a']_n$. Therefore ϕ is well-defined (but note we had to assume $n|m$). Moreover ϕ is a homomorphism since:

$$\phi([a]_m + [b]_m) = \phi([a+b]_m) = [a+b]_n = [a]_n + [b]_n = \phi([a]_m) + \phi([b]_m)$$

for all $[a]_m, [b]_m \in \mathbb{Z}_m$ Similarly we find that $\phi([a]_m[b]_m) = \phi([a]_m)\phi([b]_m)$. And $\phi([1]_m) = [1]_n$ by definition, so ϕ is a homomorphism.

Example 5.8.6 Recall that $\mathbb{R}[x]$ denotes the ring of polynomials in the variable x with coefficients in $\mathbb{R}$ (Example 5.2.2). If α is any real number, then we can define a function

$$v_\alpha : \mathbb{R}[x] \to \mathbb{R}$$

by $v_\alpha(f(x)) = f(\alpha)$ which we could call *evaluation-at-α*. Then v_α is a homomorphism since

$$v_\alpha(f(x) + g(x)) = f(\alpha) + g(\alpha) = v_\alpha(f(x)) + v_\alpha(g(x)).$$

Similarly $v_\alpha(f(x)g(x)) = f(\alpha)g(\alpha) = v_\alpha(f(x))v_\alpha(g(x))$. Finally $v_\alpha(1) = 1$, so v_α is a homomorphism.

Example 5.8.7 Consider the function $\phi : \mathbb{R} \to \mathbb{R} \times \mathbb{R}$ (the product ring) defined by $\phi(x) = (x, 0)$. Then ϕ satisfies conditions a) and b), but not c), so ϕ is *not* a homomorphism.

Now let's note a few general properties of homomorphisms.

Proposition 5.8.8 *Suppose that $\phi : R \to S$ is a homomorphism of rings. Then $\phi(n_R) = n_S$ for all $n \in \mathbb{Z}$.*

Proof. Recall that $n_R = n \cdot 1$. According to Defn. 5.8.1, ϕ is a homomorphism of groups (under $+$) and $\phi(1_R) = 1_S$, so by Prop. 4.12.10 (with additive notation), we have

$$\phi(n_R) = \phi(n \cdot 1_R) = n \cdot \phi(1_R) = n \cdot 1_S = n_S$$

for all $n \in \mathbb{Z}$. □

The proof of the following proposition is left as an exercise:

Proposition 5.8.9 *If $\phi : R \to S$ and $\psi : S \to T$ are homomorphisms of rings, then so is $\psi \circ \phi : R \to T$.*

Recall that if R is a ring, then the *unit group* of R is the group

$$R^\times = \{\, r \in R \,|\, rr' = r'r = 1_R \text{ for some } r' \in R \,\}$$

under the multiplication operation of the ring (See Def. 5.5.1 and Thm. 5.5.3).

Proposition 5.8.10 *Suppose that $\phi : R \to S$ is a homomorphism of rings. Then the restriction of ϕ defines a homomorphism of groups from $R^\times$ to $S^\times$. (In particular, if $r \in R^\times$, then $\phi(r) \in S^\times$.)*

Proof. If $r \in R^\times$, then there is an element $r' \in R$ such that $rr' = r'r = 1_R$. Since ϕ is a homomorphism, this implies that

$$\phi(r)\phi(r') = \phi(rr') = \phi(1_R) = 1_S.$$

Similarly $\phi(r')\phi(r) = 1_S$. Therefore $\phi(r)$ has an inverse in S, namely $\phi(r')$, so $\phi(r) \in R^\times$.

Restricting the domain of ϕ to $R^\times$ therefore defines a function

$$\phi^\times : R^\times \to S^\times$$

(i.e., $\phi^\times$ is defined by $\phi^\times(r) = \phi(r)$ for $r \in R^\times$). Since ϕ is a homomorphism of rings, $\phi(rr') = \phi(r)\phi(r')$ for all $r, r' \in R$. Therefore $\phi^\times(rr') = \phi^\times(r)\phi^\times(r')$ for all $r, r' \in R^\times$. □

Recall that if a homomorphism of groups is bijective, it is called an isomorphism (Definition 4.12.12). There's a similar notion for rings:

Definition 5.8.11 If R and S are rings, then a function $\phi : R \to S$ is called an **isomorphism** (of **rings**) if ϕ is a bijective homomorphism (of rings). In that case we say R is **isomorphic** to S.

Example 5.8.12 Let S denote the subring of $M_2(\mathbb{R})$ consisting of the diagonal matrices (Example 5.4.5), and let R denote the product ring $\mathbb{R} \times \mathbb{R}$ (see Example 5.6.1). Consider the function $\phi : R \to S$ defined by $\phi((x, y)) = \begin{pmatrix} x & 0 \\ 0 & y \end{pmatrix}$. We check that ϕ is a homomorphism. Let $r = (x, y)$ and $r' = (x', y')$ be elements of $R = \mathbb{R} \times \mathbb{R}$. Then 1)

$$\begin{aligned} \phi(r + r') = \phi((x + x', y + y')) &= \begin{pmatrix} x + x' & 0 \\ 0 & y + y' \end{pmatrix} \\ &= \begin{pmatrix} x & 0 \\ 0 & y \end{pmatrix} + \begin{pmatrix} x' & 0 \\ 0 & y' \end{pmatrix} = \phi(r) + \phi(r'), \end{aligned}$$

and similarly 2) $\phi(rr') = \phi(r)\phi(r')$. Since $1_R = (1, 1)$ and $1_S = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, we also have 3) $\phi(1_R) = 1_S$. So ϕ is a homomorphism. Since it is also obviously bijective, ϕ is an isomorphism.

We'll have an important example of a ring isomorphism in the next section. For now, let's note some general properties of isomorphisms.

Proposition 5.8.13 *Suppose that $\phi : R \to S$ is an isomorphism of rings.*

1. *The inverse function $\phi^{-1} : S \to R$ is also an isomorphism of rings.*

2. *If $\psi : S \to T$ is also an isomorphism of rings, then so is the composite $\psi \circ \phi : R \to T$.*

The proofs are almost identical to Prop. 4.12.16 and Prop. 4.12.17 and left as an exercise.

Corollary 5.8.14 *If $\phi : R \to S$ is an isomorphism of groups, then its restriction $R^\times \to S^\times$ is an isomorphism of groups.*

Proof. Let ψ denote the inverse of ϕ, a ring isomorphism by Prop. 5.8.13. According to Prop. 5.8.10, the restrictions of ϕ and ψ define group homomorphisms from $R^\times$ to $S^\times$ and vice-versa. Denote these $\phi^\times$ and $\psi^\times$ as in the proof of Prop. 5.8.10 (so these are the "same" functions as ϕ and ψ but viewed with smaller domain and codomain). Since $\psi(\phi(r)) = r$ for all $r \in R$ and $\phi(\psi(s)) = s$ for all $s \in S$, it follows that $\psi^\times(\phi^\times(r)) = r$ for all $r \in R^\times$ and $\phi^\times(\psi^\times(s)) = s$ for all $s \in S^\times$. Therefore $\phi^\times$ and $\psi^\times$ are inverse functions of each other, so they are bijective, and are therefore isomorphisms. □

Example 5.8.15 Consider the isomorphism $\phi : R \to S$ of Example 5.8.12. The unit group of $R = \mathbb{R} \times \mathbb{R}$ is $R^\times = \mathbb{R}^\times \times \mathbb{R}^\times$ (an exercise). The restriction of ϕ defines an isomorphism from this group to

$$S^\times = \left\{ \left(\begin{array}{cc} x & 0 \\ 0 & y \end{array} \right) \middle| \, x, y \in \mathbb{R}^\times \right\}.$$

5.9 The Chinese Remainder Theorem

Suppose that m and n are positive integers. According to Example 5.8.5, since $m|mn$, there is a homomorphism $\phi : \mathbb{Z}_{mn} \to \mathbb{Z}_m$ defined by $\phi([a]_{mn}) = [a]_m$. Similarly, we have the homomorphism $\phi' : \mathbb{Z}_{mn} \to \mathbb{Z}_n$ defined by $\phi'([a]_{mn}) = [a]_n$. It follows that the function

$$\psi : \mathbb{Z}_{mn} \to \mathbb{Z}_m \times \mathbb{Z}_n \quad \text{defined by } \psi([a]_{mn}) = ([a]_m, [a]_n)$$

is a homomorphism. (In general if $\phi : R \to S$ and $\phi' : R \to S'$ are homomorphisms, then so is the function $\psi : R \to S \times S'$ defined by $\psi(r) = (\phi(r), \phi'(r))$; this is left as an exercise.) Note that the two rings $\mathbb{Z}_{mn}$ and $\mathbb{Z}_m \times \mathbb{Z}_n$ have the same number of elements (namely mn), but computing a few examples shows

that $|psi$ may or may not be an isomorphism. For example, if $m = n = 2$, then the table of values:

$[a]_4$	$\psi([a]_4)$
$[0]_4$	$([0]_2, [0]_2)$
$[1]_4$	$([1]_2, [1]_2)$
$[2]_4$	$([0]_2, [0]_2)$
$[3]_4$	$([1]_2, [1]_2)$

shows that ψ is not bijective in this case. On the other hand if $m = 2$ and $n = 3$, then the table

$[a]_6$	$\psi([a]_6)$
$[0]_6$	$([0]_2, [0]_3)$
$[1]_6$	$([1]_2, [1]_3)$
$[2]_6$	$([0]_2, [2]_3)$
$[3]_6$	$([1]_2, [0]_3)$
$[4]_6$	$([0]_2, [1]_3)$
$[5]_6$	$([1]_2, [2]_3)$

shows that ψ is bijective in this case. It turns out to depend on whether m and n are relatively prime.

