Introduction to abstract algebra

Svetlana Makarova

Maximilien Péroux

Email address: murmuno@gmail.com

Email address: peroux@msu.edu

Contents

Bibliography 167

Preface

There are starred sections ($\star$) that can be skipped on your first reading of the book; you should read them only if you really want to.

These notes are still incomplete and some sections are missing. Last edit: May 19, 2024

CHAPTER 1

Sets

1.1. Reminder about math notation

In these notes, we will be using various mathematical symbols to express mathematical ideas more concisely.

These are commonly used logical operators:

- the *implication operator* is written as $\Rightarrow$ and reads "implies";
- $\Leftrightarrow$ reads "if and only if" or "exactly when".

These two symbols are called *quantifiers*:

- the *universal quantifier* is written as $\forall$ and reads "for all" or "for any";
- the *existential quantifier* is written as $\exists$ and reads "exist(s)" or "there is/are".

On a related note, we write $\nexists$ to mean "there is no such".

We will also use $\coloneqq$ when we define the object on the left as the expression on the right. For example, $x \coloneqq 2y$ reads as "let x be equal to $2y$".

1.2. Semi-axiomatic definitions

Most mathematicians think of set theory as the foundation of mathematics [Stacks, Tag 0009]. Every mathematical object, when you unravel its definitions, can be expressed in terms of sets. For example, what is a linear operator between two vector spaces? It is a function $\phi : V \to W$ that satisfies certain properties. A function can be defined as a graph – a sub*set* in the Cartesian product $V \times W$. A vector space is a *set* of elements with the additional structure of addition and multiplication by scalars... And so it goes. It would take a considerable amount of work to write all of the definitions formally until we manage to boil everything down to sets, but it should be possible to do so for every mathematical object.

Definition 1.2.1. By a *set* we mean a collection of elements. If x is an element of a set X, we write it as $x \in X$; otherwise, we write $x \notin X$.

We will proceed with listing some of the Zermelo–Fraenkel axioms of set-theory. The motivation for building set theory axiomatically arose in 1901, when a British philosopher and mathematician Bertrand Russell discovered that "naïve set theory" has contradictions. Naïvely, we could try to consider the "set of all sets" S, and define its subset X whose elements are all sets $Y \in S$ which are not elements of themselves, namely $Y \notin Y$. But then we can ask whether X is an element of X. If it is, then by the description of elements in X we have $X \notin X$, which is a contradiction. But if it is not, so $X \notin X$, then again by definition of X we get that it should belong to itself. In symbols, the paradox can be formulated as follows:

$$\text{Let } X := \{Y \in S \mid Y \notin Y\}, \text{ then } X \in X \iff X \notin X.$$

So we conclude that we should not allow to consider the "set of all sets", and that it is safer to build set theory starting from "smaller sets", and get new sets by means of allowable constructions on existing sets. This is a rough idea of what our list of axioms should do.

Here is the list of some axioms and properties of sets that we will be using, the list is based on *Zermelo–Fraenkel axioms* (ZF axioms):

- **(i)** There is exactly one set with no elements, it is called *the empty set* and is denoted by $\varnothing$.
- **(ii)** Two sets are equal (that is, they are the same set) if they have the same elements.
- **(iii)** If X is a set, then we can specify a subset Y of elements of X that satisfy a certain property ϕ, and Y is a set. We will write it using *set-builder* notation: $Y = \{x \in X \mid \phi(x)\}$.
- **(iv)** If X and Y are sets, then there exists a set S that has X and Y as its elements.
- **(v)** The union over the elements of a set S exists. For example:

$$\bigcup \{\{1,2\},\{2,3\},\{4,5\}\} = \{1,2,3,4,5\}.$$

- **(vi)** There exists an infinite set.
- **(vii)** For any set X, there exists a set that contains every subset of X, and it is denoted by 2^X or $\mathcal{P}(X)$.

Remark 1.2.2. Given a set X of sets, the axiom **(v)** allows us to define a set $\bigcup_{S \in X} S$, which is the union of all sets in X.

It is not too important for this course to know the full list of axioms, but rather know that it exists. In the full list there is the "axiom of foundation", not formulated above, which is used to prove in particular that there does not exist a set which is an element of itself, so the formulation of Russell's paradox becomes nonsensical.

We will also assume *the axiom of choice*, which has been historically controversial, yet useful for streamlining arguments in pure mathematics and eliminating technical assumptions.

Axiom 1.2.3 (Axiom of Choice)**.** Let X be a set of sets whose members are all nonempty. Then there exists a function

$$f : X \longrightarrow \bigcup_{S \in X} S$$

from X to the union of the members of X, called a "choice function", such that for all $Y \in X$ one has $f(Y) \in Y$.

The axiom of choice is logically independent from ZF, and most mathematicians assume it in their work. ZF with the added axiom of choice is called ZFC. The axiom of choice only matters for infinite sets. Some mathematicians try to avoid using the axiom of choice because it is not "constructible": it only states that a choice function exists, but it doesn't describe it.

1.3. Main constructions

Definition 1.3.1. We write $Y \subseteq X$ or $Y \subset X$ when $\forall y \in Y : y \in X$, and we say then that Y is a *subset* of X.

For example, for any set X, we have $\varnothing \subset X$ and $X \subset X$.

Definition 1.3.2. We say that $Y \subseteq X$ is a *proper subset* of X if Y is not equal to X, and in this case we write $Y \subsetneq X$.

The empty set $\varnothing$ is the only set that does not have proper subsets. If X is any nonempty set, $\varnothing$ will be a proper subset.

Example 1.3.3. The following list contains the standard notation for various sets of numbers:

- $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ – set of natural numbers;
- $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ – set of integers;
- $\mathbb{Q}$ – set of rational numbers;
- $\mathbb{R}$ – set of real numbers;
- $\mathbb{C}$ – set of complex numbers.

These sets fit into a sequence of inclusions:

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}.$$

Definition 1.3.4. Given two sets X and Y, we want to introduce the following constructions:

- The *union* $X \cup Y$ can be defined as $X \cup Y \coloneqq \bigcup\{X, Y\}$ as in the ZF Axiom **(v)**, or in other words
$$X \cup Y \coloneqq \{x \mid x \in X \text{ or } x \in Y\},$$
using the set-builder notation from the ZF Axiom **(iii)**.
- The *intersection* $X \cap Y$ is defined as the set of all elements that X and Y share:
$$X \cap Y \coloneqq \{x \mid x \in X \text{ and } x \in Y\}.$$
- The *difference* $X \setminus Y$ is the set of those elements of X that do not belong to Y:
$$X \setminus Y \coloneqq \{x \in X \mid x \notin Y\}.$$
- The *Cartesian product* $X \times Y$ is the set of pairs:
$$X \times Y \coloneqq \{(x, y) \mid x \in X \text{ and } y \in Y\}.$$
The name comes from the analogy with cartesian coordinates on the plane: we can informally think of X as "the first coordinate" and Y as "the second coordinate".
- The *power set*, denoted $\mathcal{P}(X)$ or 2^X, is defined as in the ZF axiom **(vii)**, or equivalently in symbols:
$$\mathcal{P}(X) = 2^X \coloneqq \{Y \mid Y \subset X\}.$$

Example 1.3.5. We can describe power sets for some small examples:

$$\mathcal{P}(\varnothing) = \{\varnothing\},$$
$$\mathcal{P}(\{1\}) = \{\varnothing, \{1\}\},$$
$$\mathcal{P}(\{1, 2\}) = \{\varnothing, \{1\}, \{2\}, \{1, 2\}\}.$$

Definition 1.3.6. The *cardinality* of a set X, which we denote by $|X|$, is the number of elements in that set. When X is not a finite set, we write $|X| = \infty$.

Exercise 1.3.7. If X is a finite set of cardinality $|X| = n$, then $|2^X| = 2^n$.

Definition 1.3.8. We say that two sets X and Y are *disjoint* if $X \cap Y = \varnothing$. When we take the union of two disjoint sets X and Y, we call it the *disjoint union* and denote it by $X \sqcup Y$. If we have two sets that possibly share elements, we may still want to take their disjoint union, but in this case we will have to add "duplicates" of the elements in the intersection, for example by decorating them with a prime or tilde. More formally, we can define the disjoint union of X_1 and X_2 to be:

$$X_1 \sqcup X_2 \coloneqq \{(x_i, i) \mid i \in \{1, 2\} \text{ and } x_i \in X_i\} \subseteq (X_1 \cup X_2) \times \{1, 2\}.$$

Example 1.3.9. Sets $X = \{1, 2, 3\}$ and $Y = \{4, 5, 6\}$ are disjoint, and their union is equal to their disjoint union:

$$X \cup Y = X \sqcup Y = \{1, 2, 3, 4, 5, 6\}.$$

However X and $Z = \{1, 7\}$ are not disjoint, and if we want to take their disjoint union, we will have to repeat 1:

$$X \sqcup Z = \{1, 2, 3, 1', 7\}.$$

So we see that when X and Z are not disjoint, their union $X \cup Z$ and disjoint union $X \sqcup Z$ are different sets.

1.4. Functions

Definition 1.4.1. Let X and Y be sets. A *function / map /mapping* f from X to Y, written as

$$f : X \to Y \text{ or } X \xrightarrow{f} Y,$$

is a subset of the direct product $f \subset X \times Y$, in which each $x \in X$ appears as the first component of exactly one ordered pair $(x, y) \in f$. Functions from X to Y form a set, and this set is denoted by $\mathrm{Hom}\, X, Y$ or $\mathrm{Hom}\, \mathsf{Set} X, Y$.

Remark 1.4.2. Informally, a function $f\colon X \to Y$ can be understood as a rule that assigns to each $x \in X$ exactly one $y \in Y$, and this y is then written as $f(x)$. When f is understood not as a rule, but as a subset of the direct product, we will usually write $\Gamma(f)$ and call it the *graph* of the function f. When we want to write the rule explicitly on elements, we can use the following notation:

$$\begin{aligned} f\colon X &\to Y, \\ x &\mapsto f(x). \end{aligned}$$

Definition 1.4.3. Given a function $f\colon X \to Y$, we call X the *domain* of f and Y the *codomain* or *range* of f.

Definition 1.4.4. Consider a function $f\colon X \to Y$.

(i) The *image* of f is the following subset of the codomain:

$$\mathrm{Im}\, f \coloneqq \{y \in Y \mid \exists x \in X\colon y = f(x)\} = \{f(x) \mid x \in X\}.$$

(ii) Given a subset $Z \subset Y$, its *preimage under* f is the following subset in the domain:

$$f^{-1}(Z) \coloneqq \{x \in X \mid f(x) \in Z\}.$$

If the subset Z only has one element, we will simplify the notation:

$$f^{-1}(y) \coloneqq f^{-1}(\{y\}).$$

An element $x \in f^{-1}(y)$ is called *a preimage* of y.

Definition 1.4.5. We say that a function $f\colon X \to Y$ is:

- *injective* if every element of Y has either one preimage or none, or equivalently, for any choice of elements $x_1, x_2 \in X$, if $f(x_1) = f(x_2)$, then $x_1 = x_2$.
- *surjective* if $\mathrm{Im}\, f = Y$, or equivalently, if every element y of Y has at least one preimage, so $\exists x \in X$ such that $f(x) = y$.
- *bijective* if it is both injective and surjective.

Example 1.4.6. Let $X = \{1, 2, 3\}$ and $Y = \{4, 5, 6\}$. The following subsets of $X \times Y$ cannot be graphs of functions:

- $S_1 = \{(1, 4), (2, 6), (3, 4), (1, 5)\}$ assigns two values to $1 \in X$;
- $S_2 = \{(1, 4), (2, 6)\}$ does not assign any value to $3 \in X$.

Example 1.4.7. The function $f\colon \mathbb{Z} \to \mathbb{Z}$ that takes x to x^3

$$\begin{aligned} f\colon \mathbb{Z} &\to \mathbb{Z}, \\ x &\mapsto x^3, \end{aligned}$$

is injective but not surjective, for instance 7 does not have a cube root and hence does not have a preimage under f.

Example 1.4.8. The function $f\colon \mathbb{R} \to \mathbb{R}$ that takes x to x^3 is bijective because every real number has exactly one cube root.

Example 1.4.9. The function $f\colon \mathbb{C} \to \mathbb{C}$ that takes x to x^3 is surjective but not injective because every nonzero number has three cube roots that differ by cube roots of unity.

1.5. Equivalence relations

Definition 1.5.1. A *relation* $\mathcal{R}$ on a set X is a subset $\mathcal{R} \subset X \times X$. We read $(x, y) \in \mathcal{R}$ as "x is related to y", and we write $x\mathcal{R}y$, or $x \sim y$ when no confusion may arise.

Definition 1.5.2. An *equivalence relation* $\mathcal{R}$ on a set X is a relation $\mathcal{R} \subset X \times X$ that satisfies these axioms:

- **(i)** *reflexivity:* $\forall x \in X : x \sim x$;
- **(ii)** *symmetry:* $\forall x, y \in X : x \sim y \Rightarrow y \sim x$;
- **(iii)** *transitivity:* $\forall x, y, z \in X : x \sim y$ and $y \sim z$ imply $x \sim z$.

When $x \sim y$ for an equivalence relation, we say that x and y are *equivalent*.

Example 1.5.3. There is only one relation on the empty set because $\varnothing \times \varnothing = \varnothing$. Given any set X, we have the *empty relation* $\mathcal{R} = \varnothing \subset X \times X$. The empty relation is an equivalence relation for $X = \varnothing$, otherwise reflexivity fails.

Example 1.5.4. $\mathcal{R}$ = "equality", or more precisely, for a set X we define $\mathcal{R} = \{(x, x) \mid x \in X\}$. This relation always is an equivalence relation.

Example 1.5.5. Let X be the set of people that are currently alive. We define the relation $x \sim y$ when x and y are of the same age. This is an equivalence relation.

Example 1.5.6. Let X be the set of people that are currently alive. Define the relation $x\mathcal{R}y$ when x is a parent of y. This is not an equivalence relation because reflexivity and symmetry fail, even though $\mathcal{R}$ is transitive.

Example 1.5.7. Let X be the set of people that are currently alive. Define the relation $x\mathcal{R}y$ when x and y share a biological parent. Reflexivity and symmetry hold for $\mathcal{R}$. However this is not an equivalence relation because x may have parents a and b, y may have parents b and c, and z may have parents c and d, so transitivity fails.

Definition 1.5.8. A *partition* of a set X is a collection of nonempty subsets of X such that every element of X is in exactly one of the subsets in the partition. These subsets are called *cells* of the partition. For $x \in X$, denote by $[x]$ the cell to which it belongs.

Theorem 1.5.9. Fix a set X.

- **(i)** An equivalence relation yields a partition of X as follows: for any $x \in X$, define the cell $[x]$ as
$$[x] = \{t \in X \mid x \sim t\}.$$
- **(ii)** Conversely, a given partition provides an equivalence relation on X:
$$x \sim y \text{ whenever } [x] = [y].$$

Moreover, these assignments are inverse to each other.

Proof. We first prove **(i)**. Given an equivalence relation on X, we see from reflexivity that every element x belongs to at least one cell, namely $[x]$, because $x \sim x$. Now we want to show that x belongs to exactly one cell, in other words, if $x \in [y]$, then $[x] = [y]$.

Let us first show that $[x] \subseteq [y]$. Assume that we have $t \in [x]$, which means $x \sim t$ by definition of $[x]$. The assumption $x \in [y]$ means $x \sim y$, which by symmetry

is equivalent to $y \sim x$. But now transitivity applies to $y \sim x$ and $x \sim t$, and we get $y \sim t$, or in other words $t \in [y]$.

The same argument applies to show $[y] \subseteq [x]$. Hence we get an inclusion of sets in both directions, and therefore $[x] = [y]$. This concludes verifying Definition 1.5.8.

Now let us prove **(ii)** by checking conditions in Definition 1.5.2. Since $[x] = [x]$, reflexivity follows. If $[x] = [y]$, then $[y] = [x]$, hence symmetry holds. Finally, if we have two equalities of sets $[x] = [y]$ and $[y] = [z]$, then $[x] = [z]$, and this shows transitivity. □

Definition 1.5.10. The cells induced by an equivalence relation are called *equivalence classes*. For example, $[x]$ is the equivalence class of x.

Definition 1.5.11. Given an equivalence relation on X, we call the set of equivalence classes the *quotient set* and denote it by $X/\sim$:

$$X/\sim := \{[x] \mid x \in X\}.$$

It comes equipped with the *canonical projection*

$$\begin{aligned} \pi \colon X &\longrightarrow X/\sim, \\ x &\longmapsto [x]. \end{aligned}$$

Remark 1.5.12. The canonical projection $\pi : X \to X/\sim$ is surjective, and $\pi(x) = \pi(y)$ if and only if $x \sim y$.

Example 1.5.13. We can partition $\mathbb{Z}$ into even integers and odd integers. Then the quotient set will be $\mathbb{Z}/\sim = \{[0], [1]\}$, and the canonical projection $\pi \colon \mathbb{Z} \to \{[0], [1]\}$ takes all even integers to $[0]$ and all odd integers to $[1]$. For example, $\pi(0) = \pi(10) = \pi(-14) = [0]$ and $\pi(1) = \pi(7) = [1]$.

1.6. ⋆ Partially ordered sets and Zorn's lemma

Definition 1.6.1. Let P be a set. Let $\leq$ be a relation on P, as in Definition 1.5.1. We say the relation $\leq$ is a *partial order* on P if it satisfies the following axioms:

- **(i)** *reflexivity:* $x \leq x$, for all $x \in P$;
- **(ii)** *antisymmetry:* if given $x, y \in P$ we have $x \leq y$ and $y \leq x$, then $x = y$;
- **(iii)** *transitivity:* if given $x, y, z \in P$ we have $x \leq y$ and $y \leq z$, then $x \leq z$.

We refer to the pair $(P, \leq)$ as a *partially ordered set*, or simply *poset*.

Example 1.6.2. A poset is a set in which one can compare elements. Most of "less or equal" comparisons the reader is familiar with form posets. For instance $(\mathbb{N}, \leq)$. One can replace $\mathbb{N}$ by integers, rationals, reals or complex numbers. One can as well replace the partial order by "greater or equal". Hence $(\mathbb{N}, \geq)$ is also a poset. "Strictly less than" is usually not a poset as reflexivity is not satisfied.

Example 1.6.3. Let S be a set. Then the power set $\mathcal{P}(S)$ together with inclusions of sets form a poset $(\mathcal{P}(S), \subseteq)$.

Example 1.6.4. Let $(P, \leq)$ be a poset. Let $S \subseteq P$ be a subset. Then $(S, \leq)$ is a poset.

Example 1.6.5. Let S be a set. Then equality forms a poset $(S, =)$.

Example 1.6.6. x is an ancestor of y is a poset.

Example 1.6.7. Any relation on the empty set $\emptyset$ is vacuously a poset.

Definition 1.6.8. Let $a, b \in \mathbb{N}$. We say a divides b and we write $a|b$ if there exists $q \in \mathbb{N}$ such that $b = aq$.

Proposition 1.6.9. The pair $(\mathbb{N}, |)$ is a poset.

PROOF. We shall prove the three axioms of a partial order.

- Since $a = 1 \cdot a$, we get $a|a$ for all $a \in \mathbb{N}$.
- Suppose $a|b$ and $b|a$. Then there exists $q, q' \in \mathbb{N}$ such that $b = aq$ and $a = bq'$. If $b = 0$, then as $a = bq'$ we get $a = 0$ and thus $a = b$. So suppose $b \neq 0$. Combining the two equations, we obtain:
$$b = aq = bq'q.$$
Therefore, if we divide by b on each side of the equation, we get $1 = q'q$. In $\mathbb{N}$, this can only happen for $q = 1$ and $q' = 1$. Thus $a = b$.
- Suppose $a|b$ and $b|c$. Then there exists $q, q' \in \mathbb{N}$ such that $b = aq$ and $c = bq'$. Therefore $c = aq'q$. Denote by $Q = q'q$. We have just proved that $c = aQ$. Therefore $a|c$. □

Definition 1.6.10. Let $(P, \leq)$ be a poset. We say the partial order $\leq$ to be a *total order in* P if in addition we have the following axiom:

- **(iv)** *totality or strong connectedness:* if $x, y \in P$, then $x \leq y$ or $y \leq x$.

In this case, we say the pair $(P, \leq)$ is a *totally ordered set* (sometimes called a *toset*).

Example 1.6.11. The poset $(\mathbb{N}, \leq)$ is totally ordered.

Example 1.6.12. The posets $(\mathcal{P}(S), \subseteq)$ and $(\mathbb{N}, |)$ are not totally ordered.

Example 1.6.13. A subset of a totally ordered set is also totally ordered.

Example 1.6.14. One can find totally ordered subsets in a poset that is not totally ordered. For instance consider the poset $(\mathbb{N}, |)$. Consider the subset of powers of two $S = \{2^n | n \in \mathbb{N}\} = \{1, 2, 4, 8, 16, \cdots\}$. Then $(S, |)$ is a totally ordered subset of $(\mathbb{N}, |)$.

Definition 1.6.15. Let $(P, \leq)$ be a poset. Let $S \subseteq P$ be a subset.

- An *upper bound of S in P* is an element $u_S \in P$ such that $s \leq u_S$ for any $s \in S$.
- A *maximal element of S* is an element $m_S \in S$ such that if $m_S \leq s$ for some $s \in S$, then $m_S = s$.

Lemma 1.6.16. Let $(P, \leq)$ be a poset. Let $S \subseteq P$ be a subset. An upper bound of S in S is a maximal element of S.

Proof. Suppose $u_S \in S$ is an upper bound of S. Then $s \leq u_S$ for all $s \in S$. Suppose $u_S \leq s$ for some $s \in S$. By antisymmetry, we get $u_S = s$. □

Example 1.6.17. Consider the poset $(S, =)$. Then every element of S is a maximal element, but S has no upper bounds in S.

Proposition 1.6.18. Let $(P, \leq)$ be a totally ordered set, and let $S \subseteq P$ be a subset. Then an element of S is an upper bound of S if and only if it is a maximal element.

Proof. By previous lemma, we are only left to show that a maximal element of S is also an upper bound of S in S. So suppose $m_S \in S$ is a maximal element of S. By reflexivity, we have $m_S \leq m_S$. Let $s \in S$ and $s \neq m_S$. Then since S is totally ordered, then $s \leq m_S$ or $m_S \leq s$. But if $m_S \leq s$, then $m_S = s$, which is not possible. Therefore $s \leq m_S$. Thus m_S is an upper bound. □

Theorem 1.6.19 (Zorn's lemma)**.** Let $(P, \leq)$ be a non-empty poset. Suppose that for every non-empty totally ordered subset $S \subseteq P$ there exists an upper bound of S in P. Then P has a maximal element.

Remark 1.6.20. Given the ZF axioms, the axiom of choice is equivalent to Zorn's lemma.

CHAPTER 2

Elementary arithmetic

2.1. Euclidean division

Theorem 2.1.1 (Euclidean division, division with remainder)**.** Let $a, d \in \mathbb{Z}$ be integers, with $d \geq 1$. Then there exist uniquely determined integers q and r such that

$$a = qd + r \quad \text{and} \quad 0 \leq r < d.$$

Proof. **Existence.** We first note that we can assume without loss of generality that $a \geq 0$. Indeed, we reduce the case $a < 0$ to $a' > 0$ by setting $a' = -a$, $q' = -q - 1$, $r' = d - r$. Then the desired equality $a = qd + r$ becomes $a' = q'd + r'$, and we have $0 \leq r' < d$.

Now for $a \geq 0$, we define $q_1 \coloneqq 0$ and $r_1 \coloneqq a$. Evidently, $a = q_1 d + r_1$. If $r_1 < d$, then the division is complete. Otherwise suppose $r_1 \geq d$ and define $q_2 \coloneqq q_1 + 1$ and $r_2 = r_1 - d$. We have $0 \leq r_2 < r_1$ and $a = q_2 d + r_2$. Now if $r_2 < d$, then the division is complete, otherwise repeat the process: set $q_{i+1} \coloneqq q_i + 1$ and $r_{i+1} = r_i - d$, observe $0 \leq r_{i+1} < r_i$ and $a = q_{i+1} d + r_{i+1}$, and check if $r_{i+1} < d$. Since we are decreasing r_i by a fixed number d at each step, the process cannot continue indefinitely while maintaining $r_i \geq 0$, and so there exists $k \in \mathbb{N}$ (which is in any case $k \leq a$) such that $a = q_k d + r_k$ and $0 \leq r_k < d$.

Uniqueness. Assume that we have two decompositions $a = qd + r$ and $a = q'd + r'$ such that $0 \leq r < d$ and $0 \leq r' < d$. To prove uniqueness, we need to show that $q = q'$ and $r = r'$.

The two equalities can be combined into $qd + r = q'd + r'$, and this is equivalent to $(q - q')d = r' - r$, so d divides $|r' - r|$. Since $0 \leq |r' - r| < d$, we conclude that $r = r'$, and therefore $(q - q')d = 0$. By assumption, $d \neq 0$, so $q = q'$ and we are done. □

Definition 2.1.2. Given the integers $a, d \in \mathbb{Z}$ with $d \geq 1$ and the result of their Euclidean division $a = qd + r$ with $0 \leq r < d$, we call q the *quotient* and r the *remainder* of the division of a by d. If $r = 0$, then d is a *divisor* of a and we write $d \mid a$. Otherwise we write $d \nmid a$.