Theorem 5.9.1 *If m and n are relatively prime positive integers, then the function*

$$\psi : \mathbb{Z}_{mn} \to \mathbb{Z}_m \times \mathbb{Z}_n$$

defined by $\psi([a]_{mn}) = ([a]_m, [a]_n)$ is an isomorphism of rings.

Proof. We already know that ψ is a homomorphism of rings, and that the two rings have the same number of elements. So it sufiices to prove that ψ is injective, in which case it must also be surjective, hance an isomorphism.

We have to show that if $\psi([a]_{mn}) = \psi([b]_{mn})$, then $[a]_{mn} = [b]_{mn}$. Now $\psi([a]_{mn}) = \psi([b]_{mn})$ means that $([a]_m, [a]_n) = ([b]_m, [b]_n)$, which means $[a]_m = [b]_m$ and $[a]_n = [b]_n$, which means that $m|(b - a)$ and $n|(b - a)$. Since $\gcd(m, n) = 1$, it follows from Cor. 2.3.5 that $mn|(b - a)$, i.e., that $[a]_{mn} = [b]_{mn}$. □

We can view the theorem as a statement about simultaneous solutions of linear congruences. In this form it is known as the **Chinese Remainder Theorem**:

Corollary 5.9.2 *Suppose that $a, b, m, n \in \mathbb{Z}$ with $m, n > 0$. If m and n are relatively prime, then there are integers $x \in \mathbb{Z}$ which simultaneously satisfy the congruences*

$$x \equiv a \bmod m \qquad \textit{and} \qquad x \equiv b \bmod n.$$

Morevoer the solution is uniquely determined modulo mn in the sense that if x_0 satisfies both congruences, then x satisfies both congruences if and only if $x \equiv x_0 \bmod mn$.

Proof. Note that x is a solution of the congruences if and only if $[x]_m = [a]_m$ and $[x]_n = [a]_n$. So letting $\psi : \mathbb{Z}_{mn} \to \mathbb{Z}_m \times \mathbb{Z}_n$ be the isomorphism of Thm. 5.9.1, we see that x is a solution if and only $\psi([x]_{mn}) = ([x]_m, [x]_n) = ([a]_m, [b]_n)$. Since ψ is bijective, there is a unique element $[x]_{mn} \in \mathbb{Z}_{mn}$ such that $\psi([x]_{mn}) = ([a]_m, [b]_n)$. □

The Corollary doesn't say how to find the solution $[x]_{mn}$ in practice. this is done as follows: Use the Euclidean Algorithm to find $r, s \in \mathbb{Z}$ such that $mr + ns = 1$. Note then that

$$\begin{array}{ll} mr \equiv 0 \bmod m, & mr \equiv 1 \bmod n, \\ ns \equiv 1 \bmod n, & ns \equiv 0 \bmod n. \end{array}$$

Therefore letting $x_0 = b(mr) + a(ns)$ gives

$$\begin{array}{rrcccl} & x_0 & \equiv & b \cdot 0 + a \cdot 1 & \equiv & a \mod m \\ \text{and} & x_0 & \equiv & b \cdot 1 + a \cdot 0 & \equiv & b \mod n. \end{array}$$

Therefore the general solution is $x \equiv bmr + ans \bmod mn$.

Example 5.9.3 Let's find all simultaneous solutions in $\mathbb{Z}$ of the congruences

$$x \equiv 121 \bmod 611 \quad \text{and} \quad x \equiv 86 \bmod 421.$$

Applying the Euclidean algorithm gives:

$$\begin{array}{rcl@{\qquad}rcl} 611 &=& 1 \cdot 421 + 190 & 26 &=& 1 \cdot 15 + 11 \\ 421 &=& 2 \cdot 190 + 41 & 15 &=& 1 \cdot 11 + 4 \\ 190 &=& 4 \cdot 41 + 26 & 11 &=& 2 \cdot 4 + 3 \\ 41 &=& 1 \cdot 26 + 15 & 4 &=& 1 \cdot 3 + 1. \end{array}$$

Therefore

$$\begin{array}{rcl} 1 = 4 - 3 &=& 4 - (11 - 2 \cdot 4) = 3 \cdot 4 - 11 \\ &=& 3(15 - 11) - 11 = 3 \cdot 15 - 4 \cdot 11 \\ &=& 3 \cdot 15 - 4(26 - 15) = 7 \cdot 15 - 4 \cdot 26 \\ &=& 7(41 - 26) - 4 \cdot 26 = 7 \cdot 41 - 11 \cdot 26 \\ &=& 7 \cdot 41 - 11(190 - 4 \cdot 41) = 51 \cdot 41 - 11 \cdot 190 \\ &=& 51(421 - 2 \cdot 190) - 11 \cdot 190 = 51 \cdot 421 - 113 \cdot 190 \\ &=& 51 \cdot 421 - 113 \cdot (611 - 421) = 164 \cdot 421 - 113 \cdot 611. \end{array}$$

Since $611 \cdot 421 = 257231$, the solution is

$$x \equiv 121 \cdot 164 \cdot 421 - 86 \cdot 113 \cdot 611 = 2416626 \equiv 101547 \bmod 257231.$$

Example 5.9.4 Now let's work another example, but where a little work is required before applying the above algorithm. We'll find all simultaneous solutions of the congruences:

$$4x \equiv 23 \bmod 57 \qquad \text{and} \qquad 22x \equiv 26 \bmod 84.$$

We first find the solutions of the individual congruences. It is easy to spot that the inverse of $[4]_{57}$ is $[-14]_{57}$, so the solution of the first congruence is

$$x \equiv -14 \cdot 23 \equiv 20 \bmod 57.$$

For the second congruence note that $\gcd(22, 84) = 2$, which divides 26, so there are solutions and the congruence is equivalent to $11x \equiv 13 \bmod 42$. The Euclidean Algorithm applies to 11 and 42 yields $1 = 5 \cdot 42 - 19 \cdot 11$, so the inverse of $[11]_{42}$ is $[-19]_{42}$, and the solution of the second congruence is

$$x \equiv -19 \cdot 13 \equiv 5 \bmod 42.$$

So we are reduced to solving

$$x \equiv 20 \bmod 57 \qquad \text{and} \qquad x \equiv 5 \bmod 42.$$

But note that 42 and 57 are not relatively prime; their gcd is 3. Since $57 = 3 \cdot 19$, and 3 and 19 are relatively prime, we can view the first congruence as equivalent to the pair of congruences

$$x \equiv 20 \bmod 3 \qquad \text{and} \qquad x \equiv 20 \bmod 19,$$

or more simply $x \equiv 2 \bmod 3$ and $x \equiv 1 \bmod 19$. Similarly the second congruence is equivalent to the pair of congruences $x \equiv 2 \bmod 3$ and $x \equiv 5 \bmod 14$. We are therefore looking for simultaneous solutions of the *three* congruences

$$x \equiv 2 \bmod 3, \qquad x \equiv 1 \bmod 19, \qquad \text{and} \qquad x \equiv 5 \bmod 14.$$

We already know that the simultaneous solution of the first two of these congruences is $x \equiv 20 \bmod 57$. So we are finally reduced to solving the pair of congruences

$$x \equiv 20 \bmod 57 \qquad \text{and} \qquad x \equiv 5 \bmod 14.$$

Since $1 = 57 - 4 \cdot 14$ and $57 \cdot 14 = 798$, the solution is

$$x \equiv 5 \cdot 57 - 20 \cdot 4 \cdot 14 = -835 \equiv -37 \bmod 798.$$

We can also apply Thm. 5.9.1 to derive a formula for the order of $\mathbb{Z}_n^\times$. The order of this group is denoted $\varphi(n)$. We can view φ as a function from $\mathbb{N}$ to $\mathbb{N}$, called *Euler's φ-function*. Thus $\varphi(n)$ is the number of integers in $\{0, 1, 2, \ldots, n-1\}$ which are relatively prime to n. Computing a few values we find:

n	1	2	3	4	5	6	7	8	9
$\varphi(n)$	1	1	2	2	4	2	6	4	6.

Recall that if p is prime then $\mathbb{Z}_p^\times = \{[0], [1], \ldots, [p-1]\}$ has order $p-1$, so $\varphi(p) = p-1$. It is also easy to compute the value of φ for prime powers p^r, with $r \geq 1$. Indeed the only integers among $\{0, 1, 2, \ldots, p^r - 1\}$ which are *not* relatively prime to p^r are precisely thos which are multiples of p, of which there are exactly p^{r-1} (namely $0, p, 2p, \ldots, p^r - p$). Therefore

$$\varphi(p^r) = p^r - p^{r-1} = (p-1)p^{r-1}.$$

For example $\mathbb{Z}_9^\times = \{[1], [2], [4], [5], [7], [8]\}$ contains all 9 elements of $\mathbb{Z}_9$ except for the 3 multiples of [3], namely [0], [3] and [6].

To find a general formula, we use the following corollary of Thm. 5.9.1:

Corollary 5.9.5 *If m and n are relatively prime, then $\mathbb{Z}_{mn}^\times$ is isomorphic to $\mathbb{Z}_m^\times \times \mathbb{Z}_n^\times$. In particular, if $\gcd(m, n) = 1$, then $\varphi(mn) = \varphi(m)\varphi(n)$.*

Proof. Thm. 5.9.1 gives an isomorphism of rings from $\mathbb{Z}_{mn}$ to $\mathbb{Z}_m \times \mathbb{Z}_n$. Cor. 5.8.14 then shows that $\mathbb{Z}_{mn}^\times$ is isomorphic to $(\mathbb{Z}_m \times \mathbb{Z}_n)^\times$. According to an exercise, this last group is the same as $\mathbb{Z}_m^\times \times \mathbb{Z}_n^\times$. □

If n is any integer greater than 1, then by the Fundamental Theorem of Arithmetic (Thm. 2.5.4) it has a prime factorization

$$n = p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k}$$

where p_1, p_2, …, p_k are distinct primes and each r_i is a positive integer. Cor. 5.9.5 then shows that

$$\varphi(n) = \varphi(m)\varphi(p_k^{r_k})$$

where $m = p_1^{r_1} p_2^{r_2} \cdots p_{k-1}^{r_{k-1}}$. Repeatedly applying Cor. 5.9.5 (i.e., by induction on k), we conclude that

$$\begin{aligned} \varphi(n) &= \varphi(p_1^{r_1})\varphi(p_2^{r_2}) \cdots \varphi(p_k^{r_k}) \\ &= (p_1 - 1)p_1^{r_1 - 1}(p_2 - 1)p_2^{r_2 - 1} \cdots (p_k - 1)p_k^{r_k - 1}. \end{aligned}$$

So for example,

$$\varphi(2200) = \varphi(2^3 \cdot 5^2 \cdot 11) = \varphi(2^3)\varphi(5^2)\varphi(11) = 2^2 \cdot 4 \cdot 5 \cdot 10 = 800.$$

Finally let's record the following corollary of Lagrange's Theorem, generalizing Fermat's Little Theorem.

Corollary 5.9.6 *Suppose that n is a positive integer and a is an integer relatively prime to n. Then*

$$a^{\varphi(n)} \equiv 1 \bmod n.$$

For example, if a is an integer not divisible by 2, 5 or 11, then $\gcd(a, 2200) = 1$, so $a^{800} \equiv 1 \bmod 2200$.

5.10 Polynomial rings

Let R be any commutative ring. The same construction that defines the polynomial ring $\mathbb{R}[x]$ (Example 5.2.2) can be used to define a commutative ring $R[x]$, called the *polynomial ring over* R (in the *variable* x). Its elements are expressions of the form

$$\sum_{i=0}^{m} a_i x^i = a_m x^m + a_{m-1} x^{m-1} + \cdots + a_1 x + a_0, \qquad \text{where } a_0, a_1, \ldots, a_m \in R.$$

While the polynomial defines a function from R to R, we don't view it as such. Instead we just work purely formally with the algebraic expressions. We define the addition and multiplication operations exactly as for $\mathbb{R}[x]$, and the proof that $R[x]$ is a ring is exactly the same as in the case $R = \mathbb{R}$.

Example 5.10.1 Taking $R = \mathbb{Z}$, the ring $\mathbb{Z}[x]$ can be viewed as the subring of $\mathbb{R}[x]$ consisting of polynomials with integer coefficients. In general, if R is a subring of S, then $R[x]$ can be identified with a subring of $S[x]$.

Example 5.10.2 We will sometimes consider the example where $R = \mathbb{Z}_n$. The elements of $\mathbb{Z}_n[x]$ are then expressions of the form

$$[a_m]x^m + [a_{m-1}]x^{m-1} + \cdots + [a_1]x + [a_0],$$

where $a_0, a_1, \ldots, a_m \in \mathbb{Z}$ and the residue classes are modulo n. Addition and multiplication operations are performed on these polynomials using arithmetic modulo n. For example, if $n = 4$,

$$f(x) = [2]x^2 + [3]x + [1] \quad \text{and} \quad g(x) = [2]x + [1],$$

then

$$\begin{aligned} f(x) + g(x) &= [2]x^2 + [1]x + [2] \\ \text{and} \quad f(x)g(x) &= [4]x^3 + [8]x^2 + [5]x + [1] = [1]x + [1]. \end{aligned}$$

Example 5.10.3 Consider the example where R itself the polynomial ring $\mathbb{R}[x]$. Since we're already using x as a variable in the notation for $\mathbb{R}[x]$, we'll instead consider the ring $R[y]$ of polynomials over $R = \mathbb{R}[x]$ in the variable y. The elements of $R[y] = (\mathbb{R}[x])[y]$ are then polynomials in the variable y, with coefficients that are themselves polynomials in the variable x. So an element of $R[y]$ is an expression of the form $\sum_{i=0}^{n} p_i(x)y^i$, where for each $i = 0, 1, \ldots, n$, the coefficient $p_i(x) \in \mathbb{R}[x]$ has the form $p_i(x) = \sum_{j=0}^{m_i} a_{i,j}x^j$ for some $a_{i,0}, a_{i,1}, \ldots, a_{i,m_i} \in \mathbb{R}$. Here m_i is the degree of $p_i(x)$, but including higher order terms to $p_i(x)$ with coefficient 0, we can assume all the m_i are the same, say m, to simplify notation. So an element of $(\mathbb{R}[x])[y]$ is an expression of the form

$$f(x,y) = \sum_{i=0}^{m} \left(\sum_{j=0}^{n} a_{i,j}x^j \right) y^i = \sum_{i=0}^{m} \sum_{j=0}^{n} a_{i,j}x^j y^i,$$

i.e., a polynomial in the two variables x and y with coefficients in $\mathbb{R}$. There is no difference between the roles of x and y in the last expression, and simply write $\mathbb{R}[x,y]$ for the ring of such polynomials. More generally, for any ring R and any integer $n \geq 1$, we can consider the polynomial ring $R[x_1, x_2, \ldots, x_n]$ over R in n variables.

Recall the definition of the *degree* of a polynomial.

Definition 5.10.4 Suppose that R is a ring and $f(x) \in R[x]$. We say that $f(x)$ has **degree** n, or $\deg(f(x)) = n$, if

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

with $a_n \neq 0_R$.

So if $f(x)$ is any non-zero polynomial, then $\deg(f(x))$ is a non-negative integer; for example $f(x) = x^4 + x + 1 \in \mathbb{R}[x]$ has degree 4.

In the case where R itself is a polynomial ring, such as in Example 5.10.3, the notation should specify the variable in which the degree is considered. For example, the element

$$\begin{aligned} f(x,y) &= x^2y^3 + xy^3 - x^4y + 3xy - 2 \\ &= (x^2 + x)y^3 + (-x^4 + 3x)y - 2 \\ &= (-y)x^4 + y^3x^2 + (y^3 + 3y)x - 2 \end{aligned}$$

of $\mathbb{R}[x, y]$ has degree 3 when viewed as a polynomial in the variable y (with coefficients in $\mathbb{R}[x]$), so we write $\deg_y f(x, y) = 3$, but the degree 4 in the variable x, i.e., $\deg_x f(x, y) = 4$.

Our convention for handling the polynomial $f(x) = 0_R$ is to define its degree as $-\infty$. This is convenient since it's consistent with formulas like those in the foollowing proposition.

Proposition 5.10.5 *Suppose that R is a commutative ring and $f(x), g(x) \in R[x]$. Let $m = \deg(f(x))$ and $n = \deg(g(x))$.*

1. *$\deg(f(x) + g(x)) \leq \max(m, n)$, and equality holds if $m \neq n$;*
2. *$\deg(f(x)g(x)) \leq m + n$, and equality holds if R is an integral domain.*

Proof. First note that if $f(x) = 0_R$, then $m = -\infty$ $f(x) + g(x) = g(x)$ has degree $\max(m, n) = n$ (with the obvious convention that $-\infty \leq n$), and $f(x)g(x)$ has degree $m + n = -\infty$ (again with an obvious convention $-\infty + n = -\infty$). So the formulas hold in this case, and similarly so if $g(x) = 0_R$.