Example 2.1.3. Notice how the sign changes the remainder when dividing by 5: $17 = 3 \cdot 5 + 2$ and $-17 = (-4) \cdot 5 + 3$.

Definition 2.1.4. A *common divisor* of two integers a and b is an integer $d \geq 1$ such that $d \mid a$ and $d \mid b$.

Definition 2.1.5. An integer $d \geq 1$ is called the *greatest common divisor* of a and b if:

(i) d is a common divisor of a and b;
(ii) if k is a common divisor of a and b, then $k \mid d$.

In this case we write $d = \gcd(a, b)$.

We will show shortly that the greatest common divisor exists for any pair of nonzero integers and that finding it can be done algorithmically, but first we will give a few examples.

Example 2.1.6. The divisors of 18 are 1, 2, 3, 6, 9, 18; the divisors of 30 are 1, 2, 3, 5, 6, 10, 15, 30; hence the common divisors of 18 and 30 are 1, 2, 3, 6. We can see that $6 = \gcd(18, 30)$.

Exercise 2.1.7. Write out the divisors of 6 and 7, then write all their common divisors and conclude that $\gcd(6,7) = 1$.

Exercise 2.1.8. Write out the divisors of -9 and 15, then write all their common divisors and conclude that $\gcd(-9,15) = 3$.

Example 2.1.9. Let $a \in \mathbb{Z} \setminus \{0\}$, then $\gcd(a,0) = |a|$.

Definition 2.1.10. A positive integer p is called *prime* if it has exactly two distinct divisors: 1 and p.

Example 2.1.11. If p and q are two distinct prime numbers, then $\gcd(p,q) = 1$.

Generalizing this example, we can give the following definition:

Definition 2.1.12. Two integers $a, b \in \mathbb{Z}$ are *relatively prime*, or *coprime*, if $\gcd(a,b) = 1$.

Theorem 2.1.13 (Bézout's identity)**.** Let $a, b \in \mathbb{Z} \setminus \{0\}$ be two nonzero integers. Then $\gcd(a,b)$ exists and can be written as

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

for some non-unique integers x and y.

Proof. Define the subset S of integers:

$$S \coloneqq \{xa + yb \mid x, y \in \mathbb{Z} \text{ and } xa + yb \geq 1\}.$$

It is clear that S is non-empty, so we can pick its smallest element and call it $d \coloneqq \min S$. We then have $d \geq 1$ and $d = xa + yb$. In order to complete the proof of the theorem, we need to show that $d = \gcd(a,b)$, and for this we will verify that $d \mid a$, $d \mid b$, and for any common divisor k of a and b, we have $k \mid d$.

We start with proving that $d \mid a$. By Theorem 2.1.1, we have

$$a = qd + r \quad \text{and} \quad 0 \leq r < d.$$

We can express r in terms of a, q and d:

$$\begin{aligned} r &= a - qd = a - q(xa + yb) \\ &= (1 - qx)a + (-qy)b. \end{aligned}$$

This shows that r is of the form $Xa + Yb$, however it cannot be in S since $r < d = \min S$, so the only option is $r = 0$ and hence $d \mid a$.

The same argument applies to showing that $d \mid b$.

Now assume that k is a common divisor of a and b, so $a = a'k$ and $b = b'k$, then $d = xa + yb = (xa' + yb')k$ and we see that $k \mid d$. □

In the remainder of this section, we will present main properties of greatest common divisor and use them to formulate Euclid's algorithm which will help us find x and y in expressing $\gcd(a,b)$ as $xa + yb$.

Exercise 2.1.14. Prove that $\gcd(a,b) = \gcd(-a,b) = \gcd(b,a)$.

2.2. Prime factorization

Lemma 2.2.1. Let $a, b \in \mathbb{Z}$ be relatively prime numbers, and let $k \in \mathbb{Z}$ be an arbitrary integer.

(i) If $a|k$ and $b|k$, then $(ab)|k$.

(ii) If $a|kb$, then $a|k$.

PROOF. **(i)** $\gcd(a, b) = 1 = ax + by$. By assumption, there are integers $r, s \in \mathbb{Z}$ such that $k = ra = sb$.

$$k = axk + byk = axsb + byra = ab(sx + ry).$$

(ii) $kb = ra$.

$$k = axk + ybk = axk + yra = a(xk + yr).$$

□

Recall: if $ab \in \mathbb{Z}$ is even, then at least one of a or b is even.

Lemma 2.2.2 (Euclid's Lemma)**.** Let p be a prime.

(i) Given $a, b \in \mathbb{Z}$, if $p|(ab)$, then $p|a$ or $p|b$.

(ii) Given $a_1, \ldots, a_r \in \mathbb{Z}$, if $p|(a_1 \cdots a_r)$, then for some index $i \in \{1, \ldots, r\}$ we have that $p|a_i$.

Remark 2.2.3. Euclid's Lemma only applies to prime numbers. For example, 6 divides $12 = 3 \cdot 4$, however 6 doesn't divide neither 3, nor 4.

Theorem 2.2.4 (Prime factorization)**.** Every integer $a \geq 1$ is the product of prime numbers, and that factorization is unique up to the order of the factors:

$$\text{if } a = p_1 \cdots p_r \text{ and } a = q_1 \cdots q_s,$$

where the p_i's and q_j's are prime, then $r = s$, and the q_j's can be relabeled in such a way that $p_i = q_i$ for all $i = 1, \ldots r$.

Theorem 2.2.5. Let $a \geq 2$ be an integer with prime factorization

$$a = p_1^{n_1} \cdots p_r^{n_r}.$$

Then any divisor d of a is of the form

$$d = p_1^{d_1} \cdots p_r^{d_r},$$

for some choice of $0 \leq d_i \leq n_i$.

Theorem 2.2.6 (Euclid's theorem)**.** There are infinitely many prime numbers.

2.3. Modular arithmetic

Previously: for $a \in \mathbb{Z}$ and $d \geq 1$, we proved that we can find unique q and r such that $a = qd + r$ and $0 \leq r < d$.

2.3.1. Congruence modulo n.

Definition 2.3.1. Two integers $a, b \in \mathbb{Z}$ are said to be *congruent modulo n* if they have the same remainder with respect to their division by $n \geq 2$. In this case, we write

$$a \equiv b \mod n.$$

Example 2.3.2. $2 \equiv 5 \equiv 14 \equiv -1 \mod 3$.

Example 2.3.3. $21 \equiv 16 \equiv -4 \mod 5$.

Lemma 2.3.4. Two numbers $a, b \in \mathbb{Z}$ are congruent modulo n if and only if n divides their difference:

$$a \equiv b \mod n \iff n|(a-b).$$

Proof. Forward direction. We can divide both a and b by n with remainder, and since $a \equiv b \mod n$, we get

$$a = qn + r \text{ and } b = q'n + r$$

for a unique choice of a triple $q, q', r \in \mathbb{Z}$, with $0 \leq r < n$. So we get $a - b = (q - q')n$ is divisible by n. Conversely, say $n|(a-b)$, then there is $k \in \mathbb{Z}$ such that $a - b = kn$. Apply Euclidean division again, and notice that we have uniquely:

$$a = qn + r \text{ and } b = q'n + r',$$

with $0 \leq r, r' < n$. Hence

$$a - b = (q - q')n + (r - r'),$$

and $0 \leq |r - r'| < n$. By uniqueness of Euclidean division, this can happen only if $r = r'$. □

Corollary 2.3.5. An integer is congruent to 0 modulo n if and only if it is divisible by n:

$$a \equiv 0 \mod n \iff n|a.$$

Theorem 2.3.6. Congruence modulo n on $\mathbb{Z}$ is an equivalence relation.

Proof. □

Thus it gives a partition on $\mathbb{Z}$. There are several common choices of notation for the corresponding equivalence classes.

Definition 2.3.7. Given $a \in \mathbb{Z}$, we write $[a]$, $[a]_n$, or $\overline{a}$ for the equivalence class with respect to congruence modulo n, and we call it the *residue class of a modulo n*:

$$[a] = [a]_n = \overline{a} = \{x \in \mathbb{Z} \mid a \equiv x \mod n\}.$$

The quotient set will be denoted by $\mathbb{Z}_n$, $\mathbb{Z}/n\mathbb{Z}$, or $\mathbb{Z}/n$.

We have $[a] = [b] \iff a \equiv b \mod n$.

Theorem 2.3.8. Let us fix an integer $a \in \mathbb{Z}$ and an integer $n \geq 2$.

(i) $[a] = [r]$ for some $0 \leq r < n$.

(ii) $[0], [1], \ldots, [n-1]$ are all distinct.

PROOF. **(i)** Euclidean division.
(ii) Suppose $[r] = [s]$ for $0 \leq r \leq s < n$, conclude $r = s$. □

In other words, we have proved that the quotient set $\mathbb{Z}/n$ has n elements, and described those:

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

2.3.2. Arithmetic modulo n.

Theorem 2.3.9. The set $\mathbb{Z}/n$ inherits the operations of addition and multiplication in $\mathbb{Z}$.

PROOF. We need to check that these operations are well-defined. □

So we have two binary operations:

$$+ : \mathbb{Z}/n \times \mathbb{Z}/n \to \mathbb{Z}/n,$$
$$\cdot : \mathbb{Z}/n \times \mathbb{Z}/n \to \mathbb{Z}/n.$$

Example 2.3.10. For $n = 6$, we can compute the following: $[3] + [5] = [8] = [2]$, $[3] \cdot [5] = [15] = [3]$, $[2] \cdot [3] = [6] = [0]$.

Exercise 2.3.11. January 27th in 2022 is a Thursday, and 2022 and 2023 are not leap years. What day of the week is January 27th in 2023? In 2024?

Proposition 2.3.12. Fix an integer $n \geq 2$. Then in $\mathbb{Z}/n$, for any three arbitrary integers $a, b, c \in \mathbb{Z}$, we have:

- **(i)** Commutativity of addition and multiplication.
- **(ii)** Associativity of addition and multiplication.
- **(iii)** Neutrality of $[0]$ and unitality of $[1]$. So $[0]$ is the zero of addition, and $[1]$ is the multiplicative identity.
- **(iv)** Opposability: $-[a] = [-a]$.
- **(v)** Distributivity.

Example 2.3.13. Multiplication table for $\mathbb{Z}/6$.
Strange behavior:

- In $\mathbb{Z}$, we have $k^2 = k \implies k = 0$ or $k = 1$. In $\mathbb{Z}/6$, it can be $k = [0], [1], [3], [4]$.
- In $\mathbb{Z}$, we have that $ab = ac$ with $a \neq 0$ implies $b = c$. However, $[4][2] = [4][5]$ and $[4] \neq [0]$, yet $[2] \neq [5]$.
- In $\mathbb{Z}$, if $ab = 0$, then $a = 0$ or $b = 0$. In $\mathbb{Z}/6$, we have $[2][3] = [0]$.

Exercise 2.3.14. In $\mathbb{Z}/12$ we can have $k^2 = [0]$ with $k \neq 0$, which is even stranger! Find a k like this.

Example 2.3.15. This example shows that arithmetic in $\mathbb{Z}/n$ can be really powerful. We will compute the remainder of 4^{119} by 7. Note: $[4]^3 = [1]$. So $[4]^{119} = [4]^{3 \cdot 39+2} = [4]^2 = [2]$.

2.3.3. Invertibility.

Definition 2.3.16. A residue class $[b] \in \mathbb{Z}/n$ is an *inverse* of $[a]$ if $[b][a] = [1]$. We then write $[b] = [a]^{-1}$. If an inverse of $[a]$ exists, we say that $[a]$ is *invertible*.

Lemma 2.3.17. If an inverse exists, it is unique.

Theorem 2.3.18. $[a]$ is invertible in $\mathbb{Z}/n$ if and only if $\gcd(a, n) = 1$.

Example 2.3.19. Find the inverse of $[16] \in \mathbb{Z}/35$. Indeed, $[16]$ is invertible modulo 35, since $\gcd(16, 35) = \gcd(2^4, 5 \cdot 7) = 1$. Now use the Euclidean algorithm:

- $35 = 2 \cdot 16 + 3$,
- $16 = 5 \cdot 3 + 1$.

We can now see that

$$1 = 16 - 5 \cdot 3 = 16 - 5 \cdot (35 - 2 \cdot 16) = 11 \cdot 16 - 5 \cdot 35,$$

which modulo 35 gives us the equality:

$$[11][16] = [1].$$

Example 2.3.20. Solve $[16]x = [9]$ in $\mathbb{Z}/35$. We have just calculated that $[16]^{-1} = [11]$, so $x = [16]^{-1} \cdot [9] = [11][9] = [99] = [29]$.

Example 2.3.21. Solving quadratic equation

$$x^2 + [3]x + [9] = 0$$

in $\mathbb{Z}/13$. Recall that over $\mathbb{Q}$, we can write $x^2 + 3x = x^2 + 3x + (\frac{1}{2} \cdot 3)^2 - (\frac{1}{2} \cdot 3)^2 = (x + \frac{1}{2} \cdot 3)^2 - (\frac{1}{2} \cdot 3)^2$. Since $\gcd(2, 13) = 1$, the residue class $[2]$ is invertible in $\mathbb{Z}/13$, and we can calculate that $[2]^{-1} = [7]$. Hence we can rewrite the original equation as follows:

$$(x + [2]^{-1}[3])^2 - ([2]^{-1}[3])^2 + [9] = 0,$$

$$(x + [7][3])^2 = ([7][3])^2 - [9] = [21]^2 - [9] = [-5]^2 - [9] = [25] - [9] = [16] = [3],$$

$$(x + [8])^2 = [3].$$

The equation $y^2 = [3]$ has two solutions in $\mathbb{Z}/13$, namely $y = \pm[4] = [4], [9]$. Therefore, $x = y - [8] = [9], [1]$.

Example 2.3.22. Note that the equation $y^2 = [3]$ has two solutions in $\mathbb{Z}/13$; however it doesn't have solutions in $\mathbb{Z}/7$.

Example 2.3.23. The equation $y^2 = [9]$ has two solutions in $\mathbb{Z}/13$ and in $\mathbb{Z}/7$, which are $[3]$ and $[6]$; however it has six solutions in $\mathbb{Z}/27$, namely $[3]$, $[6]$, $[12]$, $[15]$, $[21]$, $[24]$.

Theorem 2.3.24 (Chinese remainder theorem)**.** Let $m, n \geq 2$ be two coprime integers. If $a, b \in \mathbb{Z}$ are arbitrary integers, then we can find $x \in \mathbb{Z}$ such that

$$x \equiv a \mod m \quad \text{and} \quad x \equiv b \mod n.$$

Proof. Since m and n are coprime, we have some $r, s \in \mathbb{Z}$ such that $1 = rm + sn$. Define

$$x = rmb + sna.$$

Now subtract a from x:

$$x - a = rmb + (sn - 1)a = rmb - rma = m \cdot r(b - a).$$

We can now see that $m|(x-a)$, so $x \equiv a \mod m$. A similar argument proves that $x \equiv b \mod n$. □

Theorem 2.3.25. For an integer $n \geq 2$, we have that the following are equivalent:

(i) Every nonzero $[a]$ in $\mathbb{Z}/n$ is invertible;
(ii) If $[a][b] = [0]$ in $\mathbb{Z}/n$, then $[a] = [0]$ or $[b] = [0]$;
(iii) The integer n is prime.

PROOF. (1) $\Rightarrow$ (2): Given $[a][b] = [0]$, either $[a] = [0]$ and we are done, or $[a] \neq [0]$, and then $[a]$ is invertible. Therefore we can get:

$$[b] = [a]^{-1}[a][b] = [a]^{-1}[0] = [0].$$

(2) $\Rightarrow$ (3): Write n as a product of two positive numbers $n = ab$, we want to argue that $a = 1$ or $b = 1$ proving that there are exactly two positive divisors of n. Taking the equality $n = ab$ modulo n, we get $[0] = [n] = [a][b]$. By assumption, then either $[a] = [0]$ or $[b] = [0]$. It follows that one of a, b should be n, so the other one whould be 1.

(3) $\Rightarrow$ (1): Now n is prime, and we study $[a] \in \mathbb{Z}/n$ which is nonzero. It follows that $\gcd(a, n) = 1$, so $[a]$ is invertible. □

Remark 2.3.26. When p is a prime, the set $\mathbb{Z}/p$ together with addition and multiplication is called a *finite field*.

Theorem 2.3.27 (Fermat's little theorem)**.** Let p be a prime and $a \in \mathbb{Z}$. Then

$$a^p \equiv a \mod p.$$

If $a \neq 0$, then it can also be written as $[a]^{p-1} = [1]$.

PROOF. It is enough to prove the theorem for $a = 0, 1, \ldots, p-1$. We will induct on a starting with $a = 0$, in which case the equality is trivial:

$$[0]^p = [0].$$

Now assume that we proved $[a]^p = [a]$, and we want to prove that $[a+1]^p = [a+1]$. We will use the formula from homework that for a prime p, we have an equality in $\mathbb{Z}/p$:

$$([a] + [b])^p = [a]^p + [b]^p.$$

Setting $[b] = [1]$, we get:

$$[a+1]^p = ([a] + [1])^p = [a]^p + [1] = [a] + [1] = [a+1].$$

This concludes the step of induction and the proof. □

This theorem also has a one-line proof using methods of *group theory*, namely Lagrange's theorem. We will return to it later.

CHAPTER 3

Groups

3.1. Binary operations

We want to generalize the notions of addition and multiplication. These two operations each take two inputs and produce one output. Rules like this are called *binary operations*.

Definition 3.1.1. A *binary operation* $*$ on a set X is a function

$$*\colon X \times X \to X,$$

which takes a pair (x, y) to the element that we will denote by $x * y$ or xy. Denote by $(X, *)$ the set X endowed with the operation $*$.

Example 3.1.2. **(i)** Usual addition or multiplication on $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$.

(ii) $M(\mathbb{R})$ – set of all matrices of all size. Then addition and multiplication are NOT binary operations.

Definition 3.1.3. A binary operation $* : X \times X \to X$ is said to be:

- *associative* if $\forall x, y, z \in X$, we have: $(x * y) * z = x * (y * z)$;
- *unital* if there exists an element $e \in X$ such that $\forall x \in X$ we have $e * x = x$ and $x * e = x$ — in this case, we say e is the *identity* of the binary operation;
- *commutative* if $\forall x, y \in X$, we have: $x * y = y * x$.

Exercise 3.1.4. Show that an identity e of a binary operation $* : X \times X \to X$ is necessarily unique: if there exists $e' \in X$ such that $\forall x \in X$ we have $e' * x = x = x * e'$, then $e = e'$. *Hint: what is $e * e'$?*

Example 3.1.5. Consider $(\mathbb{N}, *)$ with $a * b = \min(a, b)$. Then $*$ is commutative and associative, but not unital. However, if we take $X = \mathbb{N} \cup \{+\infty\}$, then taking minimum will also be unital with $e = +\infty$.

Example 3.1.6. Consider $(\mathbb{N}, *)$ with $a * b = a$. Then $*$ is associative, but neither commutative nor unital.

Remark 3.1.7. It may be helpful to view compositions of maps as a diagram, to see better where each set map goes. When the composition of set maps on the diagram is the same regardless of the path we choose, this diagram is said to *commute*, and is called a *commutative diagram*. With this, we can illustrate axioms for associativity, unitality and commutativity as follows.

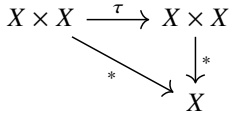
Associativity: The axiom of associativity $(x * y) * z = x * (y * z)$ is equivalent to the commutativity of the diagram:

$$\begin{array}{ccc} X \times X \times X & \xrightarrow{*\times \mathrm{id}_X} & X \times X \\ {\scriptstyle \mathrm{id}_X \times *}\downarrow & & \downarrow{\scriptstyle *} \\ X \times X & \xrightarrow[*]{} & X, \end{array}$$

Unitality: Given $e \in X$, define the map $e_1 : X \to X \times X$ by $e_1(x) = (e, x)$ and the map $e_2 : X \to X \times X$ by $e_2(x) = (x, e)$. The axiom of unitality $e * x = x * e$ is equivalent to the commutativity of the diagram:

$$\begin{array}{ccc} X & \xrightarrow{e_1} & X \times X \\ & \searrow & \downarrow{\scriptstyle *} \\ & & X \end{array} \qquad \begin{array}{ccc} X & \xrightarrow{e_2} & X \times X \\ & \searrow & \downarrow{\scriptstyle *} \\ & & X \end{array}$$

Commutativity: Define $\tau : X \times X \to X \times X$ by $\tau(x, y) = (y, x)$. The axiom of commutativity is equivalent to the commutativity of the diagram:



Exercise 3.1.8. Show that an element x_0 in a set X is equivalent to the data of a map $\{\star\} \to X$. Conclude that the map $e_1 : X \to X \times X$ can be decomposed as a map $X \cong X \times \{\star\} \to X \times X$.

Theorem 3.1.9. Given an associative binary operation $(X, *)$, we can write unambiguously

$$a_1 * a_2 * \cdots * a_n$$

for all $n \geq 3$.

PROOF. Induction on n. □

Proposition 3.1.10. If $(X, *)$ is unital, then the identity is unique.

PROOF. Let $e, e' \in X$ such that $e * x = x = x * e$ and $e' * x = x = x * e'$ for all $x \in X$. Then:

$$e = e * e' = e'$$

□

3.2. Monoids

Definition 3.2.1. A *monoid* $(M, *, e)$ is an associative binary operation $(M, *)$ that admits an identity e. We say that it is a *commutative monoid* if $(M, *)$ is commutative.

Example 3.2.2. **(i)** $(M_n(\mathbb{R}), +, 0)$ – commutative monoid.
(ii) $(M_n(\mathbb{R}), \cdot, I)$ – noncommutative monoid.

Example 3.2.3. Let S be any set, then $(2^S, \cup, \varnothing)$ and $(2^S, \cap, S)$ are commutative monoids.

For a finite set, a binary operation can be defined by means of *Cayley table* – like the multiplication table we had for $\mathbb{Z}/n$.

When the monoid comes from multiplication on some set, we write $(M, \cdot, 1)$ – *multiplicative notation*. When it come from addition on some set, we write $(M, +, 0)$ – *additive notation*. All monoids written in additive notation are assumed to be commutative; multiplicative notation can be applied to both commutative and noncommutative settings.

Definition 3.2.4. Let $(M, \cdot, 1)$ be a monoid. If $n \in \mathbb{N}$, define a^n as follows:

- $a^0 = 1$;
- $a^n = a \cdot a^{n-1}$ for $n \geq 1$.

For example, $a^1 = a$, $a^2 = aa$, and so on. With additive notation, we would write na instead, so $2a = a + a$.

Theorem 3.2.5. Given a monoid $(M, \cdot, 1)$, consider $a, b \in M$ and $n, m \in \mathbb{N}$. Then:
(i) $a^n a^m = a^{n+m}$;
(ii) $(a^n)^m = a^{nm}$;
(iii) if $ab = ba$, then $(ab)^n = a^n b^n$.

Proof. Induction. □

Theorem 3.2.6. If $(X, *)$ admits an identity, then the identity is unique.

Definition 3.2.7. Let $(M, \cdot, 1)$ be a monoid. An element $a \in M$ *admits an inverse* if there exists $b \in M$ such that $ab = 1 = ba$. If such an element exists, then a is said to be *invertible* in, or a *unit* of M. The element b is then called the *inverse* of a and is denoted by a^{-1}. Given a monoid, we will denote the subset of all units by $M^\times$.

In additive notation, the inverse is denoted by $-a$, so $a + (-a) = 0 = (-a) + a$, and the additive inverse is often called the *opposite* element to a.

Theorem 3.2.8. Let $(M, \cdot, 1)$ be a monoid. If $a \in M^\times$ is invertible, then its inverse is unique.

Theorem 3.2.9. Let $(M, \cdot, 1)$ be a monoid.
(i) 1 is a unit, and $1^{-1} = 1$.
(ii) If a is a unit, then $(a^{-1})^{-1} = a$.
(iii) If $a, b \in M^\times$, then $ab \in M^\times$, and $(ab)^{-1} = b^{-1}a^{-1}$.
(iv) If $a \in M^\times$, then $a^n \in M^\times$, and $(a^n)^{-1} = (a^{-1})^n$.

3.3. Groups

Definition 3.3.1. A *group* $(G, *, e)$ is a monoid in which every element is a unit. We will often say "a group G", suppressing the notation for the group operation and the identity, when this is clear from the context.

In other words, a group is defined by:

- a binary operation $* : G \times G \to G$,
- an element $e \in G$,

such that:

- $*$ is associative;
- e is the identity of $*$, that is $\forall g \in G : g * e = g = e * g$;
- every element of G is invertible.

Definition 3.3.2. A group $(G, *, e)$ is said to be *Abelian* if it is commutative as a monoid.

Definition 3.3.3. The *order* of a finite group G is its cardinality, and it is denoted by $|G|$.

When the group is infinite, we can say that it has infinite order, but this notion doesn't play a role in various theories of infinite groups.

Example 3.3.4. $\{1\}$ is a group – the *trivial group*.

Example 3.3.5.
(i) $(\mathbb{Z}, +, 0)$, $(\mathbb{Q}, +, 0)$, $(\mathbb{R}, +, 0)$, $(\mathbb{C}, +, 0)$ are all Abelian groups.
(ii) $(\mathbb{N}, +, 0)$ is a commutative monoid, but not a group, because none of the nonzero elements have the opposite.