Now assume that $f(x)$ and $g(x)$ are both non-zero, so

$$f(x) = a_m x^m + \cdots a_0 \quad \text{and} \quad g(x) = b_n x^n + \cdots b_0.$$

If $m > n$, then the leading term of $f(x) + g(x)$ is $a_m x^m$, so its degree is $m = \max(m, n)$. Similarly if $n > m$, we find the degree is $n = \max(m, n)$. If $n = m$, then $a_m + b_m$ may be 0_R and the remaining terms of $f(x) + g(x)$ have lower exponent, so the degree is at most $m = \max(m, n)$.

Consider now the degree of

$$f(x)g(x) = a_m b_n x^{m+n} + (a_{m-1}b_n + a_m b_{n-1})x^{m+n-1} + \cdots + a_0 b_0.$$

If R is an integral domain, then $a_m \neq 0_R$ and $b_n \neq 0_R$ implies that $a_m b_n \neq 0_R$, so the degree of $f(x)g(x)$ is $m + n$. Without the assumption that R is an integral domain, we still have that the degree is at most $m + n$. □

Note that in Example 5.10.2 we had $\deg(f(x)g(x)) = 1$, which is strictly less than $\deg(f(x)) + \deg(g(x)) = 2 + 1 = 3$. (Of course $\mathbb{Z}_4$ is not an integral domain since $[2][2] = [0]$.)

Corollary 5.10.6 *Suppose that R is an integral domain. Then*

1. *$R[x]$ is an integral domain;*
2. *$(R[x])^\times = R^\times$.*

Proof. 1) We have to prove that if $f(x)$ and $g(x)$ are non-zero elements of $R[x]$, then $f(x)g(x)$ is also not zero. This is immediate from the part 2) of Prop. 5.10.5, since $\deg(f(x)) \geq 0$ and $\deg(g(x)) \geq 0$ implies that $\deg(f(x)g(x)) = \deg(f(x)) + \deg(g(x)) \geq 0$.

2) Suppose that $f(x) \in (R[x])^\times$. This means that there is a polynomial $g(x) \in R[x]$ such that $f(x)g(x) = 1_R$. Note that $\deg(f(x)) \geq 0$ and $\deg(g(x)) \geq 0$ (since they are non-zero polynomials). Applying part 2) of Prop. 5.10.5 again gives that

$$\deg(f(x)) + \deg(g(x)) = \deg(f(x)g(x)) = \deg(1_R) = 0,$$

which is only possible if $\deg(f(x)) = \deg(g(x)) = 0$. In other words $f(x) = a$ and $g(x) = b$ for some $a, b \in R$. Since $ab = 1_R$, we in fact have $f(x) = a \in R^\times$. □

Example 5.10.7 Since $\mathbb{R}$ is an integral domain (in fact, a field), Cor. 5.10.6 shows that $\mathbb{R}[x]$ is an integral domain and $(\mathbb{R}[x])^\times = \mathbb{R}^\times$. Applying Cor. 5.10.6 again (now with $R = \mathbb{R}[x]$), shows that $\mathbb{R}[x, y] = R[y]$ is also an integral domain and that $(\mathbb{R}[x, y])^\times = (\mathbb{R}[x])^\times = \mathbb{R}^\times$. In fact, we see that in general if R is an integral domain, then so is $R[x_1, x_2, \ldots, x_n]$ and its unit group is $R^\times$ (i.e., the constant poynomials where the constant is in $R^\times$).

Example 5.10.8 Consider the ring $\mathbb{Z}_n[x]$. If n is *prime*, then $\mathbb{Z}_n$ is an integral domain (in fact a field), so Prop. 5.10.6 implies that $\mathbb{Z}_n[x]$ is a domain and its unit group is $\mathbb{Z}_p^\times$. On the other hand if n is *composite*, then $\mathbb{Z}_n$ is not an integral domain and clearly neither is $\mathbb{Z}_n[x]$. It may also be the case that there are units in $\mathbb{Z}_n[x]$ which are not in $\mathbb{Z}_n$. For example, if $n = 4$ then

$$([2]x + [1])([2]x + [1]) = [4]x^2 + [4]x + [1] = [1],$$

so $[2]x + [1]$ is a unit in $\mathbb{Z}_4[x]$.

We will now focus on the case where R is not just an integral domain, but a *field*. Recall this means that every non-zero element of R has a multipicative inverse; for example $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{Z}_p$ (for p prime) are fields. If R is a field, then the polynomial ring $R[x]$ turns out to have some nice properties similar to the ring $\mathbb{Z}$.

Recall that the division algorithm for integers (Thm. 2.1.2) states that if $a \in \mathbb{Z}$ and $b \in \mathbb{N}$, then there are unique integers q and r such that

$$a = bq + r \quad \text{and} \quad 0 \leq r < b.$$

Let's state it slightly differently to emphasize the analogy with the version for polynomial rings which we're about to prove: If $a, b \in \mathbb{Z}$ and $b \neq 0$, then there are unique integers q and r such that

$$a = bq + r \quad \text{and} \quad 0 \leq r < |b|.$$

There is a similar *Division Algorithm* for polynomial rings, with the role of the absolute value played by the degree.

Theorem 5.10.9 *Suppose that R is a field and $f(x), g(x) \in R[x]$ with $g(x) \neq 0_R$. Then there are unique polynomials $q(x), r(x) \in R[x]$ such that*

- *$f(x) = g(x)q(x) + r(x)$, and*
- *either $r(x) = 0_R$ or $\deg(r(x)) < \deg(g(x))$.*

Proof. First we prove existence of $q(x)$ and $r(x)$ as in the theorem.

Let us first take care of the case where $g(x) = b$ is a non-zero constant polynomial. Since R is assumed to be a field, there is an element $b^{-1} \in R$, and we can set $q(x) = b^{-1}f(x)$ and $r(x) = 0_R$ to get $f(x) = g(x)q(x) + r(x)$.

Now suppose that $n = \deg g(x) > 0$. We view $g(x)$ as fixed and prove the existence of $q(x)$ and $r(x)$ by induction on $m = \deg f(x)$. If $m < n$ (or $f(x) = 0_R$), then we can take $q(x) = 0$ and $r(x) = f(x)$. So suppose that $m \geq n$ and that the existence part of the theorem holds (for our $g(x)$) with $f(x)$ replaced by any polynomial of degree *less* than m (including the zero polynomial). Write $f(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_0$ and $g(x) = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_0$ with $a_m \neq 0_R$ and $b_n \neq 0_R$. Then the polynomial

$$\begin{aligned} h(x) &= f(x) - a_m b_n^{-1} x^{m-n} g(x) \\ &= (a_m x^m + a_{m-1}x^{m-1} + \cdots + a_0) \\ &\quad -(a_m x^m + a_m b_n^{-1} b_{n-1} x^{n-1} + \cdots + a_m b_n^{-1} b_0) \\ &= (a_{m-1} - a_m b_n^{-1} b_{n-1})x^{m-1} + \cdots + (a_0 - a_m b_n^{-1} b_0) \end{aligned}$$

has degree at most $m-1$, so the induction hypothesis implies that there are polynomials $s(x), r(x) \in R[x]$ such that

- $h(x) = g(x)s(x) + r(x)$, and
- either $r(x) = 0_R$ or $\deg(r(x)) < \deg(g(x))$.

Therefore

$$\begin{aligned} f(x) &= a_m b_n^{-1} x^{m-n} g(x) + h(x) \\ &= a_m b_n^{-1} x^{m-n} g(x) + g(x)s(x) + r(x) = g(x)q(x) + r(x) \end{aligned}$$

where $q(x) = a_m b_n^{-1} x^{m-n} + s(x)$. This proves the existence of $q(x)$ and $r(x)$ as in the theorem.

Now we prove uniqueness. Suppose that

$$f(x) = g(x)q(x) + r(x) = g(x)s(x) + t(x)$$

with $q(x)$, $r(x)$, $s(x)$ and $t(x)$ in $R[x]$ and the degrees of $r(x)$ and $t(x)$ being less than $n = \deg(g(x))$ (including the possibility that $r(x)$ or $t(x)$ be 0_R). We then have

$$g(x)(q(x) - s(x)) = t(x) - r(x),$$

so if $q(x) - s(x)$ is non-zero, then it has some degree $k \geq 0$, and $t(x) - r(x)$ has degree $n+k \geq n$. On the other hand, since $t(x)$ and $r(x)$ have degree less than n, so does $t(x) - r(x)$. This contradiction shows that $q(x) - s(x) = 0_R$, so $q(x) = s(x)$. It follows that $t(x) - r(x) = 0$ as well, so $r(x) = t(x)$. □

You probably already know the "algorithm" for finding $q(x)$ and $r(x)$ by long division of polynomials, which resembles long division of integers. Let's do an example, which also illustrates the idea of the proof: the successive terms $c_i x^i$ in the quotient are gotten by dividing the leading of $f(x)$ by that of $g(x)$, and then replacing $f(x)$ by the polynomial $f(x) - c_i x^i$ which has lower degree than $f(x)$.

Example 5.10.10 Let $f(x) = x^5 + 2x^3 + x^2 + 1$ and $g(x) = x^3 + x$ in $\mathbb{Q}[x]$. Long division gives

$$\begin{array}{rr|l}
 & & x^2 + 0x + 1 \\
\hline
x^3 + x &) & x^5 + 0x^4 + 2x^3 + x^2 + 0x + 1 \\
 & & x^5 + x^3 \\
\hline
 & & x^3 + x^2 + 0x + 1 \\
 & & x^3 + 0x^2 + x \\
\hline
 & & x^2 - x + 1.
\end{array}$$

So $q(x) = x^2 + 1$ and $r(x) = x^2 - x + 1$. (Note that this is easy to *check* by calculating $g(x)q(x) + r(x)$ and making sure it agrees with $f(x)$.)

Recall that if $m, n \in \mathbb{Z}$, then we say m divides n (written $m|n$) if $m = nk$ for some $k \in \mathbb{Z}$. There is a similar notion of divisibility for any commutative ring R:

Definition 5.10.11 Suppose that R is a ring and $r, s \in R$. Then we say r is **divisible** by s (in R) if $r = st$ for some $t \in R$. If r is divisible by s, we write $s|r$.

(If R is not commutative, then there are obvious notions of *left divisibility* and *right divisibility*.)

Example 5.10.12 If $R = \mathbb{Z}$, this is just the familiar definition.

Example 5.10.13 In $R = \mathbb{Z}_8$, we have that $[6]|[4]$ since $[4] = [6][2]$.

Example 5.10.14 The only element of R divisible by 0_R is 0_R.

Example 5.10.15 If $s \in R^\times$, then every element $r \in R$ is divisible by s since $r = st$ with $t = s^{-1}r$. In particular if R is a field, then $s|r$ for all $r, s \in R$, unless $s = 0_R$ anad $r \neq 0_R$.

Example 5.10.16 We will focus on divisiblity in polynomial rings $R[x]$ where R is a field. For example, the polynomial $x^4 - 1 \in \mathbb{R}[x]$ is divisible by $x^2 + 1$ since $x^4 - 1 = (x^2 + 1)(x^2 - 1)$.

Some basic properties of divisibility, in particular parts 1) and 3) of Prop. 2.1.1, are valid for any commutative ring. For example, if $s|r$ and $t|s$, then $r = ss'$ and $s = tt'$ for some $s', t' \in R$. Therefore $r = ss' = tt's'$ is divisible by t. If R is an integral domain, then there is an analogue of part 2) as well for the polynomial ring $R[x]$. Indeed if $g(x)|f(x)$ and $f(x) \neq 0_R$, then Prop. 5.10.5 part 2) shows that $\deg(g(x)) \leq \deg(f(x))$.

Recall that if R is a field and $f(x)$ and $g(x)$ are polynomials in $R[x]$ with $g(x) \neq 0_R$, then the *Division Algorithm* (Thm. 5.10.9) yields polynomials $q(x), r(x) \in R[x]$ such that

- $f(x) = g(x)q(x) + r(x)$, and
- $\deg(r(x)) < \deg(g(x))$.

Moreover the polynomials $q(x)$ and $r(x)$ are unique in the sense that they are the only polynomials in $R[x]$ satisfying these criteria. So if $g(x)|f(x)$, then we get $f(x) = g(x)q(x)$ and $r(x) = 0_R$. (Recall our convention that $\deg(r(x)) = -\infty$.)

Considering the special case where $g(x) = x - \alpha$ for some $\alpha \in R$, we have the following corollary of the Division Algorithm.

Corollary 5.10.17 *Suppose that R is a field, $f(x) \in R[x]$ and $\alpha \in R$. Then*

1. *$f(x) = (x - \alpha)q(x) + f(\alpha)$ for some $q(x) \in R[x]$;*
2. *$(x - \alpha)|f(x)$ if and only if $f(\alpha) = 0$.*

Proof. The Division Algorithm applied with $g(x) = x - \alpha$ states that there are polynomials $q(x), r(x) \in R[x]$ such that $f(x) = (x-\alpha)q(x) + r(x)$ and $\deg(r(x)) < \deg(x-\alpha) = 1$, so either $r(x) = 0_R$ or $\deg(r(x)) = 0$. In either case we have $r(x) = \beta$ for some $\beta \in R$ (i.e., $r(x)$ is a constant polynomial). Substituting α for x gives

$$f(\alpha) = (\alpha - \alpha)q(\alpha) + \beta = \beta.$$

this proves part 1). Part 2) follows immediately since

$$(x-\alpha)|g(x) \Leftrightarrow r(x) = 0_R \Leftrightarrow f(\alpha) = 0.$$

□

Example 5.10.18 Consider $f(x) = x^n + 1 \in \mathbb{R}[x]$. If n is even, then the remainder of $f(x)$ on division by $x+1 = x-(-1)$ is $f(-1) = (-1)^n + 1 = 2$, and $f(x)$ is not divisible by $x+1$. On the other hand, if n is odd, then $f(-1) = (-1)^n + 1 = 0$, so $f(x)$ is divisible by $x+1$. In fact, if n is odd then

$$x^n + 1 = (x+1)(x^{n-1} - x^{n-2} + \cdots - x + 1).$$

Example 5.10.19 Let $f(x) = x^4 - [1]$ in $\mathbb{Z}_5[x]$. Then $f([1]) = [1]^4 - [1] = [0]$, so $f(x)$is divisible by $x - [1]$. In fact $f([2]) = [2]^4 - [1] = [15] = [0]$, so $f(x)$ is also divisible by $x - [2]$, and by $x - [3]$ since $f([3]) = f([-2]) = [0]$, and by $x - [4]$ since $f([4]) = f([-1]) = [0]$.

In fact for any prime p and any $[a] \in \mathbb{Z}_p^\times$, Fermat's Little Theorem implies that $[a]^{p-1} - [1] = [0]$, so the polynomial $x^{p-1} - [1]$ is divisible by $x - [a]$.

Definition 5.10.20 Suppose that R is a commutative ring, $f(x) \in R[x]$ and $\alpha \in R$. We say that α is a **root** of $f(x)$ if $f(\alpha) = 0_R$.

Corollary 5.10.21 *Suppose that R is a field and $f(x) \in R[x]$ is a polynomial of degree n. If $f(x) \neq 0_R$, then $f(x)$ has at most n roots in R.*

Proof. We prove the corollary by induction on n. If $n = 0$, then $f(x) = a_0$ is a non-zero constant polynomial, and therefore has no roots.

Suppose now that $n > 0$ and that the corollary is true for polynomials of degree $n-1$. Suppose that $f(x) \in R[x]$ is a polynomial of degree n and that $\alpha \in R$ is a root of $f(x)$. Since $f(\alpha) = 0_R$, Cor. 5.10.17 implies that $f(x)$ is divisible by $x - \alpha$, i.e., that $f(x) = (x-\alpha)g(x)$ for some $g(x) \in R[x]$. Note that $\deg(g(x)) = n-1$ since

$$n = \deg(f(x)) = \deg((x-\alpha)g(x)) = \deg(x-\alpha) + \deg(g(x)) = 1 + \deg(g(x)).$$

So by the induction hypothesis $g(x)$ has at most $n-1$ roots.

To complete the proof of the corollary, we will show that if $\beta \in R$ is a root of $f(x)$ different from α, then β is root of $g(x)$. So suppose that $\beta \in R$, $\beta \neq \alpha$ and $f(\beta) = 0_R$. Then

$$0_R = f(\beta) = (\beta - \alpha)g(\beta),$$

but $\beta - \alpha \neq 0_R$. Since R is a field (and in particular an integral domain), it follows that $g(\beta) = 0_R$, so β is a root of $g(x)$. □

We will asssume from now on that R is a field.

Example 5.10.22 Suppose that $f(x) \in R[x]$ has degree 1; i.e., $f(x)$ is *linear*. Then $f(x) = a_1x + a_0$ for some $a_0, a_1 \in R$ with $a_0 \neq 0$. Since R is a field, $a_0 \in R^\times$ and $a_0x + a_1 = 0_R$ has a unique solution, namely $x = -a_0^{-1}a_1$; thus $f(x)$ has exactly one root.

Example 5.10.23 A polynomial $f(x) \in R[x]$ can have fewer than n roots in R (where $n = \deg(f(x))$). In fact it might not have any roots; take for example, $f(x) = x^2 + 1 \in \mathbb{R}[x]$. The polynomial $x^4 - 1 \in \mathbb{R}[x]$ has two roots, namely 1 and -1. Note that either of these polynomials has n roots in $\mathbb{C}$, since i and $-i$ are roots. An example of a polynomial in $\mathbb{C}[x]$ with fewer than n roots is $f(x) = x^2$; its only root in $\mathbb{C}$ is 0. (There is of course a notion of a *repeated* root, of which this is an example.)