Example 3.3.6. $\{1, -1\} \subset (\mathbb{R}\backslash\{0\}, \cdot, 1)$ is a group. Its multiplication table is similar (up to relabeling) to that of $(\mathbb{Z}/2, +, [0])$. We will later say that these groups are isomorphic.

Example 3.3.7. If V is a vector space, then $(V, +, 0)$ is an Abelian group.

Example 3.3.8. $(S^1, \cdot, 1) \subset (\mathbb{C}, \cdot, 1)$ is an Abelian group – the *group circle*.

$$S^1 = \{z \in \mathbb{C} : |z| = 1\} = \{e^{i\phi} : \phi \in \mathbb{R}\}.$$

Example 3.3.9. $(\mathbb{Z}/n, +, [0])$ is an abelian group.

Theorem 3.3.10. If $(M, \cdot, 1)$ is a monoid, then $(M^\times, \cdot, 1)$ is a group.

Proof. By Theorem 3.2.9, $M^\times$ is closed under multiplication and taking inverses. Moreover, since M is a monoid, the operation on $M^\times$ is associative and unital, and since we only took units, every element in $M^\times$ is invertible. □

Example 3.3.11 (Symmetric groups)**.** Let X be any set, and consider $M_X = \operatorname{Hom} X, X$. Then $(M_X, \circ, \mathrm{id}_X)$ is a monoid. The set $M_X^\times$ is the group of invertible elements in M_X – bijections $M_X \to M_X$. When $X = \{1, \ldots, n\}$, we write $\mathfrak{S}_n = (M_X^\times, \circ, \mathrm{id}_X)$, and we call it the *symmetric group on n elements*. If we think of $\mathfrak{S}_n$ as permutations of n elements, we can combinatorially conclude that the order $|\mathfrak{S}_n|$ is $n!$.

Theorem 3.3.12. If $(G, *, e_G)$ and $(H, \star, e_H)$ are groups, then $(G \times H, * \times \star, (e_G, e_H))$ is also a group.

$$\begin{aligned}(G \times H) \times (G \times H) &\xrightarrow{*\times\star} G \times H \\ ((g,h),(g',h')) &\mapsto (g * g', h \star h')\end{aligned}$$

Example 3.3.13. $\mathbb{Z}/2 \times \mathbb{Z}/2 = \{(0,0),(1,0),(0,1),(1,1)\}$

Theorem 3.3.14. Let G be a group and $g, h, f \in G$. Then:

(i) $gh = gf \implies h = f$,
(ii) $hg = fg \implies h = f$.

Theorem 3.3.15. If G is a finite group, then for any $g \in G$, for some $n \geq 1$, we have $g^n = e$.

Definition 3.3.16. Given $n \geq 1$, the *cyclic group of order* n is $C_n = \{1, a, a^2, \ldots, a^{n-1}\}$ with operation $a^r a^s = a^{r+s}$ such that $a^n = a^0 = 1$.

Example 3.3.17. We can write the cyclic group of order n in another form $\mu_n = \{e^{2k\pi i/n} \mid k = 0, \ldots, n-1\} \subset \mathbb{C}\backslash\{0\}$. Sometimes $e^{2\pi i/n}$ is denoted by ω, and elements of μ_n are called *nth roots of unity.*

3.4. Subgroups

Definition 3.4.1. Let $(G, *, e)$ be a group. A subset $H \subseteq G$ is called a *subgroup* of G, and denoted by $H \leq G$, if:

(i) H is closed under $*$, that is for any $h, h' \in H$, we also have $h * h' \in H$;
(ii) $(H, *)$ is a group.

Example 3.4.2. $\mu_n \leq S^1 \leq \mathbb{C} \setminus \{0\}$.

Example 3.4.3. $\mathbb{Q} \setminus \{0\} \subset \mathbb{R}$ is a subset, but not a subgroup of $(\mathbb{R}, +)$.

Definition 3.4.4. Let G be a group. The *trivial subgroup* is $\{e\} \leq G$.

Example 3.4.5. $G \leq G$.

Definition 3.4.6. A *proper subgroup* of G is a subgroup $H \leq G$ such that $H \neq G$. We write it as $H < G$.

Example 3.4.7. $\mu_n < S^1 < \mathbb{C} \setminus \{0\}$.

Example 3.4.8. $\mathbb{Z} < \mathbb{Q} < \mathbb{R} < \mathbb{C}$.

Example 3.4.9. Subgroups of $\mathbb{Z}/4$ and $\mathbb{Z}/2 \times \mathbb{Z}/2$.

Proposition 3.4.10 (Subgroup criterion)**.** Let $(G, *, e)$ be a group. A nonempty subset $H \subseteq G$ is a subgroup of G if and only if the following three conditions hold:

(i) H is closed under $*$;
(ii) $e \in H$;
(iii) if $h \in H$, then $h^{-1} \in H$.

Example 3.4.11. Let $n \in \mathbb{N}$. Define

$$n\mathbb{Z} = \{nk \mid k \in \mathbb{Z}\}.$$

Then $n\mathbb{Z} \leq \mathbb{Z}$. One can check it using the subgroup criterion.

In fact, any subgroup of $\mathbb{Z}$ is of the form $n\mathbb{Z}$ for some $n \in \mathbb{N}$ – we won't prove this fact just yet.

Proposition 3.4.12 (Finite subgroup criterion)**.** Let $(G, *, e)$ be a group. A nonempty finite subset $H \subseteq G$ is a subgroup of G if and only if H is closed under $*$.

Example 3.4.13. Take $G = (\mathbb{Z}/15, \cdot)^\times$. We know that the units in the monoid $(\mathbb{Z}/15, \cdot)^\times$ are exactly those elements that are coprime with 15, so as a set $G = \{1, 2, 4, 7, 8, 11, 13, 14\}$. We will use the finite subgroup criterion to check when the following examples are subgroups of G.

- $H = \{1, 7\}$ is not a subgroup because $7^2 = 4 \notin H$.
- $H' = \{1, 4, 7\}$ is not a subgroup because $7 \cdot 4 = 13 \notin H'$.
- $H'' = \{1, 4, 7, 13\}$ is a subgroup because $7 \cdot 13 = 1 \in H''$, and all elements are powers of 7.

Notice that in the previous example we constructed the smallest subgroup of $\mathbb{Z}/15$ that contains 7.

3.5. Generators and order of elements

3.5.1. Cyclic groups revisited.

Definition 3.5.1. Let G be a group. Define the *cyclic subgroup generated by* g:

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

Exercise 3.5.2 (Easy)**.** Check that $\langle g \rangle \leq G$ and that $\langle g \rangle$ is Abelian.

Example 3.5.3. Let $(G, *, e)$ be a group, then $\langle e \rangle = \{e\}$.

Example 3.5.4. $n\mathbb{Z} = \langle n \rangle \leq \mathbb{Z}$.

Definition 3.5.5. If in a group G, there is an element $g \in G$ such that $G = \langle g \rangle$, then G is a *cyclic group generated by* g, and g is called a *generator* of G.

Exercise 3.5.6. Let $G = (\mathbb{Z}/n, +)$. Show that an element a is a generator if and only if $\gcd(a, n) = 1$.

Example 3.5.7. $\mathbb{Z} = \langle 1 \rangle = \langle -1 \rangle$.

Definition 3.5.8. Let G be a group. The *order of an element* $g \in G$ is the smallest integer $n \geq 1$ such that $g^n = e$, in which case we write $|g| = n$ or $\operatorname{ord} g = n$. If there is no such integer $n \geq 1$, then we say that g has *infinite order* and write $|g| = \operatorname{ord} g = \infty$.

Example 3.5.9. In $\mathbb{Z}$, $|1| = \infty$. In $\mathbb{Z}/n\mathbb{Z}$, $|1| = n$. In $(\mathbb{Z}/15, \cdot)^\times$, $|7| = 4$. In any group $(G, *, e)$, $|e| = 1$.

Proposition 3.5.10. For any element g of a group G, we have $|g| = |g^{-1}|$.

Proposition 3.5.11. If G is a finite group, then any element $g \in G$ has finite order.

How can we guess the order of an element in general?

Theorem 3.5.12. Let $(G, *, e)$ be a group and $g \in G$ be an element of finite order $|g| = n$. Then:

(i) $g^k = e \iff n|k$;
(ii) $g^k = g^m \iff k \equiv m \mod n$;
(iii) $\langle g \rangle = \{e, g, g^2, \ldots, g^{n-1}\}$.

Proof. For the forward direction in (1), use Euclidean division. Suppose $g^k = e$, and divide k by n with remainder:

$$k = qn + r, \quad 0 \leq r < n.$$

Then:

$$g^r = e^q * g^r = g^{nq} * g^r = g^{qn+r} = g^k = e.$$

Since $g^r = e$ and $0 \leq r < n = |g|$, the only possibility is that $r = 0$, and so $n|k$.

For part (2), observe that $g^k = g^m$ if and only if $g^{k-m} = e$, and apply part (1).

For part (3), first check that $\langle g \rangle \subset \{e, g, g^2, \ldots, g^{n-1}\}$ and then observe that all of the elements $e, g, g^2, \ldots, g^{n-1}$ are distinct. □

Theorem 3.5.13. Let $(G, *, e)$ be a group and $g \in G$ be an element of infinite order. Then:

(i) $g^k = e \iff k = 0$;

(ii) $g^k = g^m \iff k = m$;
(iii) $\langle g \rangle = \{g^k \mid k \in \mathbb{Z}\}$.

Corollary 3.5.14. Let $(G, *, e)$ be a group and $g \in G$. Then $|g| = |\langle g \rangle|$.

Corollary 3.5.15. Let $(G, *, e)$ be a group and $g \in G$ of finite order $|g| = n$. If $d|n$, for some $d \geq 1$, then $|g^d| = \frac{n}{d}$.

PROOF. Let $k = \frac{n}{d}$ and $h = g^d$. We want to show that $|h| = k$.

First notice that $h^k = e$. But is it the smallest? Suppose we have $h^r = e$ for some $r \geq 1$, so $g^{dr} = h^r = e$, and by Theorem 3.5.12 we have $n|dr$, in particular $dr \geq n = dk$, and therefore $r \geq k$, so k is indeed the minimum. □

Corollary 3.5.16 (Important!)**.** Every subgroup of a cyclic group G is cyclic. (Note that G is not necessarily finite.)

PROOF. □

Corollary 3.5.17. In particular, every subgroup of $\mathbb{Z}$ is of the form $n\mathbb{Z}$ for some $n \in \mathbb{N}$.

Example 3.5.18. Find all subgroups of $\mathbb{Z}$ containing $187\mathbb{Z}$. Notice that $187 = 11 \cdot 17$. If $187\mathbb{Z} \leq n\mathbb{Z}$, then $187 \in n\mathbb{Z}$, and so $n|187$. We can have $n = 1, 11, 17, 187$.

Theorem 3.5.19. Let G be a cyclic group of finite order n, $G = \langle g \rangle$. Then $G = \langle g^k \rangle$ if and only if $\gcd(k, n) = 1$.

Example 3.5.20. Generators of $\mathbb{Z}/12$ are $1, 5, 7, 11$.

Corollary 3.5.21. Let $G = \langle g \rangle$ be a cyclic group of order n. Then $\langle g^a \rangle = \langle g^b \rangle$ if and only if $\gcd(a, n) = \gcd(b, n)$.

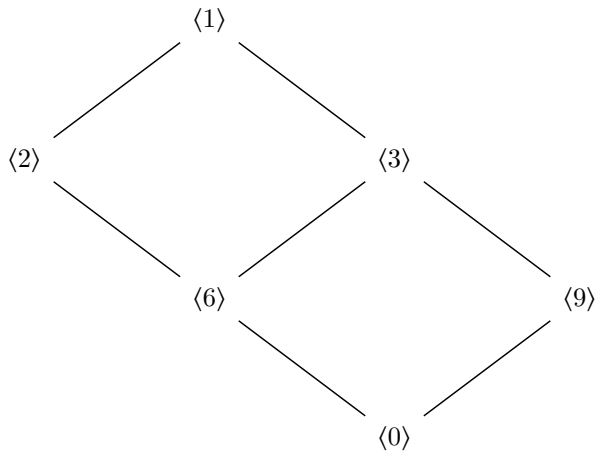
Theorem 3.5.22 (Important!)**.** Let $G = \langle g \rangle$ be a cyclic group of order n.

(i) If $H \leq G$, then $H = \langle g^d \rangle$ for some nonnegative $d|n$. In particular, $|H|$ divides $|G|$.
(ii) If $k|n$, then $\langle g^{\frac{n}{k}} \rangle$ is the unique subgroup of G of order k.

PROOF. **(i)** Since $H \leq G$, we have $H = \langle g^m \rangle$ for some $m \geq 1$ by Corollary 3.5.16. Let $d = \gcd(m, n)$. Then $\langle g^m \rangle = \langle g^d \rangle$ by Corollary 3.5.21.
(ii) Let $K \leq G$ be some subgroup of order $|K| = k$. By part 1, it is of the form $K = \langle g^d \rangle$ for $d|n$. But $|g^d| = \frac{n}{d} = |K|$, so $d = \frac{n}{k}$.

□

Example 3.5.23. Consider $G = \mathbb{Z}/18$. We can use the previous theorem to draw all the subgroups of this group.



3.5.2. Generators. Not all groups are cyclic, for example $\mathbb{Z}/5 \times \mathbb{Z}/5$ is not. However, we can still find two elements $(1,0)$ and $(0,1)$ whose linear combinations give the whole group. The choice of these elements is not unique though: we could have chosen $(4,1)$ and $(0,1)$, or $(4,4)$ and $(1,0)$.

Definition 3.5.24. Let $S \neq \varnothing$ be a nonempty subset of a group G. Define

$$\langle S \rangle = \{x_1^{k_1} \cdots x_m^{k_m} \mid x_i \in S, k_i \in \mathbb{Z}, m \geq 1, \text{the } x_i \text{ may repeat}\} \subset G.$$

The subset $\langle S \rangle$ is called *the subgroup of G generated by S*.

- If $S = \{g\}$, then $\langle S \rangle = \langle g \rangle$ as before.
- Similarly, if $S = \{g_1, \ldots, g_n\}$, then we will denote $\langle S \rangle$ by $\langle g_1, \ldots g_n \rangle$, omitting the curly braces.

Theorem 3.5.25. Let G be a group and $S \subset G$ a subset of G. Then:

(i) $\langle S \rangle$ is a subgroup of G and contains the set S;
(ii) if $H \leq G$ such that $S \subset H$, then $\langle S \rangle \leq H$.

Proof. **(i)** Use the subgroup criterion (Proposition 3.4.10): check that $\langle S \rangle$ is closed under the group operation, taking inverses and contains the identity element e.
(ii) Direct check.

□

Definition 3.5.26. If $G = \langle S \rangle$ for a finite subset $S \subset G$, then we say that G is *finitely generated*.

Example 3.5.27. Of course, if G is finite, then it is finitely generated, because $G = \langle G \rangle$.

Example 3.5.28. $\mathbb{Z} = \langle 1 \rangle$ is finitely generated.

Example 3.5.29. $\mathbb{Z} \times \mathbb{Z} = \langle (1,0), (0,1) \rangle$ is finitely generated.

Example 3.5.30. $(\mathbb{Q}, +, 0)$ is not finitely generated.

3.6. Group homomorphisms

3.6.1. Definitions. Suppose that we have two groups G and H, and now we want to see in what way they can "interact". In general, if we have a set map $f : G \to H$, we have no reason to expect that it would be compatible with the group operations on G and H, and this is the very least that we would like to expect from a "suitable" map between the groups.

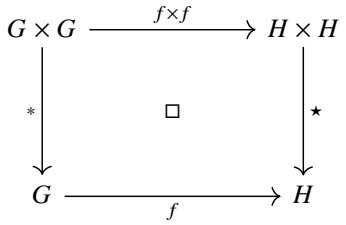
Definition 3.6.1. Let $(G, *, e_G)$ and $(H, \star, e_H)$ be two groups. We say that a set map

$$f : G \to H$$

is a *group homomorphism* (or *homomorphism of groups,* or just *homomorphism*) if

$$\forall a, b \in G : f(a * b) = f(a) \star f(b).$$

In other words, the following diagram should commute:



Remark 3.6.2. We denote the set of all group homomorphisms by $\operatorname{Hom}\mathsf{Grp}G, H$, or by $\operatorname{Hom} G, H$ when it is clear that we are only looking at group homomorphisms and not just set maps. When we want to underline that we are looking at set maps that don't necessarily preserve group structure, we will write $\operatorname{Hom}\mathsf{Set}G, H$.

When we want to point out which operations are preserved, we can write

$$f : (G, *, e_G) \to (H, \star, e_H).$$

When we use multiplicative notation, we can also write $f(ab) = f(a)f(b)$.

Example 3.6.3. $\mathrm{id}_G : G \to G$ is a homomorphism.

Example 3.6.4. If $H \leq G$ is a subgroup and $i : H \hookrightarrow G$ is the inclusion $i(h) = h$, then i is a homomorphism.

Example 3.6.5. The shift map $f : \mathbb{Z} \to \mathbb{Z} : x \mapsto x + 1$ is not a homomorphism:

$$f(x + y) = x + y + 1 \neq$$
$$\neq f(x) + f(y) = x + 1 + y + 1 = x + y + 2.$$

Example 3.6.6. $f : \mathbb{Z} \to \mathbb{Z} : x \mapsto 2x$ is a homomorphism.

Example 3.6.7. The unique map to the trivial group $G \to \{e\}$ is a homomorphism.

Definition 3.6.8. Let G and H be two groups, then the composite

$$\begin{array}{ccccc} G & \longrightarrow & \{e_H\} & \hookrightarrow & H \\ g & \longmapsto & & & e_H \end{array}$$

is called the *trivial homomorphism.*

Proposition 3.6.9. Let $f : G \to H$ be a group homomorphism. Then for any $g \in G$ and $k \in \mathbb{Z}$:

(i) $f(e_G) = e_H$;
(ii) $f(g^{-1}) = f(g)^{-1}$;
(iii) $f(g^k) = f(g)^k$.

Remark 3.6.10. In particular, if $f(e_G) \neq e_H$, then f is not a homomorphism.

PROOF. **(i)** $f(e_G) \star f(e_G) = f(e_G * e_G) = f(e_G) = f(e_G) \star e_H$. Use cancellation to get $f(e_G) = e_H$.
(ii) $f(g^{-1}) \star f(g) = f(g^{-1} * g) = f(e_G) = e_H$.
(iii) Induction on k for $k \geq 0$, then use the current progress and part (2) for $k < 0$.

□

Proposition 3.6.11. If $f : G \to H$ and $g : H \to K$ are group homomorphisms, then $g \circ f$ is a homomorphism as well.

PROOF. Straightforward check of $(g \circ f)(ab) = (g \circ f)(a)(g \circ f)(b)$. □

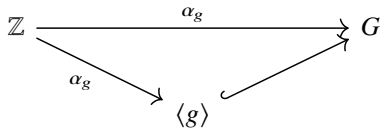
Example 3.6.12. Let G be a group and $g \in G$ any element. Define the *exponent map*:

$$\begin{aligned} \alpha_g : \mathbb{Z} &\to G \\ k &\mapsto g^k \end{aligned}$$

Then α_g is a homomorphism:

$$\alpha_g(k + l) = g^{k+l} = g^k g^l = \alpha_g(k)\alpha_g(l).$$

Notice that α_g factors through the cyclic subgroup $\langle g \rangle \leq G$:



As a particular example, we can take $G = \mathbb{Z}$ and $g = r \in \mathbb{Z}$, then:

$$\begin{aligned} \alpha_r : \mathbb{Z} &\to \mathbb{Z} \\ k &\mapsto kr \end{aligned}$$

Example 3.6.13. $\gamma : \mathbb{Z} \to \mathbb{Z}/n$, $\gamma(a) = [a]$ is a homomorphism.

Recall that for a set map $f : G \to H$, we have defined image $\operatorname{Im} f = f(G)$ and preimage $f^{-1}(K)$ for any subset $K \subseteq H$.

Proposition 3.6.14. Let $f : G \to H$ be a group homomorphism.

(i) For any subgroup $K \leq G$, its image $f(K)$ is a subgroup of H. In particular, $f(K) \leq \operatorname{Im} f = f(G) \leq H$.
(ii) For any subgroup $K \leq H$, its preimage $f^{-1}(K)$ is a subgroup of G.

Slogan 3.6.15. This proposition can be understood as f "preserves subgroups in both directions".

PROOF. Use the subgroup criterion for both parts: check that the identity is in the image/preimage and that it is closed under the group operation and taking inverses. □

There is a special name for the preimage of the trivial subgroup.

Definition 3.6.16. For a group homomorphism $f : G \to H$, the preimage of the trivial subgroup $\{e_H\}$ is called the *kernel* of f, and is denoted by $\operatorname{Ker} f$.

3.6.2. First instance of categorical thinking. You may have noticed a pattern that when we defined sets, we immediately defined set maps, and now that we defined groups, we have just defined group homomorphisms. In general, in mathematics, just like in physics or in real world, studying isolated objects is not as fruitful as looking at interactions between them, so for any type of mathematical objects, you will also see a definition of "admissible interactions" between them – the most general term for them is *morphisms*. A collection of objects of your choice together with morphisms forms a *category* – and of course, there are certain natural axioms that you want to introduce on objects and morphisms, like associativity of composition and existence of identity morphisms. You can now skip ahead and read Section 8.1.1 for a more precise definition, or wait until later in you mathematical journey to pick up category theory.

3.6.3. Surjections and injections. Since a group homomorphism is in particular a set map, we can talk about it being surjective or injective. We will now see how these properties of homomorphisms interact with properties of groups.

Proposition 3.6.17. Let $f : G \to H$ be a surjective homomorphism.

(i) If G is Abelian, then H is Abelian.
(ii) If G is cyclic, then H is cyclic.
(iii) If G is finitely generated, then H is finitely generated.

Proof. **(i)** Let $a, b \in H$. Write them as images of some elements from G and use it commute.
(ii) Take the image of a generator of G – it will be a generator of H.
(iii) Left as exercise.

□

Exercise 3.6.18. Prove Proposition 3.6.17(3).

Proposition 3.6.19. Let $\alpha, \beta : G \to H$ be group homomorphisms. Suppose that $G = \langle S \rangle$ for some $S \subseteq G$. Then $\alpha = \beta$ if and only if $\forall x \in S : \alpha(x) = \beta(x)$.

Slogan 3.6.20. Two homomorphisms coincide if and only if they coincide on generators.

Proof. The "only if" part is tautological.

Now assume that $\forall x \in S : \alpha(x) = \beta(x)$. Take any $g \in G$ and write it as a product of generators, then apply α to it and use properties of homomorphisms to rewrite it as $\beta(g)$. □

Recall that for sets, a set map $f : X \to Y$ is surjective if and only if $\operatorname{Im} f = Y$ – this remains true for group homomorphisms. However we can give a nicer characterization of injective group homomorphisms.

Proposition 3.6.21. Let $f : G \to H$ be a group homomorphism. Then f is injective if and only if $\operatorname{Ker} f = \{e_G\}$.

Proof. Recall that

$$\operatorname{Ker} f = f^{-1}(e_H) = \{g \in G \mid f(g) = e_H\}.$$

Note that we always have $e_G \in \operatorname{Ker} f$.

Let's start with the forward direction and assume that f is injective. Then for any $a \in \operatorname{Ker} f$, we have

$$f(a) = e_H = f(e_G),$$

and therefore $\operatorname{Ker} f = \{e_G\}$.

Conversely, suppose that $\operatorname{Ker} f = \{e_G\}$. Let us test the definition of injectivity: take $a, b \in G$ such that $f(a) = f(b)$. Then we can write:

$$e_H = f(a)f(b)^{-1} = f(a)f(b^{-1}) = f(ab^{-1}).$$

This means that $ab^{-1} \in \operatorname{Ker} f = \{e_G\}$, and hence $ab^{-1} = e_G$, or equivalently, $a = b$. □

Example 3.6.22. $\gamma : \mathbb{Z} \to \mathbb{Z}/n$, $\gamma(a) = [a]$, $\operatorname{Ker} \gamma = n\mathbb{Z}$.

Example 3.6.23. $f : \mathbb{Z} \to n\mathbb{Z}$, $f(k) = nk$, for $n \neq 0$. $\operatorname{Ker} f = \{0\}$, thus f is injective. In fact, it is also surjective – and hence an *isomorphism* – we will define this term next.

3.6.4. Isomorphisms.

Definition 3.6.24. Let $f : G \to H$ be a group homomorphism. We say that it is an *isomorphism of groups* (or just an *isomorphism*) if f is bijective. If there exists an isomorphism $f : G \to H$, we write $G \cong H$ and say that G and H are isomorphic.

Here is the three-step way to checking that $f : G \to H$ is an isomorphism:

Step 1. Check that f is a homomorphism.
Step 2. Check that $\operatorname{Im} f = H$.
Step 3. Check that $\operatorname{Ker} f = \{e_G\}$.

⚠ **Warning 3.6.25.** If $G \cong H$, it doesn't mean that the isomorphism is unique.

Example 3.6.26. $G = (\{-1, 1\}, \cdot, 1)$ and $H = \mathbb{Z}/2$ are isomorphic via $f : G \to H$ that is defined on elements as: $f(1) = [0]$ and $f(-1) = [1]$.

Example 3.6.27. For any group G, we have $G \cong G$ via $\mathrm{id}_G : G \to G$.

Example 3.6.28. $f : \mathbb{Z} \to \mathbb{Z}$, $f(a) = -a$ is an isomorphism that is not identity.

Proposition 3.6.29. Define $\iota : G \to G$ as $\iota(g) = g^{-1}$. Then ι is an isomorphism if and only if G is Abelian.

Proof. Note that ι is always bijective. For the only if part, if G is not Abelian, then ι is not even a homomorphism. □

Theorem 3.6.30. Given any set of groups, isomorphism $\cong$ defines an equivalence relation. In particular, if $f : G \to H$ is an isomorphism, then $f^{-1} : H \to G$ is an isomorphism.

Proof. Reflexivity is provided by identity homomorphisms.

Prove symmetry. Say that $f : G \to H$ is an isomorphism, so f is a bijection. Then there exists the inverse bijection $g : H \to G$ – we will prove that it is a homomorphism. Consider $a, b \in H$, then we have the following equalities, using $f \circ g = \mathrm{id}_H$ and that f is a homomorphism:

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

Since f is a bijection, we conclude that the elements $g(ab) \in G$ and $g(a)g(b) \in G$ are equal.

Transitivity is provided by composition. □

The following corollary is directly implied by the theorem.

Corollary 3.6.31. A homomorphism $\varphi : G \to H$ is an isomorphism if and only if there exists a homomorphism $\psi : H \to G$ such that $\varphi\psi = \mathrm{id}_H$ and $\psi\varphi = \mathrm{id}_G$.

Goal of group theory is to classify groups up to isomorphism – this is difficult!

Proposition 3.6.32. If $\varphi : G \to H$ is an injective homomorphism, then $G \cong \operatorname{Im}\varphi$.

Proposition 3.6.33. Any cyclic group is isomorphic to either $\mathbb{Z}$ or $\mathbb{Z}/n$ for some $n > 0$.

Theorem 3.6.34. Two finite groups are isomorphic if and only if they have the same Cayley tables up to relabelling.

Proof. Relabelling is a set map that provides an isomorphism. □

Example 3.6.35. Up to isomorphism, there is only one group of order one, one group of order two, and one group of order three.

Example 3.6.36. Up to isomorphism, there are two groups of order four, and you described them in your homework:

$$\mathbb{Z}/4 \quad \text{and} \quad \mathbb{Z}/2 \times \mathbb{Z}/2.$$

3.6.5. Invariants. Given two groups G and H, how can we tell when they are not isomorphic? It is not always easy, but there are certain properties that are invariant under isomorphisms, meaning that they don't change.

Example 3.6.37 (Cardinality)**.** We can see that if there is no bijection between G and H as sets, then they cannot be isomorphic as groups. For example, $\mathbb{Z}/6$ and $\mathbb{Z}/12$ are not isomorphic.

Example 3.6.38 (Being Abelian)**.** The groups $\mathfrak{S}_3$ and $\mathbb{Z}/6$ have the same cardinality. However, the former is not Abelian, while the latter is, so they cannot be isomorphic.

Example 3.6.39 (Being cyclic)**.** Both $\mathbb{Z}/4$ and $\mathbb{Z}/2\times\mathbb{Z}/2$ have the same cardinality and are Abelian. However, the former is cyclic, and the latter cannot be generated by just one element, which shows that they are not isomorphic.

Example 3.6.40 (Being finitely generated)**.** $\mathbb{Z}$ and $\mathbb{Q}$ are in bijection, but they cannot be isomorphic because $\mathbb{Z}$ is finitely generated (even cyclic) and $\mathbb{Q}$ is not finitely generated. Similarly, $\mathbb{Z} \times \cdots \times \mathbb{Z}$ cannot be isomorphic to $\mathbb{Q}$.

Example 3.6.41 (Having elements of a certain order)**.** Define the subgroup of the group circle $\mu_\infty := \{e^{2\pi i\alpha} \mid \alpha \in \mathbb{Q}\} \leq \mathbb{S}^1$ that consists of only those elements whose angle is a rational multiple of $2\pi i$. Then we can notice that both $\mathbb{Q}$ and μ_∞ are countably infinite, Abelian, not cyclic, not finitely generated. However, every element in μ_∞ has finite order, while every nonzero element in $\mathbb{Q}$ has infinite order. Since an isomorphism would preserve the order of elements, we conclude that $\mathbb{Q}$ and μ_∞ are not isomorphic.

Example 3.6.42 (Number of elements of a certain order)**.** We can also compute the number of elements of a certain order k, and this will be invariant under isomorphisms. For example, another way to see that $\mathfrak{S}_3$ and $\mathbb{Z}/6$ are not isomorphic is to compute how many elements of order two each of these has. In $\mathfrak{S}_3 = \{e, r, r^2, sr, sr^2\}$, with $r^3 = s^2 = e$ and $rsr = s$, we have two elements of order 2: sr, sr^2. However, $\mathbb{Z}/6$ has only one element of order 2.

3.6.6. Automorphisms.

Definition 3.6.43. Let G be a group. An *automorphism* of G is an isomorphism $G \to G$. The set of all automorphisms of G is called the *group of automorphisms* of G and is denoted by

$$\operatorname{Aut} G = \{f : G \to G \mid f \text{ is an isomorphism}\}.$$

Theorem 3.6.44. $(\operatorname{Aut} G, \circ, \mathrm{id}_G)$ is a group.

For the proof, recall that for any set X we have defined the group $\mathfrak{S}_X$ of permutations of X as units in the monoid $\operatorname{Hom}_{\mathsf{Set}} X, X$ under composition:

$$\mathfrak{S}_X = (\operatorname{Hom}_{\mathsf{Set}} X, X, \circ, \mathrm{id}_X)^\times.$$

Proof. Note that $\operatorname{Aut} G$ is a subset of $\mathfrak{S}_G$, so we can apply the subgroup criterion (Proposition 3.4.10) to check that it is in fact a subgroup. □

Proposition 3.6.45. Let G be a group and $a \in G$ an element. Define the following map:

$$\begin{aligned} \sigma_a : G &\to G \\ g &\mapsto aga^{-1} \end{aligned}$$

Then σ_a is an automorphism of G.

Proof. We first show that σ_a is a homomorphism:

$$\sigma_a(gh) = agha^{-1} = aga^{-1}aha^{-1} = \sigma_a(g)\sigma_a(h).$$

Now, we want to show that σ_a is injective. For this, take any $g \in \operatorname{Ker} \sigma_a$:

$$\sigma_a(g) = e \quad \Leftrightarrow \quad aga^{-1} = e \quad \Leftrightarrow \quad ag = a \quad \Leftrightarrow \quad g = e.$$

So $\operatorname{Ker} \sigma_a = \{e\}$, and therefore σ_a is injective by Proposition 3.6.21.

To show surjectivity, we need for each $g \in G$ find an element $x \in G$ such that $\sigma_a(x) = g$. This is equivalent to the following:

$$\sigma_a(x) = g \quad \Leftrightarrow \quad axa^{-1} = g \quad \Leftrightarrow \quad x = a^{-1}ga.$$

So $g = \sigma_a(a^{-1}ga)$, and therefore σ_a is surjective. □

⚠ **Warning 3.6.46.** We need to conjugate by a and not just multiply by a on one side because the latter does not define a homomorphism: $g \mapsto ag$ does not send the identity element to itself.

Theorem 3.6.47. Let G be a group. Then the following map

$$\begin{aligned} \theta : G &\to \operatorname{Aut} G \\ a &\mapsto \sigma_a \end{aligned}$$

is a homomorphism.

PROOF. To show that θ is a homomorphism, we need to prove that for every $a, b \in G$:

$$\theta(ab) = \theta(a)\theta(b),$$

or equivalently:

$$\sigma_{ab} = \sigma_a \circ \sigma_b.$$

Check the equality of maps on elements $g \in G$:

$$(\sigma_a \circ \sigma_b)(g) = \sigma_a(\sigma_b(g)) = \sigma_a(bgb^{-1}) = abgb^{-1}a^{-1} = (ab)g(ab)^{-1} = \sigma_{ab}(g).$$

□

Definition 3.6.48. We define the *group of inner automorphisms* of G as the subgroup of $\operatorname{Aut} G$ of the possible σ_a, and denote it by $\operatorname{Inn} G$:

$$\operatorname{Inn} G \coloneqq \operatorname{Im} \theta \leq \operatorname{Aut} G.$$

Definition 3.6.49. The subgroup $Z(G) \coloneqq \operatorname{Ker} \theta \leq G$ is called the *center* of G.

Remark 3.6.50. If G is Abelian, then $Z(G) = G$.

Example 3.6.51. $\operatorname{Aut}(\mathbb{Z}/4) = \mathbb{Z}/2$. The only nontrivial automorphism is $f : \mathbb{Z}/4 \to \mathbb{Z}/4$ defined as $f(a) = -a$.

Example 3.6.52. $\operatorname{Aut}(\mathbb{Z}/2\times\mathbb{Z}/2) = \mathfrak{S}_3 = \{\mathrm{id}, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\}$. We define $\sigma((1,0)) = (0,1)$, $\sigma((0,1)) = (1,1)$, $\sigma((1,1)) = (1,0)$, and $\tau((0,1)) = (1,0)$, $\tau((1,0)) = (0,1)$, $\tau((1,1)) = (1,1)$.

3.7. Cayley's theorem and symmetric groups

3.7.1. Permutation groups and Cayley's theorem. So far, we have been studying groups as certain kinds of binary operations – we have a lot of examples, but they are all quite abstract. In this section, we will show that any group can be understood as a group of certain permutations of a set, which give a more concrete model to have in mind.

Definition 3.7.1. Recall that, for a set X, we denote the group of all bijections $X \to X$ by $\mathfrak{S}_X$. A *permutation group* is a subgroup of $\mathfrak{S}_X$.

Let G be a group. We can define a homomorphism

$$\Phi\colon G \longrightarrow \mathfrak{S}_G$$
$$g \longmapsto \mu_g,$$

where μ_g is the left multiplication map that is defined on elements as follows:

$$\mu_g(x) = gx.$$

Theorem 3.7.2 (Cayley)**.** The homomorphism $\Phi\colon G \to \mathfrak{S}_G$ is injective, so G is isomorphic to a group of permutations. In particular, if $|G| = n$, then G is isomorphic to a subgroup of $\mathfrak{S}_n$.

Proof. Take $g \in \operatorname{Ker}\Phi$, which by definition of Φ means that $\Phi(g) = \mathrm{id}_G$. In turn, it means that

$$\forall x \in G\colon \mu_g(x) = gx = x.$$

By the cancellation property, we can conclude that $g = e$ is the identity element, hence $\operatorname{Ker}\Phi = \{e\}$, and Φ is injective. □

This shows that studying symmetric groups is important for the theory of finite groups.

3.7.2. Symmetric groups.

Notation 3.7.3. Recall that we denote by $\mathfrak{S}_n$ the group of bijections of the set $\{1, \dots, n\}$. Elements of this group are called permutations, and the group $\mathfrak{S}_n$ is called the *symmetric group of degree n*. We will denote the identity permutation by ε. A permutation $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ can be compactly defined via the following table:

$$\begin{pmatrix} 1 & 2 & \cdots & n \\ \sigma(1) & \sigma(2) & \cdots & \sigma(n) \end{pmatrix}.$$

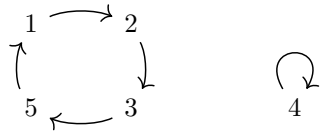
Example 3.7.4. Consider the following element of $\mathfrak{S}_5$:

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

On elements, this permutation acts like this:

$$\sigma(1) = 2, \quad \sigma(2) = 3, \quad \sigma(3) = 5, \quad \sigma(4) = 4, \quad \sigma(5) = 1,$$

or pictorially:



Example 3.7.5. We can write out all elements of the symmetric group $\mathfrak{S}_3$ of degree three:

$$\mathfrak{S}_3 = \left\{ \varepsilon = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} \right\}.$$

Let us compute a composition:

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

We can also compute the inverse easily by "reading from bottom to top":

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}.$$

Example 3.7.6. Take the following elements σ and τ in $\mathfrak{S}_5$:

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

We can compute that $\sigma\tau \neq \tau\sigma$:

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

By generalizing this example, we can see that $\mathfrak{S}_n$ is not Abelian for any $n \geq 3$.

Proposition 3.7.7. $|\mathfrak{S}_n| = n!$.

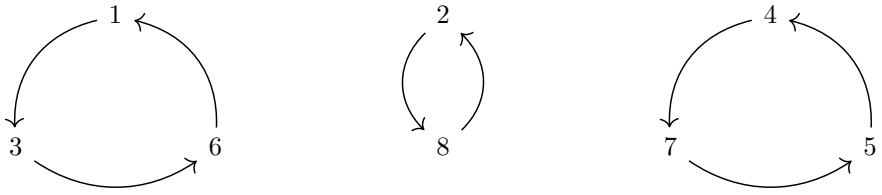
3.7.3. Orbits and cycles.

Definition 3.7.8. Given $\sigma \in \mathfrak{S}_n$, define a partition of $\{1, \ldots, n\}$:

$$a \sim b \quad \Longleftrightarrow \quad \exists k \in \mathbb{Z} : b = \sigma^k(a).$$

The *orbits* of σ are the equivalence classes defined by this partition. An orbit is called *trivial* if its size is one.

Example 3.7.9. We can enumerate the three orbits of $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 8 & 6 & 7 & 4 & 1 & 5 & 2 \end{pmatrix}$: $\{1, 3, 6\}$, $\{2, 8\}$, $\{4, 5, 7\}$.



Notice that each circle determines a permutation in $\mathfrak{S}$, for example the first orbit corresponds to the following permutation:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 2 & 6 & 4 & 5 & 1 & 7 & 8 \end{pmatrix}.$$

Definition 3.7.10. A permutation $\sigma \in \mathfrak{S}_n$ is called a *cycle* if it has exactly one nontrivial orbit. The size of the nontrivial orbit is called the *length* of σ; notice that it must be at least 2. If k is the length, we say that σ is a k-cycle. If σ is a k-cycle and a is some element that is moved by σ, we can use a more economical notation for the permutation: $\sigma = (a\, \sigma(a)\, \sigma^2(a) \cdots \sigma^{k-1}(a))$.

Example 3.7.11. $(1\,3\,6)$ is the cycle corresponding to the first orbit in the previous example. This notation is not unique, for example $(1\,3\,6) = (3\,6\,1) = (6\,1\,3)$.

Definition 3.7.12. Two permutations $\sigma,\ \tau \in \mathfrak{S}_n$ are *disjoint* if no element in $\{1, \ldots, n\}$ is moved by both σ and τ.

Theorem 3.7.13. Every permutation $\sigma \in \mathfrak{S}_n$ is a product of disjoint cycles.

PROOF. Consider the orbits $O_1, \ldots, O_r$ of σ. □

Example 3.7.14.

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 8 & 6 & 7 & 4 & 1 & 5 & 2 \end{pmatrix} = (1\,3\,6)(2\,8)(4\,5\,7).$$

Convention 3.7.15. In our decomposition into cycles, we omit orbits of size 1.

Definition 3.7.16. A 2-cycle is called a *transposition*.

Proposition 3.7.17. Every cycle is a product of (non-disjoint) transpositions. Therefore, every permutation in $\mathfrak{S}_n$ is a product of transpositions, in a non-unique way.

PROOF. $(a_1\, a_2\, \cdots\, a_k) = (a_1\, a_2)(a_2\, a_3) \cdots (a_{k-1}\, a_k)$. □

Theorem 3.7.18. Suppose that an element $\sigma \in \mathfrak{S}_n$ is written as a product of transpositions in two ways:

$$\sigma = \sigma_r \cdots \sigma_1 = \tau_s \cdots \tau_1.$$

Then $r \equiv s \mod 2$.

PROOF. View $\mathfrak{S}_n \cong \mathfrak{S}_A$, where A is the set of rows of the identity matrix I_n. A permutation σ permutes the rows of I_n, and changes the sign of the determinant depending on the number of transpositions of rows. □

Definition 3.7.19. A permutation σ is called *even* if it can be decomposed as a composition of an even number of transpositions; and *odd* if the number of transpositions is odd. This defines the following homomorphism:

$$\operatorname{sgn}\colon \mathfrak{S}_n \to \{1, -1\},$$

where $\operatorname{sgn}\sigma = 1$ if σ is even, and $\operatorname{sgn}\sigma = -1$ if σ is odd.

3.8. Cosets

3.8.1. Definitions. Recall that for $n \geq 2$, we defined the set $\mathbb{Z}/n$. In this section, we will offer a new perspective on the construction. Namely, note that we have a subgroup $n\mathbb{Z} \leq \mathbb{Z}$, and we can write the residue classes in $\mathbb{Z}/n = \mathbb{Z}/n\mathbb{Z}$ as follows:

- $[0] = n\mathbb{Z}$;
- $[1] = 1 + n\mathbb{Z} = \{1 + kn | k \in \mathbb{Z}\}$;
- $[r] = r + n\mathbb{Z} = \{r + kn | k \in \mathbb{Z}\}$.

Definition 3.8.1. Let G be a group and $H \leq G$ be its subgroup. We exhibit two partitions of G induced by H induced by the following two equivalence relations:

- $a, b \in G$ are left related if $a^{-1}b \in H$, and we write $a \sim_L b$;
- $a, b \in G$ are right related if $ab^{-1} \in H$, and we write $a \sim_R b$.

Theorem 3.8.2. The two relations $\sim_L$ and $\sim_R$ are equivalence relations on G.

Proof. □

Definition 3.8.3. Let $H \leq G$ and $a \in G$. Define:

(i) $aH = \{ah | h \in H\} \subset G$ — *left coset* of H generated by a.
(ii) $Ha = \{ha | h \in H\} \subset G$ — *right coset* of H generated by a.

Lemma 3.8.4. Let $H \leq G$ and $a, b \in G$, then:

(i) $a \sim_L b \quad \Leftrightarrow \quad b \in aH \quad \Leftrightarrow \quad aH = bH$;
(ii) $a \sim_R b \quad \Leftrightarrow \quad b \in Ha \quad \Leftrightarrow \quad Ha = Hb$.

This justifies the names "left" and "right".

Definition 3.8.5. Given $H \leq G$, we introduce the following notation for the sets of left and right cosets:

(i) $G/H \coloneqq G/\sim_L = \{aH \mid a \in G\}$ – left cosets;
(ii) $H \setminus G \coloneqq G/\sim_R = \{Ha \mid a \in G\}$ – right cosets.

⚠ **Warning 3.8.6.** G/H and $H \setminus G$ are not always groups!

Example 3.8.7. $G = \mathbb{Z}/6$, $H = \{0, 3\}$. We can notice that here $G/H = H \setminus G$.

Remark 3.8.8. If G is an abelian group with a subgroup $H \leq G$, and $a \in G$ is an element, then $aH = Ha$.

Example 3.8.9. $G = \mathfrak{S}_3$, $H = \langle (1\,2) \rangle = \{\varepsilon, (1\,2)\}$.

(i) Some of the left cosets: H, $(1\,2\,3)H = \{(1\,2\,3), (1\,3)\}$, $(1\,3\,2)H = \{(1\,3\,2), (2\,3)\}$.
(ii) Some of the right cosets: H, $H(1\,2\,3) = \{(1\,2\,3), (2\,3)\}$, $H(1\,3\,2) = \{(1\,3\,2), (1\,3)\}$.

Here $G/H \neq H \setminus G$.

3.8.2. Lagrange's theorem.

Theorem 3.8.10. Let $H \leq G$ and $a \in G$.

(i) There are bijections between H, aH and Ha.
(ii) There is a bijection between G/H and $H \setminus G$.

Proof. **(i)** A set map $\varphi : H \to aH$ that sends h to $\varphi(h) = ah$ is a bijection with the inverse $\psi : aH \to H$, $\psi(g) = a^{-1}g$.

(ii) Notice that $aH = bH \quad \Leftrightarrow \quad Ha^{-1} = Hb^{-1}$. Use this observation to prove that $\varphi : G/H \to H \setminus G$ that sends aH to Ha^{-1} is a well-defined map with the inverse $\psi : H \setminus G \to G/H$, $\psi(Hb) = b^{-1}H$.

□

Definition 3.8.11. Let $H \leq G$. The *index* of H in G is the cardinality of G/H, and it is denoted by $[G : H]$, or $(G : H)$, or $|G : H|$.

Example 3.8.12. $[\mathbb{Z} : 3\mathbb{Z}] = |\mathbb{Z}/3| = 3$.

Theorem 3.8.13 (Lagrange)**.** Let G be a finite group and $H \leq G$. Then:

(i) $|G| = |H| \cdot [G : H]$;
(ii) $|H|$ divides $|G|$.

Proof. We showed that G/H is a partition of G, and every coset $aH \in G/H$ has the same number of elements as $|H|$. □

3.8.3. Corollaries of Lagrange's theorem.

Corollary 3.8.14. Let G be a finite group of order $n = |G|$. Then $\forall g \in G : g^n = e$.

In particular, this corollary implies Fermat's little theorem.

Corollary 3.8.15. Let G be a finite group of prime order p. Then G is cyclic and has no nontrivial proper subgroups.

Definition 3.8.16. We define the *dihedral group* D_n, for $n \geq 1$, as follows:

$$D_n = \{e, r, r^2, \ldots, r^{n-1}, s, sr, sr^2, \ldots, sr^{n-1}\},$$

where $\operatorname{ord} r = n$, $\operatorname{ord} s = 2$, and $rsr = s$.

We can think of D_n as group of symmetries of a regular two-sided n-gon. Then r would be rotation by $\frac{2\pi}{n}$ and s would be a reflection around a fixed axis.

Theorem 3.8.17. Let G be a group of order $2p$, where p is prime. Then either G is cyclic or $G \cong D_p$.

Proof. We first note that the case $p = 2$ is trivial because there are only two groups of order four: $\mathbb{Z}/4$ is cyclic and $\mathbb{Z}/2 \times \mathbb{Z}/2 \cong D_2$. So we assume for the rest of the argument that $p > 2$.

Suppose that G is not cyclic, then we will show that $G \cong D_p$. Notice that for any $g \in G$, the order $\operatorname{ord} g$ can be 1, 2 or p. It cannot be $2p$ because G is not cyclic.

Claim 1. There is $r \in G$ such that $\operatorname{ord} r = p$.

Proof 1. Suppose there was no such r, then for any $g \in G$, we have $g^2 = e$, and therefore G is abelian. Then for any two nontrivial $g \neq h \in G$, we get a subgroup of order four: $\{e, g, h, gh\}$. This is a contradiction.

We can therefore pick $r \in G$ such that $\operatorname{ord} g = p$, and define $H = \langle r \rangle$.

Claim 2. Pick an element $s \in G$ such that $s \notin H$, then $\operatorname{ord} s = 2$.

Proof 2. $G = H \sqcup sH$, $s^2 \in H$. If $\operatorname{ord} s = p$, then $s = s^{p+1} = (s^2)^{\frac{p+1}{2}} \in H$, which is a contradiction. Thus $\operatorname{ord} s = 2$.

Claim 3. $rsr = s$.

Proof 3. We have $sr \notin H$, so by the above argument we get that $\operatorname{ord} sr = 2$, that is $srsr = e$. □

We can now easily classify all groups of order up to 7, and for a prime p, of orders p and $2p$.

3.9. Quotient groups

3.9.1. Normal subgroups. Given a group G and its subgroup $H \leq G$, we had two partitions $G/H = \{aH \mid a \in G\}$ and $H \setminus G = \{Ha \mid a \in G\}$. We will now give a name to the situation when $G/H = H \setminus G$, in other words $aH = Ha$ for all $a \in G$.

Definition 3.9.1. A subgroup $N \leq G$ is *normal* if $\forall g \in G\colon gN = Ng$. In this case we write $N \trianglelefteq G$. When N is a proper normal subgroup, we write $N \triangleleft G$.

Example 3.9.2. $\{e\} \trianglelefteq G$, $G \trianglelefteq G$.

Example 3.9.3. If G is abelian, then every subgroup is normal.

Example 3.9.4. G can be non-abelian, but such that all of its subgroups are normal. For example, we can take the *quaternion group* of order 8 that is described as follows:

$$\begin{aligned} G = \{ &\pm 1, \pm i, \pm j, \pm k\} \\ &i^2 = j^2 = k^2 = -1 \\ &ij = k = -ji \\ &jk = i = -kj \\ &ki = j = -ik \end{aligned}$$

Example 3.9.5. $G = \mathfrak{S}_3$. $H = \{\varepsilon, (1\,2)\}$ is not normal. $N = \{\varepsilon, (1\,2\,3), (1\,3\,2)\}$ is normal.

Example 3.9.6. In fact, for every n, the subgroup $\mathfrak{A}_n = \ker(\mathrm{sgn}) \leq \mathfrak{S}_n$ of even permutations is normal.

Proposition 3.9.7. Let $H \leq G$. Then the following are equivalent:

(i) $H \trianglelefteq G$;
(ii) $\forall g \in G\colon gHg^{-1} \subseteq H$;
(iii) $\forall g \in G\colon gHg^{-1} = H$.

Here $gNg^{-1} = \{gxg^{-1} \mid x \in N\}$.