Example 5.10.24 If p is prime, then the polynomial $f(x) = x^p - x \in \mathbb{Z}_p[x]$ has p roots, since by Fermat's Little Theorem, every element $[a] \in \mathbb{Z}_p$ satisfies $f([a]) = [0]$.

A polynomial is *irreducible* if it can't be written as a product of polynomials of lower degree. (Recall that we are assuming R is a field.)

Definition 5.10.25 Suppose that $f(x) \in R[x]$ is a polynomial of degree $n > 0$. We say that $f(x)$ is **irreducible** (in $R[x]$) if the following holds:

$$g(x)|f(x) \Rightarrow \deg(g(x)) \in \{0, n\};$$

otherwise we say $f(x)$ is **reducible** (in $R[x]$).

Example 5.10.26 A polynomial of degree 1 (i.e., a *linear* polynomial) is automomatically irreducible.

Example 5.10.27 A polynomial $f(x) \in R[x]$ of degree 2 is reducible if and only if it is divisible by a polynomial $g(x)$ of degree 1. Since $g(x)$ has a root in R (Example 5.10.22), it follows that if $f(x)$ is reducible then $f(x)$ has a root in R. Conversely, if $f(x)$ has a root $\alpha \in R$, then $(x-\alpha)|f(x)$, so $f(x)$ is reducible.

Let's consider the polynomial $f(x) = x^2 + 1_R \in R[x]$ for various fields R.

- If $R = \mathbb{R}$, then $f(x)$ is irreducible.
- If $R = \mathbb{C}$, then $f(x)$ has roots $\pm i$, and $f(x) = (x+i)(x-i)$ is reducible.
- If $R = \mathbb{Z}_2$, then $f(x)$ has $[1]$ as a root, so $f(x)$ is reducible. (In fact $f(x) = (x-[1])^2$ and $[1]$ is a repeated root.)
- If $R = \mathbb{Z}_3$, then $f(x)$ is irreducible.
- If $R = \mathbb{Z}_5$, then $f(x)$ has roots $[2]$ and $[3]$ and $f(x) = (x-[2])(x-[3])$ is reducible.

Example 5.10.28 Suppose $f(x) \in R[x]$ is a polynomial of degree 3. Then $f(x)$ is reducible if and only if $f(x) = g(x)h(x)$ where $g(x)$ and $h(x)$ have degrees 1 and 2. So just as in Example 5.10.27, we see that $f(x)$ is reducible if and only if $f(x)$ has a root in R. So for example

$$f(x) = x^3 + x + [1]_2 \in \mathbb{Z}_2[x]$$

is irreducible since $f([0]) = f([1]) = [1]$. The polynomial

$$f(x) = x^3 + x + [1]_3 \in \mathbb{Z}_3[x]$$

is reducible since $f([1]) = [0]$.

Example 5.10.29 Suppose $f(x) \in R[x]$ is a polynomial of degree $n > 3$. If $f(x)$ has a root $\alpha \in R$, then of course $(x-\alpha)|f(x)$ by Cor. 5.10.17, so $f(x)$ is reducible. (So for example, $x^4 - 1$ is reducible in $\mathbb{R}[x]$.) On the other hand, $f(x)$ can be reducible without having a root. For example,

$$f(x) = x^4 + 3x^2 + 2 = (x^2+1)(x^2+2)$$

is reducible in $\mathbb{R}[x]$, but it has no root in $\mathbb{R}$.

It turns out that the polynomial ring $R[x]$ (where R is a field) has many properties in common with the ring $\mathbb{Z}$. There is for example a Euclidean Algorithm which computes greatest (in degree) common divisors, but we don't have time for this.

The notion of an irreducible polynomial $f(x) \in R[x]$ is similar to that of a prime number $p \in \mathbb{Z}$ in the sense that neither can be written as a product of "smaller" factors. To carry the analogy further, we impose a further condition on $f(x)$ which is in some sense like the requirement that p be positive.

Definition 5.10.30 A polynomial $f(x) \in R[x]$ is **monic** if

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0,$$

i.e., if $a_n = 1_R$ where $n = \deg(f(x))$.

We can then view monic irreducible polynomials in $R[x]$ as playing a role similar to that of prime numbers in $\mathbb{Z}$. In particular, there is the following analogue of the Fundamental Theorem of Arithmetic:

Theorem 5.10.31 *Suppose that $f(x) \in R[x]$ is a monic polynomial of positive degree. Then*

$$f(x) = p_1(x)p_2(x)\cdots p_k(x)$$

for some monic irreducible polynomials $p_1(x), p_2(x), \ldots, p_k(x) \in R[x]$. Moreover this expression is unique except for the possibility of permuting the factors.

We will not give the proof, but it is very similar to the proof of the Fundamental Theorem of Arithmetic. Instead we close with a few examples:

Example 5.10.32 The polynomial $f(x) = x^4 - 1 \in \mathbb{R}[x]$ of Example 5.10.23 has the factorization

$$f(x) = (x-1)(x+1)(x^2+1)$$

into monic ireducibles in $\mathbb{R}[x]$. The polynomial $x^4 + 3x^2 + 2 \in \mathbb{R}[x]$ of Example 5.10.29 has factorization $(x^2+1)(x^2+2)$.

Example 5.10.33 The polynomial $f(x) = x^3 + x + [1] \in \mathbb{Z}_3[x]$ of Example 5.10.28 has factorization

$$f(x) = (x - [1])(x^2 + x - [1]);$$

note that $x^2 + x - [1]$ is irreducible in $\mathbb{Z}_3[x]$ since it has degree 2 and has no roots in $\mathbb{Z}_3$.

Example 5.10.34 The polynomial $f(x) = x^p - x \in \mathbb{Z}_p[x]$ of Example 5.10.24 has factorization

$$f(x) = x(x - [1])(x - [2])\cdots(x - [p-1])$$

since $[0], [1], \ldots, [p-1]$ are all roots of $f(x)$.

5.11 The unit group of $\mathbb{Z}_p$

As an application of Cor. 5.10.21, we will prove an important property of the group $\mathbb{Z}_p^\times$. Recall that $\mathbb{Z}_p^\times$ is an abelian group of order $p-1$; we will prove that it is in fact cyclic. We first need some counting lemmas.

To begin with, we determine the number of elements of each order in the cyclic group $\mathbb{Z}_n$. Recall that we already showed in Cor. 4.8.9 that $\mathbb{Z}_n$ has exactly $\varphi(n)$ generators (where φ denotes Euler's φ-function); that is a special case of the following lemma:

Lemma 5.11.1 *Let n be a positive integer and let d be a positive divisor of n. Then the group $\mathbb{Z}_n$ has exactly $\varphi(d)$ elements of order d.*

Proof. Recall from Cor. 4.8.8 that the order of $[a]_n$ in $\mathbb{Z}_n$ is $n/\gcd(a,n)$; therefore $[a]_n$ has order d if and only if $\gcd(a,n) = n/d$. We must therefore determine the number of elements $[a]_n \in \mathbb{Z}_n$ such that $\gcd(a,n) = d'$, where $d' = n/d$. Recall that if $\gcd(a,n) = d'$, then $\gcd(a/d', n/d') = 1$, so setting $b = a/d'$, we have $\gcd(b,d) = 1$. Conversely if b is any integer relatively prime to d, then $\gcd(bd', n) = d'\gcd(b,d) = d'$. Moreover if b_1 and b_2 are two such integers, then $b_1 \equiv b_2 \bmod d$ if and only if $b_1 d' \equiv b_2 d' \bmod n$, so $[b]_d \mapsto [d'b]_n$ defines a bijection between the congruence classes $[b]_d$ such that $\gcd(b,d) = 1$ and the congruence classes $[a]_n$ such that $\gcd(a,n) = d'$, Therefore the number of such classes $[a]_n$ is precisely the order of $\mathbb{Z}_d^\times$, which is $\varphi(d)$. □

Corollary 5.11.2 *Let n be a positive integer, and let $d_1, d_2, \ldots, d_k$ be the positive divisors of n. Then $n = \varphi(d_1) + \varphi(d_2) + \cdots + \varphi(d_n)$.*

Proof. The order of each element of $\mathbb{Z}_n$ is positive divisor of n, and for each positive divisor d of n, Lemma 5.11.1 shows that there are $\varphi(d)$ elements of $\mathbb{Z}_n$ order d. Since the total number of elements of $\mathbb{Z}_n$ is n, it follows that $n = \varphi(d_1) + \varphi(d_2) + \cdots + \varphi(d_n)$. □

For example, the orders of elements of $\mathbb{Z}_{10}$ are given in the table:

element	[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
order	1	10	5	10	5	2	5	10	5	10.

Note that there are $\varphi(1) = 1$ element of order 1, $\varphi(2) = 1$ of order 2, $\varphi(5) = 4$ of order 5 and $\varphi(10) = 4$ of order 10, and that $1 + 1 + 4 + 4 = 10$.