Slogan 3.9.8. Normal subgroups are those invariant under conjugation. Recall that $\Theta\colon G \to \operatorname{Aut} G$ was defined as $g \mapsto \sigma_g$, and $\sigma_g(x) = gxg^{-1}$. Then by the above proposition, we have that H is normal in G if and only if $\forall g \in G\colon \sigma_g(H) = H$.

Theorem 3.9.9. Let $\varphi : G \to H$ be a homomorphism. Then $\ker \varphi$ is a normal subgroup of G.

We will show later that any normal subgroup can be realized as the kernel of some homomorphism.

Example 3.9.10. $Z(G) = \ker \Theta \trianglelefteq G$.

Example 3.9.11. $\langle r \rangle \trianglelefteq D_n$.

Exercise 3.9.12. Generalize the previous example by showing that if $H \leq G$ and $[G : H] = 2$, then $H \trianglelefteq G$.

Definition 3.9.13. A group G is called *simple* if it has exactly two normal subgroups: $\{1\}$ and G.

Example 3.9.14. Cyclic groups of prime order are simple.

Exercise 3.9.15. There are no other finite simple Abelian groups.

Example 3.9.16 (hard)**.** $\mathfrak{A}_n$ is simple for $n \geq 5$, but $\mathfrak{A}_4$ is not simple.

To sum up, if we have $N \trianglelefteq G$, then $G/N = N \setminus G$.

3.9.2. Quotient groups.

Proposition 3.9.17. Let $N \leq G$ be a subgroup. Then N is normal if and only if the map

$$\begin{aligned} G/N \times G/N &\xrightarrow{*} G/N \\ (aN, bN) &\mapsto (ab)N \end{aligned}$$

is well-defined.

Proof. For the forward direction, suppose $N \trianglelefteq G$ is a normal subgroup. To show that the operation $aN * bN = (ab)N$ is well-defined, we need to show that for any other choice of elements $an \in aN$ and $bm \in bN$ (with $n, m \in N$), we have $aN * bN = anN * bmN$, i.e. $(ab)N = (anbm)N$. Since N is a normal subgroup, we have

$$b^{-1}nb = n' \quad \text{for some } n' \in N.$$

Therefore, we get

$$anbm = ab(b^{-1}nb)m = (ab)(nm).$$

Now notice that $nm \in N$, hence $(nm)N = N$ and $(anbm)N = (ab)(nm)N = (ab)N$, as desired.

We now prove the converse: for $g \in G$, we want to show that $gN = Ng$. Let $a \in gN$, then $gN * g^{-1}N = N$, so $ag^{-1} \in N$ and therefore $a \in Ng$. This shows that $gN \subseteq Ng$. Similarly one can show $Ng \subseteq gN$. □

Theorem 3.9.18. Let $N \trianglelefteq G$ be a normal subgroup of G, then G/N has a natural group structure.

Proof. We have already showed that the candidate multiplication map is well-defined. Then it is straightforward to check the axioms.

- Associativity: $aN * (bN * cN) = aN * (bc)N = (a(bc))N = ((ab)c)N = (ab)N * cN = (aN * bN) * cN$.
- Identity: $eN * aN = (ea)N = aN = (ae)N = aN * eN$.
- Inverse: for $aN \in G/N$, its inverse $(aN)^{-1}$ is $a^{-1}N$, since $aN * a^{-1}N = (aa^{-1})N = eN$, and similarly $a^{-1}N * aN = eN$.

□

Exercise 3.9.19. If G is an abelian group with a normal subgroup N, then G/N is abelian as well.

Definition 3.9.20. Let $N \trianglelefteq G$, then G/N is the *quotient group* of G by N.

Example 3.9.21. For $n\mathbb{Z} \trianglelefteq \mathbb{Z}$, we had $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}/n$.

Example 3.9.22. $G/\{1\} \cong G$, $G/G \cong \{1\}$.

Example 3.9.23. $\mathfrak{S}_n/\mathfrak{A}_n \cong \{1, -1\}$ has order two.

Example 3.9.24. Take $G = D_4$ and $Z = \{e, r^2\}$. Then $D_4/Z \cong \mathbb{Z}/2 \times \mathbb{Z}/2$.

In general, $|G/N| = \frac{|G|}{|N|}$ by Lagrange's theorem.

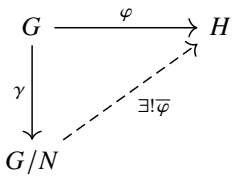
3.9.3. Universal property. Given an equivalence relation, we can consider the canonical projection map from the given set to the quotient set. In particular, for subgroups $N \leq G$ we get a set map

$$\gamma\colon G \to G/N\colon g \mapsto gN$$

that is a group homomorphism if and only if N is normal.

Notice that if $\gamma(g) = eN = N$, then $gN = N$, so $g \in N$. Thus $\ker\gamma = N$. We will show that γ is the "universal" homomorphism with kernel N.

Theorem 3.9.25 (Universal property of quotient group)**.** Let G be a group with a normal subgroup N. Let $\varphi : G \to H$ be a homomorphism such that $N \subseteq \ker\varphi$. Then there exists a unique homomorphism $\overline{\varphi}\colon G/N \to H$ such that $\overline{\varphi}\gamma = \varphi$, i.e. the following diagram commutes:



In particular, for any $g \in G$, we have $\overline{\varphi}(gN) = \varphi(g)$.

Proof. The formula $\overline{\varphi}(gN) = \varphi(g)$ is well-defined, hence we showed the existence part. Uniqueness. Homomorphism. □

Remark 3.9.26 (Key takeaway)**.** In order to define a homomorphism $G/N \to H$, one can start with a homomorphism $\varphi\colon G \to H$ such that $N \subseteq \ker\varphi$.

Example 3.9.27. $\mathbb{Z}/4 \to \mathbb{Z}/2\colon [x]_4 \mapsto [x]_2$. To show that this is a well-defined homomorphism, we can start with $\varphi : \mathbb{Z} \to \mathbb{Z}/2$, $\varphi(x) = [x]_2$ and observe that $4\mathbb{Z} \subseteq \ker\varphi$.

3.10. Isomorphism theorems

3.10.1. First isomorphism theorem.

Remark 3.10.1. Let us take a group G with a normal subgroup N. Assume that there exists a group G' together with a homomorphism $q : G \to G'$ such that for every homomorphism $\varphi : G \to H$, if $N \leq \operatorname{Ker}\varphi$, then there exists a unique homomorphism $\overline{\varphi} : G' \to H$ such that $\overline{\varphi}q = \varphi$. Then we can apply the universal property of the quotient G/N and get that $G' \cong G/N$.

Let us look at the case when N is the kernel of some homomorphism.

Theorem 3.10.2 (First isomorphism theorem)**.** Let $\varphi : G \to H$ be a homomorphism. Then $\operatorname{Im}\varphi \cong G/\operatorname{Ker}\varphi$.

Proof. We can define a natural map $q : G \to \operatorname{Im}\varphi$ that takes g to $q(g) \coloneqq \varphi(g)$. By the universal property of the quotient, we have a unique homomorphism $\overline{q} : G/\operatorname{Ker}\varphi \to \operatorname{Im}\varphi$. This homomorphism is clearly surjective.

So it is enough to show that $\overline{q}$ is injective. □

Remark 3.10.3. If you are asked to show that $G/N \cong H$, you can find a surjective homomorphism $G \to H$ with kernel N.

Example 3.10.4. $\mathbb{R}/\mathbb{Z} \cong \mathbb{S}^1$.

3.10.2. Second isomorphism theorem.

Definition 3.10.5. Given $H, K \leq G$, we define the following subset of G:

$$HK \coloneqq \{hk \mid h \in H, k \in K\}.$$

Lemma 3.10.6. HK is a subgroup $\iff HK = KH \iff KH$ is a subgroup.

Exercise 3.10.7. Prove this lemma.

Lemma 3.10.8. Assume that N and M are normal subgroups of G and $N \cap M = \{e\}$. Then $NM \cong N \times M$.

Exercise 3.10.9. Prove this lemma.

Theorem 3.10.10 (Second isomorphism theorem)**.** Let G be a group, $H \leq G$ be a subgroup and $N \trianglelefteq G$ be a normal subgroup. Then:

- **(i)** $N \trianglelefteq HN \leq G$;
- **(ii)** $H \cap N \trianglelefteq H$;
- **(iii)** $HN/N \cong H/(H \cap N)$.

Proof. **(i)** Exercise.

(ii) We know that $H \cap N$ is a subgroup of H, so we are left to show that it is a normal subgroup, i.e. that for every $h \in H$ we have $h(H \cap N)h^{-1} \subseteq H \cap N$. Fix $h \in H$ and take any $n \in H \cap N$. Then $hnh^{-1} \in N$ because N is normal in G, but also $hnh^{-1} \in H$ since H is a subgroup.

(iii) Define $\varphi : H \to HN/N$ as $\varphi(h) = hN$. Check that this is a homomorphism. We can check that its kernel is $\operatorname{Ker}\varphi = H \cap N$. So we get the desired result by the first isomorphism theorem.

□

Example 3.10.11. $a, b \geq 1$ integers. Define $m \coloneqq \operatorname{lcm}(a, b)$ and $d \coloneqq \gcd(a, b)$. Let $H = a\mathbb{Z}$ and $N = b\mathbb{Z}$. Then $H + N = d\mathbb{Z}$, $H \cap N = m\mathbb{Z}$, so $d\mathbb{Z}/b\mathbb{Z} \cong a\mathbb{Z}/m\mathbb{Z}$.

3.10.3. Third isomorphism theorem.

Lemma 3.10.12. Let $\varphi\colon G \to H$ be a homomorphism.

(i) If $N \trianglelefteq H$, then $\varphi^{-1}(N) \trianglelefteq G$.
(ii) If $N \trianglelefteq G$ and φ is surjective, then $\varphi(N) \trianglelefteq H$.

PROOF. **(i)** To show the claim, we need to check that for every $x \in \varphi^{-1}(N)$ and $g \in G$ we have $gxg^{-1} \in \varphi^{-1}(N)$. We know that $\varphi(x) \in N$, so

$$\varphi(xgx^{-1}) = \varphi(x)\varphi(g)\varphi(x)^{-1} \in N$$

as N is normal in H. Hence $gxg^{-1} \in \varphi^{-1}(N)$.

(ii) We need to show that for every $x \in N$ and $h \in H$, we have $h\varphi(x)h^{-1} \in \varphi(N)$. As φ is surjective, there exists $g \in G$ such that $\varphi(g) = h$, so we have

$$h\varphi(x)h^{-1} = \varphi(g)\varphi(x)\varphi(g)^{-1} = \varphi(gxg^{-1}).$$

Since N is normal in G, we have $gxg^{-1} \in N$, and hence

$$h\varphi(x)h^{-1} = \varphi(gxg^{-1}) \in \varphi(N).$$

□

Theorem 3.10.13 (Third isomorphism theorem)**.** Let G be a group and $N, M \trianglelefteq G$ be two normal subgroups such that $N \leq M$. Then:

(i) $M/N \trianglelefteq G/N$;
(ii) $G/M \cong {}^{(G/N)}\!/_{(M/N)}$;
(iii) in particular, $[G : N] = [G : M][M : N]$.

PROOF. **(i)** The quotient map $\gamma : G \to G/N$ is a surjective homomorphism, and since $M \trianglelefteq G$, we have by Lemma 3.10.12 that $M/N = \gamma(M)$ is a normal subgroup of G/N.

(ii) Let us denote the quotient map from G/N by

$$\gamma'\colon G/N \to {}^{(G/N)}\!/_{(M/N)}.$$

The composition F

$$F \coloneqq \gamma' \circ \gamma\colon G \to {}^{(G/N)}\!/_{(M/N)}$$

is a surjective homomorphism. Let us describe its kernel:

$$\begin{aligned}\operatorname{Ker} F &= \{g \in G \mid \gamma'(\gamma(g)) = e\} = \gamma^{-1}\left(\gamma'^{-1}(e)\right)\\ &= \gamma^{-1}(M/N) = M.\end{aligned}$$

So by the first isomorphism theorem, we get the desired isomorphism:

$${}^{(G/N)}\!/_{(M/N)} = \operatorname{Im} F \cong G/\operatorname{Ker} F = G/M.$$

□

Example 3.10.14. $12\mathbb{Z} \leq 6\mathbb{Z} \leq \mathbb{Z}$, then ${}^{\mathbb{Z}/12\mathbb{Z}}\!/_{6\mathbb{Z}/12\mathbb{Z}} \cong \mathbb{Z}/6\mathbb{Z}$.

3.10.4. Fourth isomorphism theorem.

Theorem 3.10.15 (Correspondence theorem, or fourth isomorphism theorem)**.** Let G be a group and $N \trianglelefteq G$ be its normal subgroup. Then there is a natural bijection between subgroups of G containing N and subgroups of G/N:

$$\{H \leq G \mid N \trianglelefteq H \leq G\} \underset{\beta}{\overset{\alpha}{\rightleftarrows}} \{H' \leq G/N\}$$

Under this bijection, normal subgroups correspond to normal subgroups.

$$\{H \trianglelefteq G \mid N \trianglelefteq H \trianglelefteq G\} \longleftrightarrow \{H' \trianglelefteq G/N\}.$$

Proof. Let $\gamma\colon G \to G/N$ be the quotient map. We define $\alpha(H) \coloneqq H/N$ and $\beta(H') \coloneqq \gamma^{-1}(H')$. It is straightforward to check that the compositions $\alpha\beta$ and $\beta\alpha$ are identities, hence define mutually inverse bijections. By Lemma 3.10.12, the maps α and β send normal subgroups to normal subgroups. □

3.10.5. Commutator subgroups. As an application of the correspondence theorem, we can answer the question, given a group G with a normal subgroup N, when G/N becomes Abelian.

Definition 3.10.16. Let $g, h \in G$. The *commutator* of g and h is the element $[g,h] = ghg^{-1}h^{-1} \in G$.

Remark 3.10.17. We can observe that $[g,h] = e$ if and only if $gh = hg$, that is when the elements g and h *commute*.

Example 3.10.18 (Parallel parking)**.** Consider the group G of movements of an arrow on a plane (think of a car on a parking lot), so we allow parallel translations, rotations, but no stretching, shearing or flipping. Denote by $g \in G$ parallel translation by 3 feet in the direction of the arrow, and by $h \in G$ rotation clockwise by 30°. Then the commutator $[g,h]$ is parallel translation of the arrow to the left (and a bit forward). Performing several times the sequence that the commutator represents, that is doing $[g,h]^n$, is what constitutes parallel parking.

Definition 3.10.19. Define the *commutator subgroup*

$$[G,G] \coloneqq \langle \{[g,h] \mid g,h \in G\} \rangle$$

as the subgroup generated by the set of all commutators.

Lemma 3.10.20. Suppose $N \trianglelefteq G$ is a normal subgroup such that the quotient G/N is abelian. Then $[G,G] \leq N$.

We see therefore that if the commutator is a normal subgroup, then the quotient by it is an Abelian group, which leads to the following.

Definition 3.10.21. The *abelianization* of G is the quotient G by the commutator subgroup $G_{\mathrm{ab}} \coloneqq {}^{G}\!/_{[G,G]}$.

In order to check that the definition makes sense, we need to verify that the commutator is indeed a normal subgroup. This, and several other properties, will be explained in the following.

Theorem 3.10.22. Let G be a group, then we have the following.

(i) $[G, G] \trianglelefteq G$;

(ii) G_{ab} is an abelian group;

(iii) given $H \leq G$, we have the equivalence:

$$[G, G] \leq H \quad \Longleftrightarrow \quad H \trianglelefteq G \text{ and } G/H \text{ is abelian;}$$

(iv) for any abelian group A, there is a natural bijection

$$\operatorname{Hom} \mathsf{Grp} G_{\mathrm{ab}}, A \cong \operatorname{Hom} \mathsf{Grp} G, A.$$

3.11. Group presentation

3.11.1. Informal motivation. So far we have been describing groups by listing all of their elements, however we would like to have a more economic way of encoding all of the structure of a group. We know what a set of generators of a group is; given this set, we can enumerate some "equations" that products of these generators satisfy. These are called *relations*. Here are some examples.

Example 3.11.1. If $e \in G$ is the identity in a group, then the relation $eg = g$ is satisfied for any $g \in G$.

Example 3.11.2. If G is an Abelian group, then for every pair $g, h \in G$ we have a relation $gh = hg$.

Example 3.11.3. $\mathbb{Z}/n$ is isomorphic to a group generated by an element a together with a relation $a^n = e$.

Example 3.11.4. $D_n = \{e, r, \dots, r^{n-1}, s, sr, \dots, sr^{n-1}\}$ is generated by r and s, subject to relations $r^n = e$, $s^2 = e$, $rsr = s$.

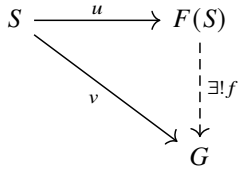
Every relation can be stated as a product of elements in G that is forced to be e: for example, with $s^2 = s$, we can rewrite $rsr = s$ as $rsrs = e$.

So we notice that to define a group, it is enough to give a list of generators and a list of relations. This two-step process consists of first defining *free groups*—groups that have no relations, and then taking quotient by the subgroup generated by the relations; and describing a group G in terms of generators and relations is called giving its *presentation*.

Example 3.11.5. In §3.11.4, you will understand that D_n admits the following presentation: $D_n = \langle r, s \mid r^n, s^2, rsrs \rangle$.

3.11.2. Free groups.

Definition 3.11.6. Let S be a set. The *free group* $F(S)$ *on the set* S is a pair $(F(S), u)$, where $F(S)$ is a group and $u\colon S \to F(S)$ is a set map, that satisfies the following universal property: for any group G and any set map $v\colon S \to G$, there exists a unique group homomorphism $f\colon F(S) \to G$ such that $fu = v$, in other words the following diagram commutes:



The above universal property can also be formulated as a natural bijection $\operatorname{Hom}\mathsf{Set}S, G \cong \operatorname{Hom}\mathsf{Grp}F(S), G$.

Theorem 3.11.7. For any set S, the free group $F(S)$ exists and is unique up to isomorphism.

Definition 3.11.8. We need to introduce some notions that will be used for the construction in the theorem.

- Given a set S, we can think of it as an *alphabet*, then an element $a \in S$ will be called a *letter*.

- Given $a \in S$ and $n \in \mathbb{Z}$, we call the formal expression a^n a *syllable*. For brevity, we write $a = a^1$.
- A finite string of syllables written in a sequence is called a *word*: $a_1^{n_1} \cdots a_k^{n_k}$.
- We also introduce the *empty word* and denote it by e.

Example 3.11.9. Let $S = \{a, b, c\}$, then the following are examples of words:

$$c^{11},$$
$$b^3a^{-1}ca^2b^{-7},$$
$$ab^5aa^{-3}a^2c^{10}b^2b^4.$$

Definition 3.11.10. In some cases, we may have used too many syllables in one word and we would want to contract it.

- Whenever in a word we have two syllables in a row that use the same letter $a^m a^n$, we replace it by one syllable a^{m+n} without changing the rest of the word. Whenever we have a syllable a^0, we replace it by e. Whenever we have the empty word e appear next to another syllable, we omit e, namely we replace both ea^n and $a^n e$ by a^n. Performing one of these operations is called a *contraction*.
- We say that a word is *reduced* if we cannot contract it anymore.

Example 3.11.11. In the previous example with $S = \{a, b, c\}$, the first two words are reduced, while the third admits a sequence of contractions:

$$ab^5aa^{-3}a^2c^{10}b^2b^4 \mapsto ab^5a^{-2}a^2c^{10}b^6 \mapsto ab^5ec^{10}b^6 \mapsto ab^5c^{10}b^6.$$

The last word in this sequence is reduced.

3.11.3. Ranks of free groups.

Definition 3.11.12. Given $F(S)$, we say that S is the *set of generators* of $F(S)$, and the cardinality of S is called the *rank* of $F(S)$.

Proposition 3.11.13. Two free groups are isomorphic if and only if they have equal ranks.

Theorem 3.11.14 (Nielsen-Schreier)**.** Every subgroup of a free group is free.

The proof of this theorem is omitted for this course, since it uses methods of algebraic topology. It is interesting to note however that it significantly relies on the axiom of choice, and this theorem can be wrong in a model of set theory where the axiom of choice doesn't hold.

3.11.4. Relations and group presentation.

Definition 3.11.15. *Relations.*

Definition 3.11.16. *Group presentation*, G is *presented* by S with relations R.

Example 3.11.17. Presentation is not unique, for example $\mathfrak{S}_3 \cong D_3$.

3.12. Classification of finitely generated abelian groups

3.12.1. Statement of the main theorem. You have an exercise to prove that every group is determined by its collection of finitely generated subgroups, so it would make sense to describe those. In general it is hard, but if we only consider abelian groups, we can get a very precise statement.

Theorem 3.12.1. Let A be a finitely generated abelian group. Then

$$A \cong \mathbb{Z}^n \times \mathbb{Z}/k_1 \times \cdots \mathbb{Z}/k_r$$

for uniquely determined $n, r \in \mathbb{N}$ and $k_1, \ldots, k_r \in \mathbb{N} \setminus \{0\}$ such that $k_i | k_{i+1}$ for each $i = 1, \ldots, r-1$.

We can give an equivalent statement of this theorem:

Theorem 3.12.2. Let A be a finitely generated abelian group. Then

$$A \cong \mathbb{Z}^n \times \mathbb{Z}/q_1 \times \cdots \mathbb{Z}/q_s,$$

for some $n, s \in \mathbb{N}$ and powers of prime numbers $q_1, \ldots, q_s$. Furthermore, the product on the right is unique up to permuting the factors.

To get Theorem 3.12.2 from Theorem 3.12.1, we should repeatedly apply Chinese remainder theorem as it appeared in an earlier homework exercise:

Exercise 3.12.3. Recall that $\mathbb{Z}/ab \cong \mathbb{Z}/a \times \mathbb{Z}/b$ if and only if $\gcd(a, b) = 1$. Use this isomorphism to derive Theorem 3.12.2 from Theorem 3.12.1.

Example 3.12.4. $\mathbb{Z}/2 \times \mathbb{Z}/6 \times \mathbb{Z}/24 \cong \mathbb{Z}/2 \times \mathbb{Z}/2 \times \mathbb{Z}/3 \times \mathbb{Z}/8 \times \mathbb{Z}/3$.

The following example shows the power of this theorem.

Example 3.12.5. Find all abelian groups of order 360 up to isomorphism. Let A be a group of order 360. Since $|A|$ is finite, no $\mathbb{Z}^n$ appears. Now take prime factorization of the order:

$$|A| = 360 = 2^3 \cdot 3^2 \cdot 5.$$

We see that we will always have a factor of $\mathbb{Z}/5$. For powers of 3, we either have $\mathbb{Z}/9$ or $\mathbb{Z}/3 \times \mathbb{Z}/3$. For powers of 2, we get three options. In total, we have the following six possibilities:

- $\mathbb{Z}/2 \times \mathbb{Z}/2 \times \mathbb{Z}/2 \times \mathbb{Z}/3 \times \mathbb{Z}/3 \times \mathbb{Z}/5$;
- $\mathbb{Z}/2 \times \mathbb{Z}/2 \times \mathbb{Z}/2 \times \mathbb{Z}/9 \times \mathbb{Z}/5$;
- $\mathbb{Z}/2 \times \mathbb{Z}/4 \times \mathbb{Z}/3 \times \mathbb{Z}/3 \times \mathbb{Z}/5$;
- $\mathbb{Z}/2 \times \mathbb{Z}/4 \times \mathbb{Z}/9 \times \mathbb{Z}/5$;
- $\mathbb{Z}/8 \times \mathbb{Z}/3 \times \mathbb{Z}/3 \times \mathbb{Z}/5$;
- $\mathbb{Z}/8 \times \mathbb{Z}/9 \times \mathbb{Z}/5$.

3.12.2. Classification of finite abelian groups.

Lemma 3.12.6. Consider two normal subgroups $N, M \trianglelefteq G$. If $G = NM$ and $N \cap M = \{e\}$, then $G \cong N \times M$.

In this lemma, we write NM to denote the subset $NM := \{nm \mid n \in N, m \in M\}$ of G.

Definition 3.12.7. Let A be an abelian group and let $N, M \leq A$ be subgroups. We write $A = N \oplus M$ and call A the *direct sum* of N and M if $N + M = A$ and $N \cap M = \{0\}$.

Notice that in an abelian group every subgroup is automatically normal, so by Lemma 3.12.6, we have $N \oplus M \cong N \times M$.

Lemma 3.12.8. Let p be a prime and A an abelian group with $|A| = p^n$, $n \geq 1$. Then:

(i) There exists an element of order p in A. In particular, there exists a subgroup of order p in A.

(ii) If there is only one subgroup $N \leq A$ of order p, then A is cyclic.

Lemma 3.12.9. Let p be a prime and A an abelian group with $|A| = p^n$, $n \geq 1$. Let $a \in A$ be an element of maximal order in A. Then there exists a subgroup $N \leq A$ such that $A = \langle a \rangle \oplus N$.

Lemma 3.12.10. Let p be a prime and A an abelian group with $|A| = p^n$, $n \geq 1$. Then

$$A \cong C_1 \oplus \cdots \oplus C_r,$$

where the C_i's are cyclic subgroups of A. In particular, we get

$$A \cong \mathbb{Z}/q_1 \oplus \cdots \oplus \mathbb{Z}/q_r$$

for $q_i = p^{m_i}$ for some $m_i \geq 1$.

Lemma 3.12.11. Let A be an abelian group such that for some natural number $n \geq 1$, we have

$$\forall a \in A \colon na = 0.$$

Suppose $n = rs$ for coprime $r, s \in \mathbb{N}$. Then $A \cong rA \oplus sA$.

Theorem 3.12.12. Let A be a finite abelian group, then

$$A \cong \mathbb{Z}/q_1 \times \cdots \times \mathbb{Z}/q_r,$$

where each q_i is some positive power of a prime number.

3.12.3. Free abelian groups.

Definition 3.12.13. We say that F is a *free abelian group* of rank $r \geq 1$ if there exists a finite subset $S = \{x_1, \ldots, x_r\} \subset F$ such that

$$\mathbb{Z}^r \to F,$$
$$(k_1, \ldots, k_r) \mapsto k_1 x_1 + \cdots + k_r x_r$$

is an isomorphism. In this case we say that S is a *basis* of F.

Lemma 3.12.14. Let $S = \{x_1, \ldots, x_r\}$ be a basis for a free abelian group F. Let $t \in \mathbb{Z}$. Then for $j \neq i$, the subset

$$T = \{x_1, \ldots, x_{j-1}, x_j + t x_i, x_{j+1}, \ldots, x_r\}$$

also is a basis of F.

Theorem 3.12.15 (Aligned bases theorem)**.** Let F be a free abelian group of rank r, and let N be a subgroup of F. Then N is a free abelian group of rank $s \leq r$. Furthermore, there exists a basis $\{x_1, \ldots, x_r\}$ of F and integers $d_1, \ldots, d_s \geq 1$ such that $d_i | d_{i+1}$ and $\{d_1 x_1, \ldots, d_s x_s\}$ is a basis of N.

3.13. Group action

3.13.1. Definition and examples.

Definition 3.13.1. Let X be a set and G be a group.

(i) A *left action* of G on X is a map

$$G \times X \to X$$
$$(g, x) \mapsto g \cdot x, \text{ or } gx$$

such that for all $x \in X$ and $g, h \in G$:
(a) $e \cdot x = x$;
(b) $g \cdot (h \cdot x) = (gh) \cdot x$.
In this case, we say that G *acts on* X *on the left*, or that X is a *left* G*-set*.

(ii) A *right action* of G on X is a map

$$X \times G \to X$$
$$(x, g) \mapsto x \star g, \text{ or } xg$$

such that for all $x \in X$ and $g, h \in G$:
(a) $x \star e = x$;
(b) $(x \star g) \star h = x \star (gh)$.
In this case, we say that G *acts on* X *on the right*, or that X is a *right* G*-set*.

Lemma 3.13.2. There is a correspondence between left and right G-sets.

Proof. Take a left action $G \times X \to X$, and define the corresponding right action $X \times G \to X$ as $x \star g \coloneqq g^{-1} \cdot x$. □

Example 3.13.3. Trivial action: $g \cdot x = x$ for any $x \in X$.

Example 3.13.4. G acts on itself by multiplication. Any subgroup $H \leq G$ acts on G by multiplication.

Example 3.13.5. Let X be a set, then $\mathfrak{S}_X \times X \to X$ is an action.

Example 3.13.6. G acts on itself by conjugation. If $N \trianglelefteq G$, then G acts on N by conjugation.

Example 3.13.7. If k is a field and V is a vector space, then in particular V is a $(k, \cdot)^\times$-set.

Example 3.13.8. $(\mathrm{Mat}_{n\times n}(\mathbb{R}), \cdot)^\times$ acts on $\mathbb{R}^n$.

Example 3.13.9. If X is a G-set and $H \leq G$, then X is also naturally an H-set.

Example 3.13.10. Take $H \leq G$, so G/H is a set, but not a group in general. Then we have a left action:

$$G \times G/H \to G/H,$$
$$(g, aH) \mapsto gaH,$$

as well as a right action:

$$H \setminus G \times G \to H \setminus G,$$
$$(Ha, g) \mapsto Hag.$$

3.13.2. Properties.

Theorem 3.13.11. Fix a group G and a set X. Then we have a natural bijection:

$$\{\text{left } G\text{-actions on } X\} \longleftrightarrow \{\text{homomorphisms } G \to \mathfrak{S}_X\}$$

$$a : G \times X \to X \longmapsto \Phi_a \colon G \to \mathfrak{S}_X$$

Here $\Phi_a(g) = a_g$ and $a_g(x) = a(g,x) = g \cdot x$.

Before proving the theorem, let us observe how it is related to previously considered constructions.

Example 3.13.12. If $X = G$ and a is multiplication (the group operation), then under the bijection we obtain the map $\Phi_a = \Phi \colon G \to \mathfrak{S}_G$ from Cayley's embedding theorem. In this case we know that Φ is injective.

Example 3.13.13. If $X = G$ and a is action by conjugation, then $\Phi_a \colon G \to \mathfrak{S}_G$ factors through $\operatorname{Inn} G \leq \operatorname{Aut} G$. In this case we know that $\operatorname{Ker} \Phi_a = Z(G)$ is the center of G.

Proof. Let X be a G-set. We first need to show that a_g is a bijection. The next thing to show is that Φ_a is a homomorphism.

Conversely, let us consider a homomorphism $\Phi \colon G \to \mathfrak{S}_X$. We need to show that $g \cdot x = \Phi(g)(x)$ defines a left action. □

Definition 3.13.14. We say that the action of G on X is *faithful* if the associated homomorphism $\Phi \colon G \to \mathfrak{S}_X$ is injective, or equivalently, if we have $g \in G$ such that $\forall x \in X \colon g \cdot x = x$, then $g = e$.

Example 3.13.15. G acting on itself by multiplication is faithful by Cayley's embedding theorem.

Example 3.13.16. G acting on itself by conjugation may be not faithful, for example if G is a nontrivial abelian group.

Definition 3.13.17. Let X be a G-set with action $a \colon G \times X \to X$. The *stabilizer* (also called *isotropy*) of $x \in X$ in G is:

$$\begin{aligned} Z_G(x) &\coloneqq \{g \in G \mid g \cdot x = x\} \\ &\supseteq \operatorname{Ker} \Phi_a \leq G. \end{aligned}$$

Exercise 3.13.18. $Z_G(x)$ is a subgroup of G for all $x \in X$.

Exercise 3.13.19. $\operatorname{Ker} \Phi_a = \bigcap\limits_{x \in X} Z_G(x)$.

3.13.3. Equivalence relation induced by an action.

Definition 3.13.20. Let X be a G-set. We say $x \sim y$ in X if $\exists g \in G \colon g \cdot x = y$.

Exercise 3.13.21. This defines an equivalence relation on X.

Definition 3.13.22. The cells of this partition are called the *orbits* of X under G. For $x \in X$, we denote the orbit of x by

$$G \cdot x = Gx = \{g \cdot x \mid g \in G\}.$$

Example 3.13.23. $H \leq G$, then G is a right H-set, with the action $G \times H \to G$ defined by $(g, h) \mapsto gh$. The orbits of this action are of the form gH, hence we see that the equivalence relation induced by the action coincides with the partition into left cosets.

Definition 3.13.24. If there is only one orbit, that is $G \cdot x = X$ for some $x \in X$, we say that the action is *transitive*. In other words, for any pair $x, y \in X$, there exists $g \in G$ such that $g \cdot x = y$.

3.13.4. Class equation.

Definition 3.13.25. Let X, Y be G-sets with actions $a\colon G \times X \to X$ and $b\colon G \times Y \to Y$. We say that a set map $\varphi\colon X \to Y$ is a *morphism of G-sets* if for any $x \in X$ and $g \in G$ we have:

$$\varphi(g \cdot x) = g \cdot \varphi(x).$$

Equivalently, the following diagram should be commutative:

$$\begin{array}{ccc} G \times X & \xrightarrow{a} & X \\ \downarrow{\scriptstyle \mathrm{id}_G \times \varphi} & & \downarrow{\scriptstyle \varphi} \\ G \times Y & \xrightarrow{b} & Y \end{array}$$

Definition 3.13.26. If a morphism of G-sets $\varphi\colon X \to Y$ is a bijection, we say that φ is an *isomorphism of G-sets*, and in this case we write $X \cong Y$.

Exercise 3.13.27. A morphism of G-sets $\varphi\colon X \to Y$ is an isomorphism if and only if there exists a morphism of G-sets $\psi\colon Y \to X$ such that $\varphi\psi = \mathrm{id}_Y$ and $\psi\varphi = \mathrm{id}_X$.

Lemma 3.13.28. Let X be a G-set. Then for every $x \in X$ there is an isomorphism of G-sets

$$G/Z_G(x) \cong G \cdot x.$$

In particular, $[G : Z_G(x)] = |G \cdot x|$, and if G is finite, then $|G \cdot x|$ divides $|G|$.

Definition 3.13.29. If $x \in X$ is such that $g \cdot x = x$ for all $g \in G$, then we say that x is a *fixed point* of the given G-action on X. We will denote the set of fixed points by

$$X^f = X^G = \{x \in X \mid \forall g \in G\colon g \cdot x = x\}$$

Theorem 3.13.30 (Class equation)**.** Let X be a finite G-set and let $x_1, \ldots, x_n$ be a set of representatives for non-singleton orbits. Then

$$|X| = |X^G| + \sum_{i=1}^{n} [G : Z_G(x_i)].$$

3.14. Sylow theorems

Sylow theorems have involved proofs, but provide a powerful tool for understanding the structure of finite groups. In particular, it will allow us to classify all finite groups of order at most 15.

3.14.1. p-subgroups.

Lemma 3.14.1. If G is a group of order $|G| = p^n$ for a prime p and $n \geq 1$, then there exists an element of order p. Consequently, there exists a subgroup in G of order p.

Lemma 3.14.2. If G is a group of order $|G| = p^n$ for a prime p and $n \geq 1$, and if X is a finite G-set, then

$$|X| \equiv |X^G| \mod p.$$

Proof. The proof is the direct application of the class equation from Theorem 3.13.30 and Lagrange's Theorem 3.8.13. More precisely, if x_i is not a fixed point, then $Z_G(x_i)$ has index greater than 1 in G, so by Lagrange's Theorem, the part $\sum_{i=1}^{n}[G : Z_G(x_i)]$ in the class equation is divisible by p. Hence $|X| - |X^G| = \sum_{i=1}^{n}[G : Z_G(x_i)]$ is divisible by p, which proves the result. □

Theorem 3.14.3 (Cauchy's theorem)**.** Let G be a finite group of order divisible by a prime number p. Then there is an element in G of order p. Consequently, G has a subgroup of order p.

Proof. We will apply Lemma 3.14.2 to a certain set X of order $|G|^{p-1}$ and the group $\mathbb{Z}/p$ acting on X. We will set it up in such a way that the fixed points will correspond to elements in G of order p.

Define the set $X = \left\{(g_1, \ldots, g_p) \in G^{\times p} \mid g_1 g_2 \cdots g_p = e\right\}$ with the $\mathbb{Z}/p$-action of rotating our sequence of elements of G to the left:

$$\mathbb{Z}/p \times X \longrightarrow X,$$
$$(a, (g_1, \ldots, g_p)) \longmapsto (g_{1+a}, \ldots, g_p, g_1, \ldots, g_a).$$

Exercise 3.14.4. Verify that this is an action.

We now observe that G^{p-1} is bijective to X via the assignment

$$f : G^{p-1} \longrightarrow X,$$
$$\left(g_1, \ldots, g_{p-1}\right) \longmapsto \left(g_1, \ldots, g_{p-1}, (g_1 \cdots g_{p-1})^{-1}\right).$$

Exercise 3.14.5. Check that f is a bijection.

Further, notice that $X^{\mathbb{Z}/p}$ consists of all the elements $(g, \ldots, g)$, which, by definition of X, must satisfy $g^p = e$. For example, $(e, \ldots, e) \in X$, so $|X^{\mathbb{Z}/p}| > 0$. But by Lemma 3.14.2, $|X^{\mathbb{Z}/p}| \equiv |X| \equiv |G|^{p-1} \mod p$, and since we assumed that $|G|$ is divisible by p, we conclude that the number of fixed points $|X^{\mathbb{Z}/p}|$ is, too. Together with the earlier observation $|X^{\mathbb{Z}/p}| > 0$, we get that $|X^{\mathbb{Z}/p}| \geq p$; in particular, there exists an element $g \neq e$ in G of order p.

□

Lemma 3.14.6. Let G be a finite group and p be a prime. Then the following are equivalent:

(i) there is $n \in \mathbb{N}$ such that $|G| = p^n$;

(ii) $\forall g \in G\, \exists r \in \mathbb{N}\colon |g| = p^r$.

Definition 3.14.7. A group G is called a *p-group* if for any element $g \in G$ there exists $r \in \mathbb{N}$ such that the order of this element is $|g| = p^r$. A subgroup $H \leq G$ is called a *p-subgroup of G* if H is itself a p-group.

With this definition, we can reformulate Lemma 3.14.6 as follows: a group G is a p-group if and only if $|G| = p^n$ for some $n \in \mathbb{N}$.

Definition 3.14.8. A *Sylow p-subgroup* of a group G is a maximal p-subgroup of G.

3.14.2. Normalizer. Given a group G, let us denote by $\mathcal{S}_G$ the set of all subgroups of G:

$$\mathcal{S}_G = \{H \mid H \leq G\}.$$

We endow $\mathcal{S}_G$ with a G-action by conjugation:

$$\begin{aligned} G \times \mathcal{S}_G &\to \mathcal{S}_G \\ (g, H) &\mapsto gHg^{-1} \end{aligned}$$

Definition 3.14.9. Given $H \in \mathcal{S}_G$, the stabilizer $Z_G(H)$ under this action is called the *normalizer* of H in G and is denoted by

$$N_G(H) \coloneqq Z_G(H) = \left\{g \in G \,\middle|\, gHg^{-1} = H\right\}.$$

Exercise 3.14.10. **(i)** $H \trianglelefteq N_G(H)$;
(ii) $N_G(H)$ is the largest subgroup of G in which H is normal.

Lemma 3.14.11. If H is a finite subgroup of an arbitrary group G and $g \in G$ is such that $gHg^{-1} \subset H$, then $g \in N_G(H)$.

Proof. The conjugation homomorphism $\varphi\colon H \to H, h \mapsto ghg^{-1}$ is well-defined by assumption, and injective. Since H is finite, it must also be surjective. By the construction, we have $\operatorname{Im}\varphi = gHg^{-1}$; so surjectivity implies that $H = \operatorname{Im}\varphi = gHg^{-1}$. This verifies that g is in $N_G(H)$. □

Lemma 3.14.12. Let H be a p-subgroup of a finite group G, then

$$[N_G(H) : H] \equiv [G : H] \mod p.$$

Proof. Apply Lemma 3.14.2 to the action

$$\begin{aligned} H \times G/H &\longrightarrow G/H, \\ (h, gH) &\longmapsto hgH. \end{aligned}$$

□

3.14.3. Sylow theorems.

Theorem 3.14.13 (First Sylow theorem)**.** Let G be a finite group of order $|G| = p^n \cdot m$, where p is prime, $n \geq 1$ and $\gcd(p, m) = 1$. Then each Sylow p-subgroup of G is of order p^n.

Proof. The idea is to use induction to show that for each r from 1 to n, there is a p-subgroup of G of order p^r.

Exercise 3.14.14. Prove the base of induction: $r = 1$.

Now suppose that we have a subgroup $H \leq G$ of order p^r, $1 \leq r < n$. We can use Lemma 3.14.12 to prove that $[N_G(H) : H]$ is divisible by p.

Observe that H is a normal subgroup of its normalizer, hence $N_G(H)/H$ is a group whose order is divisible by p. By Cauchy's Theorem 3.14.3, the quotient group has a subgroup $K \leq N_G(H)/H$ of order p. We can lift it to a subgroup of $N_G(H) \leq G$ of order p^{r+1} and containing H. □

Definition 3.14.15. Let $H \leq G$ be a subgroup and $g \in G$ be an element of a group G. Then gHg^{-1} is called the *conjugate* of H by g. If $H' \leq G$ is a subgroup for which there exists $g \in G$ such that $H' = gHg^{-1}$, then we say that the subgroups H and H' are *conjugate*.

Exercise 3.14.16. If G is a finite group and $P \leq G$ is a Sylow p-subgroup, then for any $g \in G$, the conjugate gPg^{-1} is also a Sylow p-subgroup of G.

Theorem 3.14.17 (Second Sylow theorem)**.** Let G be a finite group of order $|G|$ divisible by a prime p. Then any two Sylow p-subgroups of G are conjugate.

Proof. Take two Sylow p-subgroups P_1, P_2 of G. We have the following action:

$$P_2 \times G/P_1 \longrightarrow G/P_1,$$
$$(y, xP_1) \longmapsto yxP_1.$$

By Lemma 3.14.2 and since $|G/P_1|$ is coprime with p, the set of fixed points $(G/P_1)^{P_2}$ is non-empty.

Pick $xP_1 \in (G/P_1)^{P_2}$, then $xP_1x^{-1} = P_2$. □

Corollary 3.14.18. Let G be a finite group of order $|G|$ divisible by a prime p. Then a Sylow p-subgroup is unique if and only if it is normal in G.

Theorem 3.14.19 (Third Sylow theorem)**.** Let G be a finite group of order $|G| = p^n \cdot m$, where p is prime, $n \geq 1$ and $\gcd(p, m) = 1$. Let n_p be the number of distinct Sylow p-subgroups in G. Then:

- **(i)** $n_p \equiv 1 \mod p$;
- **(ii)** $n_p | m$;
- **(iii)** $n_p = [G : N_G(P)]$ for any Sylow p-subgroup P of G.

Proof. Let P be a Sylow p-subgroup and let X be the set of all Sylow p-subgroups of G, so $n_p = |X|$. We have the action of P on X by conjugation.

Exercise 3.14.20. If $T \in X^P$, prove that P is a subgroup of the normalizer $N_G(T)$ of T.

So if $T \in X^P$, then P is a Sylow p-subgroup of $N_G(T)$. But since T is a normal Sylow p-subgroup of $N_G(T)$, we conclude by Corollary 3.14.18 that $T = P$, hence $X^P = \{P\}$ and $|X^P| = 1$. Now Lemma 3.14.2 implies part (i) of the theorem.

For the remaining two parts, consider the action $G \times X \to X$ by conjugation.

Exercise 3.14.21. Use the second Sylow theorem to conclude that this action only has 1 orbit.

Now we can use Lemma 3.13.28 to prove part (iii). This also shows that $n_p | p^n \cdot m$. From part (i), we derive that n_p and p are coprime, which then implies part (ii). □

3.14.4. Examples of using Sylow theorems.

Example 3.14.22. Let us prove that there are no simple groups of order 15. In other words, we should prove that every group of order 15 contains a normal subgroup.

Let G denote a group of order $15 = 3 \cdot 5$. Then $n_5 \equiv 1 \mod 5$ and $n_5|3$, hence $n_5 = 1$. So there exists a unique Sylow 5-subgroup of G, and this subgroup is necessarily normal.

Exercise 3.14.23. Let p and q be distinct prime numbers. Prove that there are no simple groups of order pq.

Example 3.14.24. Let us prove that there are no simple groups of order 56.

Let G denote a group of order $56 = 2^3 \cdot 7$. Then $n_7 \equiv 1 \mod 7$ and $n_7|8$. So we get two cases: when $n_7 = 1$ and when $n_7 = 8$. In the former case, the Sylow 7-subgroup is unique and hence normal, so we should only study the case $n_7 = 8$. So we have 8 distinct subgroups $P_1, \dots, P_8$ of G of order 7, these subgroups are necessarily cyclic and simple because the order is prime. Hence the intersection of any two is trivial $\{e\}$. The set of all elements of order 7 is the union of $P_i \setminus \{e\}$, and since these subgroups intersect trivially, the union is disjoint, and we get that the number of elements of order 7 is

$$|P_1 \setminus \{e\} \cup \dots \cup P_8 \setminus \{e\}| = |P_1 \setminus \{e\}| + \dots + |P_8 \setminus \{e\}| = 8 \cdot |P_i \setminus \{e\}| = 8 \cdot 6 = 48.$$

Then the number of elements whose order is neither 1 nor 7 is

$$|G \setminus (P_1 \cup \dots \cup P_8)| = 56 - 1 - 48 = 7.$$

It shows that there can be only one Sylow 2-subgroup in G, so this subgroup is normal.

Exercise 3.14.25. Classify all groups of order at most 15.

- We know all groups of order at most 5: $\{e\}$, $\mathbb{Z}/2$, $\mathbb{Z}/3$, $\mathbb{Z}/4$, $\mathbb{Z}/2 \times \mathbb{Z}/2$, $\mathbb{Z}/5$.
- For a group G of order $6 = 2 \cdot 3$, we know that there exists exactly one Sylow 3-subgroup P, and so it must be normal. So for any elements $g \in G$ and $x \in P \setminus \{e\}$, we will have $gxg^{-1} \in P$. Now, gxg^{-1} is not equal to e, so there are two options: either $gxg^{-1} = x$, in which case we get $G \cong \mathbb{Z}/6$, or $gxg^{-1} = x^2$, in which case we get $G \cong \mathfrak{S}_3$.
- For each of the prime values $p = 7, 11, 13$, there is only one group $\mathbb{Z}/p$ up to isomorphism.
- $\dots$

CHAPTER 4

Rings

4.1. ⋆ Monoids revisited

In the last chapter, we focused a lot on groups, but many concepts can be generalized to monoids (Section 3.2).

Definition 4.1.1. Let $(M, *, e_M)$ and $(N, \star, e_N)$ be monoids. We say a set map $f\colon M \to N$ is a *monoid homomorphism* (or a *homomorphism of monoids*, or just a *homomorphism*) if:

(i) $\forall a, b \in M\colon f(a * b) = f(a) \star f(b)$;
(ii) $f(e_M) = e_N$.

We say that f is a *monoid isomorphism* if f is bijective.

Therefore, a group homomorphism is just a monoid homomorphism between groups. Notice though that in the case of group homomorphisms, we did not require $f(e_M) = e_N$ as it was automatic (see Proposition 3.6.9). In fact we have the following.

Proposition 4.1.2. Let $(M, *, e_M)$ be a monoid. Let $(G, \star, e_G)$ be a group. Let $f : M \to G$ be a set map that satisfy **(i)** above. Then **(ii)** holds automatically.

Proof. We have $f(e_M) = f(e_M * e_M) = f(e_M) * f(e_M)$. Multiplying by $f(e_M)^{-1} \in G$ on each side, we obtain $e_G = f(e_M)$. □

Definition 4.1.3. Let $(M, *, e_M)$ be a monoid. Define its *opposite monoid* $M^{\mathsf{op}} = (M, *^{\mathsf{op}}, e_M)$ as follows. For any $a, b \in M$, we have $m *^{\mathsf{op}} n := n * m$. This definition can be carried over to define *opposite groups*.

Proposition 4.1.4. $M = M^{\mathsf{op}}$ if and only if M is a commutative monoid.

Proposition 4.1.5. Let G be a group. Then $G \cong G^{\mathsf{op}}$.

Proof. Define the desired isomorphism by:

$$\begin{aligned} \varphi\colon G &\longrightarrow G^{\mathsf{op}} \\ g &\longmapsto g^{-1}. \end{aligned}$$

We can check that it is a group homomorphism:

$$\varphi(g) *^{\mathsf{op}} \varphi(h) = \varphi(h) * \varphi(g) = h^{-1}g^{-1} = (gh)^{-1} = \varphi(gh). \qquad \square$$

Remark 4.1.6. The previous proposition is not true for monoids. For example, consider the monoid $(\operatorname{Hom}\mathsf{Set}X, X, \circ, \mathrm{id}_X)$, where X is a set.

Definition 4.1.7. A *submonoid* N of a monoid $(M, *, e_M)$ is a subset $N \subseteq M$ such that:

(i) if $n, n' \in N$, then $n * n' \in N$;
(ii) $e_M \in N$.

One could define quotient of a monoid by a submonoid, but greater care is needed and we shall not discuss this.

We now generalize the definition of monoids.

Definition 4.1.8. A *monoid with many objects* $\mathcal{M}$ consists of the following data.

(i) A set $\mathrm{Ob}(\mathcal{M})$. Its elements are called objects of $\mathcal{M}$. We write $X \in \mathcal{M}$ instead of $X \in \mathrm{Ob}(\mathcal{M})$.
(ii) For each ordered pair (X, Y) of objects in $\mathcal{M}$, we have a set $\mathcal{M}(X, Y)$.

(iii) Given objects X, Y, Z in $\mathcal{M}$, we have a function called the *composing map*:

$$\begin{aligned}\mathcal{M}(Y,Z)\times\mathcal{M}(X,Y) &\longrightarrow \mathcal{M}(X,Z)\\ (m,n) &\longmapsto m*n.\end{aligned}$$

The above data is subject to the following axioms.

Associativity: Given $m\in\mathcal{M}(X,Y)$, $n\in\mathcal{M}(Y,Z)$, and $q\in\mathcal{M}(Z,W)$, we require:

$$(q*n)*m = q*(n*m).$$

Unitality: For each object $X\in\mathcal{M}$, there exists a element $\mathrm{id}_X\in\mathcal{M}(X,X)$, called the *identity element on* X, such that:

- $\mathrm{id}_X * m = m$, for all $m\in\mathcal{M}(Y,X)$;
- $n*\mathrm{id}_X = n$, for all maps $n\in\mathcal{M}(X,Y)$.