If p is prime, then the only positive divisors of p are 1 and $p-1$, and in this case Cor. 5.11.2 becomes $p = 1 + (p-1)$. If $k \geq 1$, then the positive

divisors of p^k are $1, p, \ldots, p^k$, and since $\varphi(p^m) = (p-1)p^{m-1}$ for $m = 1, \ldots, k$, the corollary becomes

$$\begin{aligned} p^k = 1 + (p^k - 1) &= 1 + (p-1)(1 + p + \cdots + p^{k-1}) \\ &= 1 + (p-1) + p(p-1) + \cdots + p^{k-1}(p-1). \end{aligned}$$

Lemma 5.11.3 *Suppose G is a group of order n with the property that for each positive divisor d of n, there are at most d elements $g \in G$ such that $g^d = e$. Then G is cyclic.*

Proof. For each $d|n$, let r_d be the number of elements of G of order d. We will prove that $r_d \leq \varphi(d)$. If $r_d = 0$, then the inequality certainly holds, so we can assume $r_d \geq 1$, i.e., that there is an element of G of order d. Let h be such an element and consider the cyclic subgroup $\langle h \rangle$ generated by h. Since $\langle h \rangle$ has order d, every element $g \in \langle h \rangle$ satisfies $g^d = e$, and there are d such elements. By hypothesis, there are at most d elements $g \in G$ satisfying $g^d = e$, so $\langle h \rangle$ contains all the elements of G such that $g^d = e$. In particular, all elements of G of order d are in the cyclic subgroup $\langle h \rangle$, and there are precisely $\varphi(d)$ such elements. Therefore $r_d \leq \varphi(d)$. (In fact we have shown that $r_d = 0$ or $\varphi(d)$.)

Now let $d_1, d_2, \ldots, d_k$ be the positive divisors of n. Since every element of G has order dividing n, and there are r_{d_i} elements of order d_i for each $i = 1, \ldots, k$, we must have

$$r_{d_1} + r_{d_2} + \cdots + r_{d_k} = n = \varphi(d_1) + \varphi(d_2) + \cdots + \varphi(d_n)$$

(as in the proof of Corollary 5.11.2). Since $r_{d_i} \leq \varphi(d_i)$ for each i, we also have for each i that

$$\begin{aligned} r_{d_i} &= n - (r_{d_1} + \cdots + r_{d_{i-1}} + r_{d_{i+1}} + \cdots + r_{d_k}) \\ &\geq n - (\varphi(d_1) + \cdots + \varphi(d_{i-1}) + \varphi(d_{i+1}) + \cdots + \varphi(d_k)) = \varphi(d_i), \end{aligned}$$

so in fact $r_{d_i} = \varphi(d_i)$ for each i. In particular $r_n = \varphi(n) \geq 1$, so G has an element of order n, hence G is cyclic. □

We can now prove:

Theorem 5.11.4 *If p is prime, then $\mathbb{Z}_p^\times$ is cyclic.*

Proof. Since $\mathbb{Z}_p$ is a field, we can apply Cor. **??** to the polynomial $x^d - [1]_p$ for each positive divisor d of $p-1$. It follows that for each such d, there are at most d roots of $x^d - [1]_p$ in $\mathbb{Z}_p$, i.e., there are at most d elements $[a]_p \in \mathbb{Z}_p$ such that $[a]_p^d = [1]_p$. Therefore there at most d elements $[a]_p \in \mathbb{Z}_p^\times$ such that $[a]^d = [1]_p$. It now follows from Lemma 5.11.3 that $\mathbb{Z}_p^\times$ is cyclic. □

Note that the theorem does not give a method of finding a generator of $\mathbb{Z}_p^\times$. We can do this by trial and error for small values of p:

- If $p = 5$, then $\mathbb{Z}_p^\times$ has order 4, and is generated by $[2]$.

- If $p = 7$, then $\mathbb{Z}_p^\times$ has order 6, but $[2]$ has order 3, so is not a generator. Trying $[3]$ instead we find that it is a generator.

- If $p = 11$, then $\mathbb{Z}_p^\times$ has order 10. To check that an element $[a]$ is a generator, it suffices to check that $[a]^d \neq [1]$ for $d = 1$, 2 and 5. We know that the only elements such that $[a]^2 = [1]$ are $[a] = [1]$ and $[a] = [p-1] = [-1]$ so we will not bother to try those. For any other $[a]$, we only need to check that $[a]^5 \neq [1]$. For example $[2]^5 = [-1]$, so $[2]$ is a generator of $\mathbb{Z}_{11}^\times$.

As to the question of whether $\mathbb{Z}_n^\times$ is cyclic for composite values of n, one can show that if p is an odd prime, then $\mathbb{Z}_{p^k}^\times$ for all $k \geq 1$; For example $\mathbb{Z}_9^\times$ is generated by $[2]$. On the other hand if p and q are distinct odd primes, then $\mathbb{Z}_{pq}^\times$ is not cyclic. We leave the proofs of these facts as exercises.

5.12 Irreducibility of polynomials

We now return to the subject of factorization of polynomials with coefficients in a field R. Recall from Theorem 5.12 that the ring $R[x]$ has a unique factorization property very similar to the Fundamental Theorem of Arithmetic for the $\mathbb{Z}$, but prime numbers are replaced by monic irreducible polynomials. In order to determine a factorization of a polynomial into irreducibles, we need to be able to decide whether a polynomial is irreducible. Recall that a polynomial of degree one is always irreducible, and a polynomial of degree two or three is irreducible if and only if it has no roots, but what about polynomials of agree at least 4?

Consider what happens for a few familiar fields R:

- If $R = \mathbb{Z}_p$ for a prime p, then there are only finitely many polynomials of each degree, so in principle one can check whether a given polynomial is divisible by any polynomial of lower degree. For example one can can check that the polynomial $f(x) = x^4 + x^3 + x^2 + x + [1] \in \mathbb{Z}_3[x]$ has no roots, and is not divisible by any of the 9 monic polynomials $x^2 + [a]x + [b]$ with $[a], [b] \in \mathbb{Z}_3$. Therefore $f(x)$ is irreducible.

- In the case $R = \mathbb{C}$, the only irreducible polynomials are those of degree one. This follows from the Fundamental Theorem of Algebra, whose proof is beyond the scope of this module. This theorem states that every non-constant polynomial $f(x) \in \mathbb{C}[x]$ has a root $\alpha \in \mathbb{C}$, hence

$f(x)$ is divisible by $x - \alpha$ for some α, so the only monic irreducible polynomials in $\mathbb{C}[x]$ are those of the form $x - \alpha$.

- It follows that if $R = \mathbb{R}$, then every irreducible polynomial has degree at most two. Indeed any $f(x) \in \mathbb{R}[x]$ can be viewed as a polynomial in $\mathbb{C}[x]$, so if $f(x)$is non-constant, it has a root $\alpha \in \mathbb{C}$. If in fact $\alpha \in \mathbb{R}$, then $f(x)$ is divisible by $x - \alpha$ in $\mathbb{R}[x]$. On the other hand if $\alpha = a + bi$ with $a, b \in \mathbb{R}$ and $b \neq 0$, then its complex conjugate $\overline{\alpha} = a - bi$ is also a root of $f(x)$ since $f(\overline{\alpha}) = \overline{f(\alpha)} = \overline{0} = 0$. Therefore both $x - \alpha$ and $x - \overline{\alpha}$ are irreducible factors of $f(x)$ in $\mathbb{C}[x]$, so $f(x)$ is divisible in $\mathbb{C}[x]$ by
$$p(x) = (x - \alpha)(x - \overline{\alpha}) = x^2 - 2ax + (a^2 + b^2).$$
Note that $p(x) \in \mathbb{R}[x]$, and dividing $f(x)$ by $p(x)$ yields the same quotient and remainder regardless of whether we apply the Division Algorithm in $\mathbb{R}[x]$ and $\mathbb{C}[x]$, so it follows that in fact $f(x)$ is divisible by $p(x)$ in $\mathbb{R}[x]$.

We will now turn to the case of $R = \mathbb{Q}$, for which it is useful to focus on polynomials in $\mathbb{Z}[x]$. We make the following definition:

Definition 5.12.1 A polynomial $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ in $\mathbb{Z}[x]$ is *primitive* if $\gcd(a_0, a_1, \ldots, a_{n-1}, a_n) = 1$.

We will also make use of the homomorphisms $\phi : \mathbb{Z}[x] \to \mathbb{Z}_p[x]$ defined as follows: if $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{Z}[x]$, then $\phi(f(x)) = [a_n]_p x^n + [a_{n-1}]_p x^{n-1} + \cdots + [a_1]_p x + [a_0]_p$. In other words, replace each coefficient by the corresponding congruence class modulo p. Writing $[f]_p(x)$ for $\phi(f(x))$, it is straightforward to check that $[f + g]_p(x) = [f]_p(x) + [g]_p(x)$ and $[fg]_p(x) = [f]_p(x)[g]_p(x)$, so ϕ is indeed a homomorphism. (Note that if $a \in \mathbb{Z}$ is identified with a constant polynomial in $\mathbb{Z}[x]$, then $\phi(a)$ is the constant polynomial $[a]_p$, consistently with previous notation; in particular $\phi(1) = [1]_p$ and $\phi(0) = [0]_p$.)