A monoid is precisely a monoid with many objects with one object, hence the justification of the name, as we gather from the example and exercise below.

Example 4.1.9. Let $(M,\star,e_M)$ be a monoid. We associate to M a monoid with many objects denoted BM, defined as follows.

- $\mathrm{Ob}(BM)=\{\cdot\}$. There is only one object and we can name it anyway we want, this is a place holder.
- $BM(\cdot,\cdot)=M$.
- The map $BM(\cdot,\cdot)\times BM(\cdot,\cdot)\to BM(\cdot,\cdot)$ is $M\times M\xrightarrow{*} M$.
- $\mathrm{id}_\cdot\in BM(\cdot,\cdot)$ is e_M.

Exercise 4.1.10. If $\mathcal{M}$ is a monoid with many objects, show that for any object X of $\mathcal{M}$, we get that $(\mathcal{M}(X,X),*,\mathrm{id}_X)$ is a monoid. In particular, if $\mathcal{M}$ has only one object, then we can associate only one monoid in this fashion.

Here is a well-known object from linear algebra that is a monoid with many objects.

Example 4.1.11. Let Mat be the monoid with many objects defined as follows.

- $\mathrm{Ob}(\mathsf{Mat})=\mathbb{N}$.
- Given $n,m\in\mathbb{N}$, define $\mathsf{Mat}(m,n)=\mathcal{M}_{n\times m}(\mathbb{R})$ to be the set of matrices with n-rows and m-columns with coefficients in $\mathbb{R}$.
- Given $n,m,s\in\mathbb{N}$, the map $\mathsf{Mat}(n,s)\times\mathsf{Mat}(m,n)\to\mathsf{Mat}(m,s)$ is defined by the usual matrix multiplication:

$$\begin{aligned}\mathcal{M}_{s\times n}(\mathbb{R})\times\mathcal{M}_{n\times m}(\mathbb{R}) &\longrightarrow \mathcal{M}_{s\times m}(\mathbb{R}),\\ (A,B) &\longmapsto AB.\end{aligned}$$

- The identity element on $n\in\mathbb{N}$ is the $n\times n$-identity matrix.

4.2. Rings and fields

4.2.1. Definition and examples. We now give the definition that generalizes the notion of a number system. The definition that we give is that of an associative unital ring, but in this course, all rings will be such, and we will refer to them as just rings for brevity.

Definition 4.2.1. A *ring* $R = (R, +, \cdot, 0_R, 1_R)$ is a set R endowed with two binary operations:

$$R \times R \xrightarrow{+} R, \qquad R \times R \xrightarrow{\cdot} R,$$

and two elements $0_R, 1_R \in R$ such that the following holds.

- **(i)** $(R, +, 0_R)$ is an abelian group.
- **(ii)** $(R, \cdot, 1_R)$ is a monoid. We often write ab as a shorthand for $a \cdot b$.
- **(iii)** (Distribution.) The two operations are compatible, namely, for all $a, b, c \in R$:

$$a(b + c) = ab + ac, \qquad (b + c)a = ba + ca,$$

If $ab = ba$ for all $a, b \in R$, we say that R is a *commutative* ring. We call 0_R the *zero of* R and 1_R the *unity* or *identity of* R.

Remark 4.2.2. Some authors do not require $(R, \cdot)$ to be unital. In this case, we shall refer to these rings as *non-unital rings*. We shall see that the datum of a non-unital ring is equivalent to that of a ring, or more precisely, the categories of non-unital rings and of rings are equivalent (see §8.1.1 for the definition).

Example 4.2.3. $R = \{0\}$ is a ring with trivially defined addition and multiplication. Here $0_R = 1_R = 0$, and $\{0\}$ is a commutative ring. This ring is called the *trivial ring* or the *zero ring*.

Example 4.2.4. $(\mathbb{Z}, +, \cdot, 0, 1)$ is a commutative ring. Similarly for $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$.

Example 4.2.5. For $n \geq 2$, $(\mathbb{Z}/n\mathbb{Z}, +, \cdot, [0], [1])$ is a commutative ring.

Example 4.2.6. Let $n \geq 1$, then square matrices $(\mathcal{M}_{n\times n}, +, \cdot, 0, I_n)$ form a ring. It is not commutative for $n > 1$.

Example 4.2.7. Polynomials of all degrees $\mathbb{R}[x] = \{a_0 + a_1x + \cdots + a_nx^n \mid n \geq 0, a_i \in \mathbb{R}\}$ form a ring, with usual addition and multiplication of polynomials. This ring is commutative.

Example 4.2.8. Generalizing the previous example, given a ring R, we can define the *polynomial ring* $R[x]$ with coefficients in R as follows. Its elements are polynomials in one variable x:

$$f = f(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + a_nx^n = \sum_{i=0}^{n} a_ix^i,$$

where $a_i \in R$ for each i. If in the expression above $a_n \neq 0_R$, then we say that f is a polynomial of *degree* n, and we write $\deg(f) = n$. Polynomials of degree 0 are also called *constant polynomials*, and may be identified with elements of R. By definition, two polynomials are equal exactly when they have equal degrees and

coefficients. We define addition and multiplication in the usual way, i.e., as follows:

$$R[x] \times R[x] \xrightarrow{+} R[x]$$
$$\left(\sum_{i=0}^{n} a_i x^i, \sum_{i=0}^{m} b_i x^i\right) \longmapsto \sum_{i=0}^{\max(n,m)} (a_i + b_i) x^i,$$

and:

$$R[x] \times R[x] \xrightarrow{\cdot} R[x]$$
$$\left(\sum_{i=0}^{n} a_i x^i, \sum_{i=0}^{m} b_i x^i\right) \longmapsto \sum_{i=0}^{n+m} \left(\sum_{j=0}^{i} a_j b_{i-j}\right) x^i$$

Its neutral elements for addition and multiplication are 0_R and 1_R, respectively, regarded as constant polynomials.

Example 4.2.9. If R is a ring and X is a nonempty set, then $\operatorname{Hom}\mathsf{Set}X, R$ is a ring with point-wise addition and multiplication.

Example 4.2.10. Let U be an open subset of $\mathbb{R}^n$, for example we can just take $U = \mathbb{R}$. We define rings of continuous $C^0(U)$, continuously differentiable $C^1(U)$, smooth $C^\infty(U)$ functions as certain subsets of $\operatorname{Hom}\mathsf{Set}U, \mathbb{R}$ with point-wise addition and multiplication.

4.2.2. Basic properties.

Proposition 4.2.11. Let R be a ring. Then $0 \cdot r = 0 = r \cdot 0$, for any $r \in R$.

Proof. We have $0 \cdot r + 0 \cdot r = (0 + 0)r = 0 \cdot r$. As $(R, +, 0)$ is a group, we can cancel the term $0 \cdot r$ on each side and obtain $0 \cdot r = 0$. □

Proposition 4.2.12. Let R be a ring. Then $1_R = 0_R$ if and only if $R = \{0\}$.

Proof. Suppose $1_R = 0_R$, and let $r \in R$. Then using the previous proposition:

$$r = r \cdot 1_R = r \cdot 0_R = 0_R$$

Thus $r = 0_R$. □

Notation 4.2.13. For $n \in \mathbb{N}$, the notation nr means $\underbrace{r + \cdots + r}_{n \text{ times}}$, while r^n means $\underbrace{r \cdot \cdots \cdot r}_{n \text{ times}}$.

Notation 4.2.14. We denote by $-r$ the inverse of r in $(R, +, 0_R)$.

Exercise 4.2.15. Prove that $(-1_R) \cdot r = -r$ by distribution.

Notation 4.2.16. If we have a negative integer $n < 0$, we define $nr = (-n)(-r)$.

Notation 4.2.17. For any ring R, we identify $n \in \mathbb{Z}$ with the element $n \cdot 1_R \in R$. For example, $2 = 1_R + 1_R$, $3 = 1_R + 1_R + 1_R$, $0 = 0_R$, and so on. Thus any element of $\mathbb{Z}$ can be viewed as an element in R.

Example 4.2.18. In the ring $\mathcal{M}_3(\mathbb{R})$, the integer 2 is identified with the scalar matrix $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$.

⚠ **Warning 4.2.19.** The elements in $\mathbb{Z}$ don't have to be distinct when regarded in R. For instance, in the ring $\mathbb{Z}/3\mathbb{Z}$, we have $3 = 6 = 0$.

Proposition 4.2.20. Let R be a ring. Let $r, s \in R$. Then:

- $(-r)s = r(-s) = -(rs)$;
- $(-r)(-s) = rs$;
- $(kr)(\ell s) = (k\ell)(rs)$, for any $k, \ell \in \mathbb{Z}$.

Definition 4.2.21. The *characteristic* of a ring R, denoted by $\mathrm{Char}(R)$, is determined by the order of 1_R in $(R, +, 0_R)$ as follows:

$$\mathrm{Char}(R) = \begin{cases} 0, & \text{if } \mathrm{ord}(1_R) = \infty, \\ n, & \text{if } \mathrm{ord}(1_R) = n \geq 1. \end{cases}$$

Example 4.2.22. The characteristic of $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are zero. The rings $\mathbb{Z}/n\mathbb{Z}$ and $(\mathbb{Z}/n\mathbb{Z})[x]$ have characteristic n.

4.2.3. Units, fields.

Definition 4.2.23. The *units* of a ring R are the units of the monoid $(R, \cdot, 1_R)$. Recall that $u \in R$ is a unit if $\exists v \in R$ such that $uv = 1_R = vu$. We denote such a v by u^{-1}. We denote by $R^\times$ the groups of units.

Definition 4.2.24. A *field* is a nontrivial commutative ring in which every nonzero element is a unit (i.e. is invertible). In other words, R is a field if $R\backslash\{0\} = R^\times$.

Example 4.2.25. The rings $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are fields. The ring $\mathbb{Z}$ is not a field as $\mathbb{Z}^\times = \{-1, 1\}$.

Example 4.2.26. Recall $\mathbb{Z}/n\mathbb{Z}^\times = \{[k] \mid 0 \leq k < n, \gcd(k, n) = 1\}$. Thus $\mathbb{Z}/n\mathbb{Z}$ is a field if and only if n is a prime number.

Definition 4.2.27. A *skew-field* or *division ring* is a not necessarily commutative ring R in which $R\backslash\{0\} = R^\times$.

Example 4.2.28. Quaternions $\mathbb{H}$ form a division ring that is not a field. We define $\mathbb{H}$ as the $\mathbb{R}$-vector space with basis $\{1, i, j, k\}$ in which the multiplication is defined by as follows:

- $1_{\mathbb{H}} = 1 + 0i + 0j + 0k$;
- $i^2 = j^2 = -1_{\mathbb{H}}$ and $k = ij = -ji$;
- we extend to any element by requiring it to be associative and $\mathbb{R}$-bilinear.

It is not commutative, as $ij = -ji$. Here is an example of how to compute the product of two elements:

$$(i + j)i = i^2 + ji = -1 + ji = -1 - k.$$

A lot of the geometry of $\mathbb{C}$ can be extended to $\mathbb{H}$. Given $x = a + bi + cj + dk$, define the conjugate to be $\overline{x} = a - bi - cj - dk$. Then we get $\overline{\overline{x}} = x$. Moreover, define the norm $N \colon \mathbb{H} \to \mathbb{R}$ to be $N(x) = x\overline{x} = a^2 + b^2 + c^2 + d^2$. One can check by direct calculation that the inverse of x is $x^{-1} = \frac{1}{N(x)}\overline{x}$. Also, $N(x) = 0$ if and only if $x = 0$.

4.2.4. Zero divisors.

Definition 4.2.29. We say that an element $r \in R$ is a *zero divisor* if there exists $s \neq 0_R$ in R such that $sr = 0_R$ or $rs = 0_R$.

In other words, r is a zero divisor if and only if

$$\text{either} \quad s\cdot\colon R \to R,\ r \mapsto sr \qquad \text{or} \qquad \cdot s\colon R \to R,\ r \mapsto rs$$

is *not* an injective map.

Example 4.2.30. The element 0_R is a zero divisor in any nonzero ring R, and it is not a zero divisor if $R = \{0\}$.

Example 4.2.31. The element 2 is a zero divisor in $\mathbb{Z}/4\mathbb{Z}$.

Example 4.2.32. The only zero divisor in each of $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ is the zero element.

Proposition 4.2.33. In $\mathbb{Z}/n$, every element is either a unit or a zero divisor.

Proof. Let a be an element in $\mathbb{Z}/n$. The case $a = 0$ is Example 4.2.30, so now assume $a \neq 0$. Denote $d = \gcd(a, n)$. We prove that a is a zero divisor if and only if $1 < d < n$.

Suppose first $1 < d < n$, then define $b := \frac{n}{d}$. In this case, $b \neq 0$ but $ba = 0$.

Suppose now that a is a zero divisor. We have $d = xa + yn$ for some $x, y \in \mathbb{Z}$ by Bézout's identity (Theorem 2.1.13). Thus $xa = d$ in $\mathbb{Z}/n$. Since a is a zero divisor, we can fix $b \neq 0$ such that $ba = 0$. So $bd = bxa = 0$ in $\mathbb{Z}/n$, which implies $d \neq 1, n$. □

⚠ **Warning 4.2.34.** The last proposition is specific to $\mathbb{Z}/n\mathbb{Z}$. In other rings, there are nonzero elements that are neither units nor zero divisors.

Theorem 4.2.35. For a ring R, the following are equivalent.

(i) For any $a, b \in R$, if $ab = 0$, then $a = 0$ or $b = 0$.
(ii) For any $a, b, c \in R$, if $ab = ac$ and $a \neq 0$, then $b = c$.
(iii) For any $a, b, c \in R$, if $ba = ca$ and $a \neq 0$, then $b = c$.

Proof. (1) ⇒ (2): suppose $ab = ac$ and $a \neq 0$. Then:

$$\begin{aligned} & ab - ac = 0 \\ \Rightarrow\ & a(b - c) = 0 \\ \Rightarrow\ & b - c = 0 \\ \Rightarrow\ & b = c. \end{aligned}$$

(2) ⇒ (1): suppose $ab = 0$. If $a \neq 0$, then $ab = 0 = a \cdot 0$, so $b = 0$.

Therefore we have shown (1) ⇔ (2). We can show (1) ⇔ (3) in a similar way. □

The equivalent conditions in Theorem 4.2.35 are so fundamental in algebra that rings satisfying them have a special name, and we discuss them next.

4.2.5. Domains.

Definition 4.2.36. A nontrivial ring R is called a *domain* if 0_R is its only zero divisor, i.e.: $\forall a, b \in R$, if $ab = 0$ then $a = 0$ or $b = 0$. A commutative domain is called an *integral domain*.

By Theorem 4.2.35, one can cancel on left and the right in a domain.

Example 4.2.37. The ring $\mathbb{Z}$ is an integral domain.

Example 4.2.38. The ring $\mathbb{Z}/n\mathbb{Z}$ is a domain if and only if n is a prime number.

Proposition 4.2.39. Every field is an integral domain.

PROOF. Let $\mathbb{F}$ be a field. Let $ab = 0$ in $\mathbb{F}$. Suppose $a \neq 0$, then $b = a^{-1}ab = a^{-1}0 = 0$. Thus $b = 0$. □

Theorem 4.2.40. The characteristic of a domain is either zero or prime.

PROOF. Let R be a ring and suppose that $\mathrm{Char}(R) = n > 1$. Suppose n is not a prime, then $n = pq$ where $1 < p < n$ and $1 < q < n$. Therefore:

$$(p \cdot 1_R)(q \cdot 1_R) = (pq)(1_R \cdot 1_R) = n \cdot 1_r = 0.$$

As $p, q < n$ and n is minimal with the property that $n \cdot 1_R = 0$, we have that $p \cdot 1_R \neq 0$ and $q \cdot 1_R \neq 0$, so they are zero divisors. Thus R is not a domain. □

Theorem 4.2.41. A finite integral domain is a field.

PROOF. Let R be an integral domain. Pick a nonzero element $a \in R$. We want to prove that it has a multiplicative inverse. For that, consider the map

$$\begin{aligned} f \colon R &\longrightarrow R, \\ r &\longmapsto ar. \end{aligned}$$

Since R is an integral domain, the map f is injective. Since R is finite, we have an injective map from a finite set to a finite set of the same cardinality, hence by the pigeonhole principle, f is also surjective. But note that $1_R \in R = \mathrm{Im}\, f$, so there is some $b \in R$ such that $1_R = f(b) = ab$. Since R is commutative, we conclude that a is a unit of R. □

4.3. Ring homomorphisms and ideals

Just as when we were studying groups, we defined group homomorphisms, so with rings, the next step after defining them as objects is to define ring homomorphisms. In a sense that can be made precise (by the Yoneda lemma), maps to or from your chosen object can fully describe it. So in math, the proverb "tell me who your friends are and I will tell you who you are" is actually a proven fact.

4.3.1. Subrings. We first define the notion of a subring, which we later see is nothing but an injective homomorphism.

Definition 4.3.1. A *subring* S of a ring R is a subset of R such that the binary operations $+, \cdot \colon R \times R \to R$ restrict to $S \times S \to S$ in such a way that $(S, +, \cdot, 0_R, 1_R)$ is a ring.

Proposition 4.3.2. Let R be a ring and let $S \subseteq R$ be a subset. Then S is a subring of R if and only if:

- $0_R, 1_R \in S$;
- for any $s, t \in S$, we have $st, s + t, s - t \in S$.

Proof. Assume first that S is a subring. Then by definition, we have $0_R, 1_R \in S$, and the addition and multiplication in R restrict to S. The latter means that in particular, for $s, t \in S$, we have $st, s + t \in S$. To see that $s - t$ is in S, we first recall that $s - t$ means $s + (-t)$. Now, since S is a ring, addition defines an abelian group structure on it, so we have that the additive inverse $-t$ of t also is an element of S. But since S is closed under addition and $s, (-t) \in S$, we get $s - t = s + (-t) \in S$.

Conversely, assume that S satisfies the conditions listed in the bullet points. We will verify the axioms from Definition 4.2.1 for S. First, observe that the second bullet point ensures that we have two binary operations $+, \cdot$ on S. Since the operations were associative and distributive in R, they will automatically be such in S. The first bullet point tells us that both operations possess identities, so both $(S, +, 0_R)$ and $(S, \cdot, 1_R)$ define structures of a monoid on S. The last property to verify is that $(S, +, 0_R)$ is an abelian group. Again, since it is a submonoid of an abelian group $(R, +, 0_R)$, commutativity is automatic. To check that S possesses additive inverses, we take $s = 0_R$ in the second bullet point to see that for any $t \in S$, we have $s - t = 0_R - t = -t \in S$. □

Proposition 4.3.3. A subring of a domain is itself a domain.

Proof. Let R be an integral domain and let $S \subseteq R$ be a subring. We want to check that S is nontrivial and the only zero divisor in S is 0. Since R is a domain, it is not the zero ring, hence $0 \neq 1$ in R by Proposition 4.2.12. Since both of these elements must belong to the subring S, we have that S is not trivial either. Now assume that $s \in S$ is a zero divisor. Then there exists a nonzero $t \neq 0$ in S such that $st = 0$. But this equality also holds in R, and since R is a domain, we conclude that $s = 0$. □

Example 4.3.4. The following is a sequence of subrings: $\mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C} \subseteq \mathbb{H}$.

Example 4.3.5. A ring R is a subring of its polynomial ring $R[x]$, in which we view every element of R as a constant polynomial in $R[x]$.

Example 4.3.6. Define $\mathbb{H}_0 = \{a + bi + cj + dk \in \mathbb{H} \mid a, b, c, d \in \mathbb{Z}\}$. It is a subring of $\mathbb{H}$. Notice that if $x \in \mathbb{H}_0$ then $\overline{x} \in \mathbb{H}_0$ and $N(x) \in \mathbb{N}$ for any $x \in \mathbb{H}_0$. Thus one can see $x \in \mathbb{H}_0^\times$ if and only if $N(x) = 1$. In fact, we obtain the group isomorphism $\mathbb{H}_0^\times \cong Q_8$, the quaternion group.

4.3.2. Homomorphisms.

Definition 4.3.7. Let R and S be rings. We say that a set map $\varphi\colon R \to S$ is a *ring homomorphism*, or a *homomorphism of rings*, if φ is a group homomorphism with respect to addition, and a monoid homomorphism with respect to multiplication. In other words, we require for all $a, b \in R$:

(i) $\varphi(a + b) = \varphi(a) + \varphi(b)$;
(ii) $\varphi(ab) = \varphi(a)\varphi(b)$;
(iii) $\varphi(1_R) = 1_S$.

We denote by $\operatorname{Hom}_{\mathsf{Ring}}(R, S)$ the set of ring homomorphisms.

Exercise 4.3.8. Deduce from the axioms of a ring homomorphism that for all $r \in R$:

- $\varphi(0_R) = 0_S$;
- $\varphi(-r) = -\varphi(r)$;
- $\varphi(kr) = k\varphi(r)$, for $k \in \mathbb{Z}$;
- $\varphi(r^n) = \varphi(r)^n$, for $n \geq 0$.

Example 4.3.9. The trivial map:

$$\begin{aligned} R &\to S, \\ r &\mapsto 0_S, \end{aligned}$$

is not a ring homomorphism unless $S = \{0\}$, as otherwise it does not send 1_R to 1_S. In particular, there exists no ring homomorphism $\{0\} \to S$ unless $S = \{0\}$.

Example 4.3.10. The identity map $\mathrm{id}_R\colon R \to R$ is a ring homomorphism.

Definition 4.3.11. We say that a ring homomorphism $\varphi\colon R \to S$ is an *isomorphism* if there exists $\psi\colon S \to R$ such that $\varphi\psi = \mathrm{id}_S$ and $\psi\varphi = \mathrm{id}_R$. We write $R \cong S$ if such an isomorphism exists, and we then say that R and S are *isomorphic*.

⚠ **Warning 4.3.12.** The notation $\cong$ can be ambiguous: it can either mean group isomorphisms or ring isomorphisms. In general, we say in words which types of isomorphisms if there can be confusions. For instance $\mathbb{R} \times \mathbb{R}$ and $\mathbb{C}$ are isomorphic as abelian groups but they are not isomorphic as rings.

Exercise 4.3.13. Prove that $\varphi\colon R \to S$ is an isomorphism if and only if it is a bijection.

Example 4.3.14. If S is a subring of R, then the inclusion map $S \hookrightarrow R$ is a ring homomorphism.

Example 4.3.15. The quotient map $\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ is a ring homomorphism.

Proposition 4.3.16. Let $\varphi\colon R \to R'$ be a ring homomorphism. Let S be a subring of R. Then the image $\varphi(S)$ of S is a subring of R'.

Proof. Follows from Proposition 4.3.2. □

Example 4.3.17. Let S be a subring of a ring R. Let $\alpha \in R$ be a fixed element. Then the evaluation map

$$\begin{aligned} \mathrm{ev}_\alpha \colon S[x] &\longrightarrow R, \\ f(x) &\longmapsto f(\alpha), \end{aligned}$$

is a ring homomorphism (we shall prove this in Example 4.8.2). We denote by $S[\alpha]$ its image. It is the smallest ring in R containing both S and α.

Example 4.3.18. In the previous example, we can consider $i \in \mathbb{C}$ and polynomials in $\mathbb{Z}$:

$$\begin{aligned} \mathrm{ev}_i : \mathbb{Z}[x] &\longrightarrow \mathbb{C} \\ f(x) &\longmapsto f(i). \end{aligned}$$

Since $i^2 = -1$, we see that $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$. This is called the *Gaussian integers*. Since it is a subring of $\mathbb{C}$, it is an integral domain.

Recall that any element $k \in \mathbb{Z}$ can be regarded as an element in any ring R. This defines a set map $\mathbb{Z} \to R$.

Theorem 4.3.19. Let R be a ring. There exists a unique ring homomorphism $\mathbb{Z} \to R$. It is injective if and only if $\mathrm{Char}(R) = 0$.

Proof. The map $\varphi \colon \mathbb{Z} \to R$ is defined by our assignment $k \mapsto k \cdot 1_R$. This forces the map to be a group homomorphism. Moreover:

$$\begin{aligned} \varphi(k\ell) &= (k\ell)1_R \\ &= (k \cdot 1_R)(\ell \cdot 1_R) \\ &= \varphi(k)\varphi(\ell). \end{aligned}$$

Notice moreover that φ is entirely determined by the assignment $\varphi(1) = 1_R$ and is thus unique. □

Because of the previous theorem, we often refer to $\mathbb{Z}$ as the *initial ring*. It is the only ring that satisfy this property as the next exercise shows.

Exercise 4.3.20. Suppose S is a ring which satisfies the following property: given any ring R, there exists a unique ring homomorphism $S \to R$. Show that $S \cong \mathbb{Z}$ as rings.

Example 4.3.21. Let $\varphi \colon R \to S$ be ring homomorphism. Then it lifts to a map between polynomial rings still denoted φ and defined as follows:

$$\begin{aligned} \varphi \colon R[x] &\longrightarrow S[x] \\ \sum_{i=0}^{n} a_i x^i &\longmapsto \sum_{i=0}^{n} \varphi(a_i) x^i. \end{aligned}$$

We shall see in Example 4.8.2 that it is a ring homomorphism.