Proposition 5.12.2 *Suppose that $f(x) \in \mathbb{Q}[x]$ be a non-zero polynomial.*

1. *There is a unique positive rational number s such that $sf(x) \in \mathbb{Z}[x]$ and $sf(x)$ is primitive.*

2. *Suppose that $f(x) \in \mathbb{Z}[x]$. Then $f(x)$ is primitive if and only if $[f]_p(x)$ is non-zero for all primes p.*

3. *Suppose that $f(x), g(x) \in \mathbb{Z}[x]$. Then $f(x), g(x)$ are both primitive if and only if $f(x)g(x)$ is primitive.*

Proof. 1. Let $f(x) = r_n x^n + \cdots + r_1 x + r_0$ with $r_0, r_1, \ldots, r_n \in \mathbb{Q}$ not all 0. Write $r_i = a_i/b_i$ with $a_i, b_i \in \mathbb{Z}$ and $b_i > 0$ for $i = 0, 1, \ldots, n$. Letting $B = b_0 b_1 \cdots b_n$, we see that $Bf(x) \in \mathbb{Z}[x]$. Write $h(x) = c_n x^n + \cdots + c_1 x + c_0$ and let $C = \gcd(c_0, c_1, \ldots, c_n)$. Then $C^{-1}h(x) \in \mathbb{Z}[x]$ and it is primitive since $\gcd(c_0/C, c_1/C, \ldots, c_n/C) = 1$. Setting $s = B/C$, we conclude that $sf(x) \in \mathbb{Z}[x]$ and $sf(x)$ is primitive.

To prove the uniqueness, suppose that s_1 and s_2 are positive rational numbers such that $g_1 = s_1 f(x)$ and $g_2 = s_2 f(x)$ are both primitive polynomials in $\mathbb{Z}[x]$. Let $t = s_1/s_2$, which we can write as m_1/m_2 for some positive integers m_1, m_2, so that $m_1 g_2(x) = m_2 g_1(x)$. Since $g_1(x)$ is primitive, the greatest common divisor of the coefficients of $m_2 g_1(x)$ is m_2; similarly since $g_2(x)$ is primitive the greatest common divisor of the coefficients of $m_1 g_2(x)$ is m_1. Therefore $m_1 = m_2$, so $s_1 = s_2$.

2. Let $f(x) = a_n x^n + \cdots + a_1 x + a_0$. Suppose that $f(x)$ is primitive. Let p be a prime. Since $\gcd(a_0, a_1, \ldots, a_n)$ is not divisible by p, there is some $i \in \{0, 1, \ldots, , n\}$ such that a_i is not divisible by p, so $[a_i]_p \neq [0]_p$, and therefore $[f]_p(x) \neq [0]_p$.

If on the other hand that $f(x)$ is not primitive, then $\gcd(a_0, a_1, \ldots, a_n)$ is divisible by some prime p. Therefore a_i is divisible by p for $i = 0, 1, \ldots, a_n$, so $[f]_p(x) = [0]_p$.

3. Suppose that $f(x)$ and $g(x)$ are primitive. Then by 2), for each prime p, $[f]_p(x)$ and $[g]_p(x)$ are non-zero. Since $\mathbb{Z}_p[x]$ is an integral domain, it follows that $[fg]_p(x) = [f]_p(x)[g]_p(x) \neq [0]_p$ for all p, so applying 2) again, we conclude that $f(x)g(x)$ is primitive.

Conversely suppose that $f(x)g(x)$ is primitive. Then $[f]_p(x)[g]_p(x) = [fg]_p(x)$ is non-zero for all primes p by 2). Therefore $[f]_p(x)$ and $[g]_p(x)$ are both non-zero for all p, so $f(x)$ and $g(x)$ are primitive, by 2) again. □

We can now prove Gauss's Lemma:

Lemma 5.12.3 *Suppose that $f(x)$ and $g(x)$ are primitive polynomials in $\mathbb{Z}[x]$. If $g(x)|f(x)$ in $\mathbb{Q}[x]$, then $g(x)|f(x)$ in $\mathbb{Z}[x]$.*

Proof. If $f(x)$ is divisible by $g(x)$ in $\mathbb{Q}[x]$, then $f(x) = g(x)h(x)$ for some $h(x) \in \mathbb{Q}[x]$, and we wish to show that $h(x) \in \mathbb{Z}[x]$. By part 1) of Prop. 5.12.2, we can write $h(x) = sk(x)$ for some positive $s \in \mathbb{Q}$ and some primitive polynomial $k(x) \in \mathbb{Z}[x]$. Now $f(x) = sg(x)k(x)$, and since $g(x)$ and $k(x)$ are both primitive, part 3) of Proposition 5.12.2 implies that $g(x)k(x)$ is primitive. Since by hypothesis, $f(x)$ is also primitive, it follows from the uniqueness in part 1) of the proposition that $s = 1$, and therefore $h(x) = k(x) \in \mathbb{Z}[x]$. □

Note that to check divisibility in $\mathbb{Q}[x]$, one can always replace polynomials by non-zero scalar multiples and hence assume the polynomials are primitive

polynomials in $\mathbb{Z}[x]$, so that Gauss's Lemma applies. Moreover if $f(x)$ is divisible by $g(x)$ in $\mathbb{Z}[x]$, then $[f]_p(x)$ is divisible by $[g]_p(x)$ in $\mathbb{Z}_p[x]$ for all primes p. We therefore have:

Corollary 5.12.4 *If $f(x)$ is a monic polynomial in $\mathbb{Z}[x]$, and $[f]_p(x)$ is irreducible in $\mathbb{Z}_p[x]$ for some prime p, then $f(x)$ is irreducible in $\mathbb{Q}[x]$.*

Example 5.12.5 Consider the polynomial $f(x) = x^4+x^3+x^2+x+1 \in \mathbb{Q}[x]$. Since $[f]_3(x)$ is irreducible in $\mathbb{Z}_3[x]$, $f(x)$ is irreducible in $\mathbb{Q}[x]$.

The following is a useful way to prove that a polynomial is irreducible. The hypothesis of the theorem is known as *Eisenstein's Criterion*.

Theorem 5.12.6 *Suppose that $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ is a primitive polynomial in $\mathbb{Z}[x]$. If there is a prime number p such that $a_0, a_1, \ldots, a_{n-1}$ are all divisible by p and a_0 is not divisible by p^2, then $f(x)$ is irreducible in $\mathbb{Q}[x]$.*

Proof. Note that since $f(x)$ is primitive, and $a_0, a_1, \ldots, a_{n-1}$ are all divisible by p, we must have $p \not| a_n$. In particular $f(x)$ has degree n.

Suppose that $f(x)$ is reducible, so $f(x) = g(x)h(x)$ for some polynomials $g(x) = b_d x^d + \cdots + b_0$, $h(x) = c_{n-d}x^{n-d} + \cdots + c_0$, with $0 < d < n$. By Gauss's Lamma, we can assume $g(x)$ and $h(x)$ are primitive polynomials in $\mathbb{Z}[x]$, so that $[f]_p(x) = [g]_p(x)[h]_p(x)$. Since $[f]_p(x) = [a_n]_p x^n$, it follows from the uniqueness of factorization in Thm. that the only possible monic irreducible factor in $\mathbb{Z}_p[x]$ of $[g]_p(x)$ and $[h]_p(x)$ is x, and therefore $[g]_p(x) = [b_d]_p x^d$ and $[h]_p(x) = [c_{n-d}]_p x^{n-d}$. Since $0 < d < n$, this implies that $[b_0]_p = [c_0]_p = [0]_p$. Since $p|b_0$ and $p|c_0$, it follows that $p^2|b_0c_0 = a_0$, contradicting the hypotheses of the theorem. □

We close with some examples:

- The polynomial $f(x) = x^4 - 2$ satisfies the hypotheses of the theorem with $p = 2$, so $f(x)$ is irreducible in $\mathbb{Q}[x]$.

- We now give a less tedious proof that the polynomial $f(x) = x^4 + x^3 + x^2 + x + 1 \in \mathbb{Q}[x]$ of Example 5.12.5 is irreducible. Since the constant term of $f(x)$ is not divisible by any prime, it cannot satisfy Eisenstein's Criterion. We will show however that $g(x) = f(x + 1)$ does satisfy the criterion, so $g(x)$ is irreducible, and it follows (by an exercise) that $f(x)$ is irreducible. Note that $f(x) = (x^5 - 1)/(x - 1)$, so

$$g(x) = ((x + 1)^5 - 1)/x = x^4 + 5x^3 + 10x^2 + 10x + 5,$$

  which satisfies Eisenstein's Criterion for $p = 5$.