4.3.3. Ideals. We saw that if A is a subgroup of an abelian group G, then the quotient G/A is an abelian group. If we consider G to be $(R, +, 0_R)$, the underlying abelian group structure of a ring R, then $R/A = \{r + A \mid r \in R\}$ is an abelian group. What structure do we need on A to guarantee that R/A is a ring, and not just a group?

Definition 4.3.22. Let R be a ring. Let I be a subgroup of $(R, +, 0_R)$.

- We say I is a *left ideal* if:
$$\forall r \in R, \forall x \in I \Rightarrow rx \in I,$$
i.e. $RI \subseteq I$, where $RI = \{rx \mid r \in R,\, x \in I\}$.
- We say I is a *right ideal* if:
$$\forall r \in R, \forall x \in I \Rightarrow xr \in I,$$
i.e. $IR \subseteq I$.
- We say I is a *two-sided ideal*, or simply an *ideal*, if I is both a left and a right ideal.

If R is a commutative ring, the three notions coincide.

Example 4.3.23. $\{0_R\}$ and R are ideals of a ring R.

Definition 4.3.24. Let R be a ring. $\{0_R\}$ is called the *trivial ideal*. An ideal $I \subset R$ that does not coincide with R is called a *proper ideal*.

Example 4.3.25. The subgroup $n\mathbb{Z}$ is an ideal of $\mathbb{Z}$. In fact, we call it the *principal ideal generated by n*.

Example 4.3.26. Let R be a commutative ring. For any element $r \in R$, we define the *principal ideal* (r) as the set of elements
$$(r) = Rr \coloneqq \{ar \mid a \in R\}.$$
More generally, given several elements $r_1, \ldots, r_n \in R$, the ideal *generated* by these elements is
$$(r_1, \ldots, r_n) \coloneqq \{a_1 r_1 + \ldots a_n r_n \mid a_i \in R\}.$$

Example 4.3.27. Let $R = \mathcal{M}_2(\mathbb{R})$ be the non-commutative ring of 2×2-matrices with coefficient in $\mathbb{R}$. Then
$$\left\{ \begin{bmatrix} a & 0 \\ b & 0 \end{bmatrix} \middle| a, b \in \mathbb{R} \right\},$$
is a left ideal of R but not a right ideal. Similarly:
$$\left\{ \begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix} \middle| a, b \in \mathbb{R} \right\},$$
is a right ideal of R but not a left ideal. In fact, the only (two-sided) ideals of R are $\{0\}$ and R.

Just as we saw in group theory for normal subgroups, every kernel of a ring homomorphism is an ideal.

Theorem 4.3.28. Let $\varphi \colon R \to S$ be a ring homomorphism. Then the kernel of φ
$$\mathrm{Ker}(\varphi) = \{r \in R \mid \varphi(r) = 0_S\},$$
is an ideal of R. More generally, if $J \subseteq S$ is an ideal of S, then the preimage $\varphi^{-1}(J)$ is an ideal of R.

Proof. We know that $\mathrm{Ker}(\varphi)$ is a subgroup. Let $r \in R$, and $x \in \mathrm{Ker}(\varphi)$. Then:
$$\varphi(rx) = \varphi(r)\varphi(x) = f(r) \cdot 0_S = 0_S.$$
Hence $rx \in \mathrm{Ker}(\varphi)$. Similarly, we can show $xr \in \mathrm{Ker}(\varphi)$. Thus $\mathrm{Ker}(\varphi)$ is an ideal of R. We leave showing that $\varphi^{-1}(J)$ is an ideal of R as an exercise. □

Now, the analogy with groups is not full: we saw that the image of a subgroup is a subgroup; however, the image of an ideal is not necessarily an ideal.

Example 4.3.29. Consider the injective homomorphism $\varphi\colon \mathbb{Z} \to \mathbb{Q}$. We know that $I = \mathbb{Z}$ is an ideal in $\mathbb{Z}$; however, $\varphi(I) = \mathbb{Z} \subset \mathbb{Q}$ is not an ideal in $\mathbb{Q}$, since, e.g., $\frac{1}{2} \cdot 1 = \frac{1}{2} \notin \varphi(I)$.

Proposition 4.3.30. Let $\varphi\colon R \to S$ be a surjective ring homomorphism and $I \subseteq R$ an ideal. Then the image $\varphi(I)$ is an ideal of S.

Proof. Suppose $y \in \varphi(I)$ and $s \in S$. Then $\exists x \in I\colon \varphi(x) = y$. As φ is surjective, there exists $r \in R$ such that $\varphi(r) = s$. So $\varphi(rx) = sy$. Since I is an ideal, we know that $rx \in I$. Therefore $sy \in \varphi(I)$. Similarly, we can show that $ys \in \varphi(I)$, so $\varphi(I)$ is an ideal. □

⚠ **Warning 4.3.31.** The requirement that the ring homomorphism be surjective is important in the last result, as seen, e.g., in Example 4.3.29.

4.3.4. Quotient rings. We will now prove that an additive subgroup of a ring is an ideal if and only if the quotient by it inherits the structure of a ring. This furthers our analogy between ideals in rings and normal subgroups in groups.

Theorem 4.3.32. Let R be a ring. Let I be a subgroup of $(R, +, 0_R)$. Then I is an ideal of R if and only if the assignment

$$\begin{aligned} R/I \times R/I &\longrightarrow R/I \\ (r + I, s + I) &\longmapsto (rs) + I, \end{aligned}$$

is well-defined. In this case $(R/I, +, \cdot, I, 1_R + I)$ is a ring, and the quotient map $\gamma\colon R \to R/I$ is a ring homomorphism.

Proof. We start with the backwards direction. Suppose the multiplication on R/I is well-defined. Let $r \in R$ and $x \in I$. Recall that $x + I = I = 0 + I$ is the identity element for the addition in R/I. Thus we obtain:

$$\begin{aligned} (rx) + I &= (r + I)(x + I) \\ &= (r + I)(0 + I) \\ &= (r \cdot 0) + I \\ &= 0 + I \\ &= I. \end{aligned}$$

Therefore we obtain $rx \in I$. We can show similarly that $xr \in I$ and thus I is an ideal.

For the forward implication, suppose I is an ideal. Suppose $r + I = r' + I$ and $s + I = s' + I$. We need to show that:

$$(rs) + I = (r's') + I.$$

We know that $r - r' \in I$ and $s - s' \in I$. Thus:

$$rs - r's' = \underbrace{r(s - s')}_{\in R(s-s') \subseteq I} + \underbrace{(r - r')s'}_{\in (r-r')R \subseteq I}.$$

Hence $rs - r's' \in I$, and so $rs + I = r's' + I$. Thus the product is well-defined. This verifies that R/I is a ring, and one checks in a straightforward way that γ is a ring homomorphism. □

Example 4.3.33. This gives another proof that $\mathbb{Z}/n\mathbb{Z}$ is a ring, as $n\mathbb{Z} \subseteq \mathbb{Z}$ is an ideal.

Definition 4.3.34. Let I be an ideal of R. We refer to R/I as the *quotient ring* of R by I.

Theorem 4.3.35 (Universal property of quotient rings)**.** Let $\varphi\colon R \to S$ be a ring homomorphism. Let I be an ideal of R such that $I \subseteq \mathrm{Ker}(\varphi)$. Then, there exists a unique ring homomorphism $\overline{\varphi}\colon R/I \to S$ such that $\overline{\varphi} \circ \gamma = \varphi$.

$$\begin{array}{ccc} R & \xrightarrow{\varphi} & S \\ {\scriptstyle\gamma}\downarrow & \nearrow_{\exists!\overline{\varphi}} & \\ R/I & & \end{array}$$

Proof. Forgetting about the multiplication structures for a moment, the universal property of group quotients already guarantees that there exists $\overline{\varphi}\colon R/I \to S$ such that $\overline{\varphi} \circ \gamma = \varphi$. In other words $\overline{\varphi}(r + I) = \varphi(r)$. Therefore, we only need to check that $\overline{\varphi}$ is a ring homomorphism. We have:

$$\begin{aligned} \overline{\varphi}(1_{R/I}) &= \overline{\varphi}(1_R + I) \\ &= \varphi(1_R) \\ &= 1_S, \end{aligned}$$

as φ is a ring homomorphism. We also have:

$$\begin{aligned} \overline{\varphi}\Big((r + I)(r' + I)\Big) &= \overline{\varphi}(rr' + I) \\ &= \varphi(rr') \\ &= \varphi(r)\varphi(r'), \quad \text{as } \varphi \text{ is a ring homomorphism,} \\ &= \overline{\varphi}(r + I)\overline{\varphi}(r' + I), \end{aligned}$$

for any $r, r' \in R$. Thus $\overline{\varphi}$ is a ring homomorphism. □

4.3.5. Constructing new rings. We have already seen several ways of constructing new rings from old ones. Let us summarize them below, and give names to some ways of combining several constructions.

4.3.5.1. *Subrings.* Given a ring R, we can construct a new ring by picking out a subset of elements $S \subseteq R$ that is a subring.

4.3.5.2. *Quotient rings.* Given a ring R and an ideal $I \trianglelefteq R$, we proved that the quotient group R/I is a ring.

4.3.5.3. *Rings of polynomials in several variables.* Given a ring R, we have already discussed what the ring of polynomials $R[x]$ with coefficients in R is: its elements are combinations $A(x) = a_0 + a_1x + \cdots + a_nx^n$ with $a_i \in R$ and $n \in \mathbb{N}$, and two elements $A(x)$ and $B(x)$ are equal if and only all of their coefficients coincide, i.e., $a_i = b_i$ for all $i \in \mathbb{N}$. Its properties, including the universal property, are discussed in more detail in Section 4.8.

Now we want to define polynomials in several variables. Roughly speaking, we think of them as finite R-linear combinations of monomials, e.g.,

$$a_0 + a_1x_1 + a_2x_3 + a_3x_1x_2x_3^5,$$

with $a_i \in R$. To define the set $R[x_1, \ldots, x_n]$ in a more concise way, we denote by x the tuple $(x_1, \ldots, x_n)$, and for any tuple of natural numbers $\delta = (\delta_1, \ldots, \delta_n) \in \mathbb{N}^n$,

we define the *monomial of multidegree δ in variables $x_1, \ldots, x_n$* as

$$x^\delta \coloneqq x_1^{\delta_1} x_2^{\delta_2} \cdots x_n^{\delta_n}.$$

Then the *ring of polynomials in n variables* is defined as the set

$$R[x_1, \ldots, x_n] \coloneqq \left\{ A(x) = \sum_{\delta \in \mathbb{N}^n} a_\delta x^\delta \;\middle|\; \begin{array}{l} a_\delta \in R, \text{ and only finitely many} \\ \text{of the } a_\delta\text{'s are nonzero} \end{array} \right\}$$

with the familiar operations of addition and multiplication of polynomials, which can be formally expressed as follows:

$$\begin{aligned} A(x) &= \sum_{\delta \in \mathbb{N}^n} a_\delta x^\delta, \\ B(x) &= \sum_{\delta \in \mathbb{N}^n} b_\delta x^\delta, \\ A(x) + B(x) &\coloneqq \sum_{\delta \in \mathbb{N}^n} (a_\delta + b_\delta) x^\delta, \\ A(x)B(x) &\coloneqq \sum_{\delta \in \mathbb{N}^n} \left(\sum_{\substack{\kappa, \lambda \in \mathbb{N}^n \\ \kappa + \lambda = \delta}} a_\kappa b_\lambda \right) x^\delta. \end{aligned}$$

One notices immediately that addition is well-defined, because it coincides with addition in the infinite direct sum of abelian groups $\bigoplus_{\delta \in \mathbb{N}^n} R$. Multiplication is well-defined because for each coefficient of a monomial in $A(x)B(x)$, we take the sum of only finitely many nonzero elements. Checking distribution is formal and amounts to noticing that $x^\kappa \cdot x^\lambda = x^{\kappa+\lambda}$.

One can also check that if R is commutative, then $R[x_1, \ldots, x_n]$ is, too.

Remark 4.3.36 ($\star$)**.** We can also define rings of polynomials in infinitely many variables in a similar way, except we take care when defining monomials. Namely, for an infinite tuple x of variables, we only allow monomials of multidegrees δ such that only finitely many coordinates in δ are nonzero. In other words, monomials are *finite* products of variables.

4.3.5.4. *Adjunction of elements.* Combining constructions from Section 4.3.5.2 and Section 4.3.5.3, we can *adjoin* elements to a ring such that they satisfy a certain equation or equations. Let R be a commutative ring. Say we want to formally add elements $\alpha_1, \ldots, \alpha_n$, for which $f_1(\alpha_1, \ldots, \alpha_n) = 0, \ldots, f_m(\alpha_1, \ldots, \alpha_n) = 0$, then the blueprint for the construction is encoded by the following formula:

$$R[\alpha_1, \ldots, \alpha_n] \big/ (f_1, \ldots, f_m),$$

where $(f_1, \ldots, f_m)$ is the ideal generated by $f_1, \ldots, f_n$ in the polynomial ring $R[\alpha_1, \ldots, \alpha_n]$, see Example 4.3.26.

Example 4.3.37 (The ring of dual numbers)**.** Let R be a ring, which for this construction is very often assumed to be a field. Then the *ring of dual numbers* is defined as follows:

$$R[\varepsilon] \big/ (\varepsilon^2).$$

For two elements $a + b\varepsilon$ and $c + d\varepsilon$ in this ring, their product is given by the formula:

$$(a + b\varepsilon)(c + d\varepsilon) = ac + (ad + bc)\varepsilon.$$

4.3.5.5. *Product rings.* Given two rings R and S, we define their product $R \times S$ as the set of pairs

$$R \times S \coloneqq \{(r, s) \mid r \in R \text{ and } s \in S\}$$

with coordinate-wise addition and multiplication:

$$(r, s) + (r', s') \coloneqq (r + r', s + s'),$$
$$(r, s) \cdot (r', s') \coloneqq (rr', ss').$$

4.3.6. Isomorphism theorems.

Theorem 4.3.38 (1st isomorphism theorem for rings)**.** Let $\varphi \colon R \to S$ be a ring homomorphism. Then there exists an isomorphism of rings:

$$R/\mathrm{Ker}(\varphi) \cong \mathrm{Im}(\varphi).$$

Proof. By Proposition 4.3.16, we have that $\mathrm{Im}(\varphi)$ is a subring of S. By the universal property of quotient rings, we have that $\overline{\varphi} \colon R/\mathrm{Ker}(\varphi) \to \mathrm{Im}(\varphi)$ is a ring homomorphism. By the universal property of quotient groups, we know that $\overline{\varphi}$ is an isomorphism of groups, in particular, it is bijective. Thus $\overline{\varphi}$ is a ring isomorphism. □

Exercise 4.3.39. Let R be a ring and let I and J be two ideals of R. Then the sets

$$I \cap J = \{x \mid x \in I \text{ and } x \in J\},$$
$$I + J = \{x + y \mid x \in I, y \in J\},$$
$$IJ = \{x_1 y_1 + \cdots + x_n y_n \mid n \geq 1, x_i \in I, y_i \in J\},$$

are ideals of R.

Example 4.3.40. In the ring $\mathbb{Z}$, we have:

$$2\mathbb{Z} \cap 3\mathbb{Z} = 6\mathbb{Z}, \quad 2\mathbb{Z} + 3\mathbb{Z} = \mathbb{Z}, \quad (2\mathbb{Z})(3\mathbb{Z}) = 6\mathbb{Z},$$

as $\gcd(2, 3) = 1$. But we also have:

$$2\mathbb{Z} \cap 4\mathbb{Z} = 4\mathbb{Z}, \quad 2\mathbb{Z} + 4\mathbb{Z} = 2\mathbb{Z}, \quad (2\mathbb{Z})(4\mathbb{Z}) = 8\mathbb{Z}.$$

Theorem 4.3.41 (2nd isomorphism theorem for rings)**.** Let R be a ring, let S be a subring of R, and I an ideal of R. Let $\gamma \colon R \to R/I$ the quotient homomorphism. Then

- the subgroup $S + I = \gamma(S)$ is a subring of R,
- the ideal I is also an ideal of $S + I$, and
- we have an isomorphism of rings:

$$(S + I)/I \cong S/(S \cap I).$$

Theorem 4.3.42 (3rd isomorphism theorem for rings)**.** Let R be a ring, let I and J be ideals of R. Suppose $J \subseteq I$. Then we have an isomorphism of rings:

$$(R/J)/(I/J) \cong R/I.$$

Theorem 4.3.43 (4th isomorphism theorem for rings, or correspondence theorem)**.** Let R be a ring and let I be an ideal of R. Denote by $\gamma \colon R \to R/I$ the quotient

homomorphism. Then there is a bijection between ideals of R containing I and ideals of R/I:

$$\left\{\begin{array}{c}\text{ideals in } R\\ \text{containing } I\end{array}\right\} \rightleftarrows \left\{\begin{array}{c}\text{ideals in}\\ R/I\end{array}\right\},$$

$$J \longmapsto \gamma(J) = J/I,$$

$$\gamma^{-1}(J) \longmapsfrom J.$$

Exercise 4.3.44. Prove the above four theorems using their group versions.

4.4. Principal, maximal, and prime ideals

4.4.1. Principal ideals.

Definition 4.4.1. Let R be a ring.

- We say a left ideal I of R is *left principal*, if $I = Rx = \{rx \mid r \in R\}$, for some $x \in I$.
- We say a right ideal I of R is *right principal*, if $I = xR = \{xr \mid r \in R\}$, for some $x \in I$.
- We say a two-sided ideal I of R is *principal* if:

$$I = RxR = \{r_1 x s_1 + \cdots + r_n x s_n \mid n \geq 1, r_i, s_i \in R\},$$

for some $x \in I$.

When R is a commutative ring, the three above notions coincide. In this case, we denote the ideal $Rx = xR = RxR$ by (x), and we refer to it as the *principal ideal generated by* x.

Example 4.4.2. In the ring $\mathbb{Z}$, the ideals $n\mathbb{Z}$ are principal. In fact. $n\mathbb{Z} = (n)$. These are the only ideals of $\mathbb{Z}$, and thus every ideal is principal.

Example 4.4.3. For any ring R, the non-proper subset R regarded as an ideal is principal: $R = (1_R)$.

Definition 4.4.4. An integral domain in which every ideal is principal is called a *principal ideal domain (PID)*.

Example 4.4.5. The ring $\mathbb{Z}$ is a principal ideal domain.

Example 4.4.6. If $\mathbb{F}$ is a field, then $\mathbb{F}[x]$ is a principal ideal domain. We shall prove this in Theorem **??**.

Example 4.4.7. The ring $\mathbb{Z}[x]$ is *not* a principal ideal domain. For example, the ideal $(2, x)$ is not principal.

Example 4.4.8. The ring $\mathbb{Z}/6$ is not a principal ideal domain because it is not a domain. However, for any $n \in \mathbb{Z}$, every ideal in the ring $\mathbb{Z}/(n)$ is principal.

4.4.2. Maximal ideals.

Definition 4.4.9. A non-trivial ring R is *simple* if the only ideals of R are $\{0\}$ and R.

Lemma 4.4.10. Let I be an ideal of R. Then the following are equivalent:

(i) I contains a unit;
(ii) $1_R \in I$;
(iii) $I = R$.

Proof. **(i)** $\Rightarrow$ **(ii)**. If $u \in I$ is a unit, then $1_R = u^{-1}u$. As $u \in I$, we get $1_R \in I$.
(ii) $\Rightarrow$ **(iii)**. Since $1_R \in I$, then $r = r \cdot 1_R \in I$ for any $r \in R$. Thus $I = R$.
(iii) $\Rightarrow$ **(i)**. As $R^\times \subseteq R = I$, we conclude that I contains all units. □

Proposition 4.4.11. A division ring is a simple ring.

Proof. Let R be a division ring, i.e., $R^\times = R\backslash\{0\}$. Let I be a nonzero ideal of R. Let $x \in I$, where $x \neq 0$, then x is a unit. Therefore, by Lemma 4.4.10, we get that $I = R$. □

The converse of this result is not true, as seen in the following example.

Example 4.4.12. The ring $\mathcal{M}_n(R)$ is simple if and only if R is simple (not obvious). In particular, if R is a field, then $\mathcal{M}_n(R)$ is simple but not a division ring.

However, in the commutative case, we can prove a converse of Proposition 4.4.11.

Theorem 4.4.13. Let R be a commutative ring. Then R is a field if and only if R is a simple ring.

PROOF. For the forward direction, recall that a field is a commutative division ring; in particular, it is simple by Proposition 4.4.11.

For the converse, suppose R is simple. Let $x \in R$, $x \neq 0$. Then $(x) = R$, as R is simple. By Lemma 4.4.10, we get $1_R \in (x)$, hence there exists $y \in R$ such that $1_R = yx$. Thus x is a unit. □

Exercise 4.4.14. Show that the center of a ring R

$$Z(R) = \{z \in R \mid zr = rz, \forall r \in R\}$$

is a subring of R. Moreover, show that if R is simple, then $Z(R)$ is a field.

Definition 4.4.15. An ideal I of a ring R is said to be *maximal* if $I \neq R$, and for any ideal J containing I, we have either $J = I$ or $J = R$.

Theorem 4.4.16. Let I be an ideal of R. Then I is maximal if and only R/I is a simple ring. In particular, if R is a commutative ring, we get that I is maximal if and only if R/I is a field.

PROOF. Apply the correspondence theorem (Theorem 4.3.43). □

Theorem 4.4.17 (Krull's Theorem)**.** Let R be a nontrivial ring. Then there exists a maximal ideal in R.

PROOF. We make use of Zorn's lemma (see Theorem 1.6.19). Define $(P, \subseteq)$ to be the poset of R formed by proper ideals of R, ordered by inclusion. Notice that it is non-empty as R is nontrivial, so $\{0\}$ is a proper ideal of R. Let $S \subseteq P$ be a totally ordered subset of P. We need to show it has an upper bound. We claim that:

$$U = \bigcup_{I \in S} I$$

is an upper bound of S. Indeed, by construction, $I \subseteq U$ for any $I \in S$; and U is an ideal of R as the union of ascending ideals is an ideal. By Zorn's lemma, there is a maximal element in P, i.e., there exists an ideal of R that is maximal. □

4.4.3. Prime ideals.

Definition 4.4.18. Let R be a commutative ring. An ideal P of R is called *prime* if $P \neq R$ and

$$rs \in P \Rightarrow r \in P \text{ or } s \in P.$$

Theorem 4.4.19. Let R be a commutative ring. Let I be a proper ideal. Then I is a prime ideal if and only if R/I is an integral domain.

PROOF. We will prove the contrapositive.

$$\begin{aligned} & \text{The ideal } I \text{ is not prime} \\ \Leftrightarrow\ & \exists r, s \in R \colon r, s \notin I, \text{ but } rs \in I \\ \Leftrightarrow\ & \exists [r], [s] \in R/I : [r], [s] \neq [0], \text{ but } [r][s] = [rs] = [0] \\ \Leftrightarrow\ & R/I \text{ is not an integral domain.} \end{aligned}$$

□

Example 4.4.20. Since $\mathbb{Z}/n\mathbb{Z}$ is an integral domain if and only if n is a prime, we get that $(n) = n\mathbb{Z}$ is a prime ideal if and only if n is a prime.

Corollary 4.4.21. In a commutative ring, a maximal ideal is a prime ideal.

PROOF. Suppose I is an ideal of a commutative ring R. Then:

$$\begin{aligned} & I \text{ is a maximal ideal} \\ \Leftrightarrow\ & R/I \text{ is a field} \\ \Rightarrow\ & R/I \text{ is an integral domain} \\ \Leftrightarrow\ & I \text{ is a prime ideal.} \end{aligned}$$

□

Example 4.4.22. In many commutative rings prime ideals are not maximal. For example, let (x) be the ideal generated by x in $\mathbb{Z}[x]$. If we apply the 1st isomorphism theorem to the ring homomorphism $\mathrm{ev}_0 \colon \mathbb{Z}[x] \to \mathbb{Z}$ (see Example 4.3.17), we obtain an isomorphism of rings $\mathbb{Z}[x]/(x) \cong \mathbb{Z}$. Since $\mathbb{Z}$ is an integral domain but not a field, we get that (x) is a prime ideal but not a maximal ideal of $\mathbb{Z}[x]$.

Corollary 4.4.23. Let R be a principal ideal domain (PID). Let $I \neq 0$ be an ideal of R. Then I is a maximal ideal if and only if I is a prime ideal.

PROOF. We only need to show that prime ideals are maximal. Suppose $I = (p)$ is prime. Let $J = (q)$ be an ideal containing I. Let us show that $J = I$ or $J = R$. Since $I \subseteq J$, we get $p \in (q)$. Thus $p = qr$ for some $r \in R$. As $qr \in I$, and I is prime, then $q \in I$ or $r \in I$.

If $q \in I$, then $(q) \subseteq I$ and thus $I = J$.

If $r \in I = (p)$, then $r = ps$ for some $s \in R$. Thus:

$$p = qr = qps.$$

Since R is a integral domain, we can cancel out p on each side and obtain $qs = 1$. Thus $q \in R^\times$, and by Lemma 4.4.10, $J = (q) = R$. □

4.5. Fractions

We now want to recall some properties that the ring of integers has that we have extensively used.

- First, we noted that $\mathbb{Z}$ is a subring of the field of rational numbers $\mathbb{Q}$, and $\mathbb{Q}$ was, in a sense, the smallest field containing $\mathbb{Z}$. Moreover, we defined $\mathbb{Q}$ formally as the set of fractions $\frac{a}{b}$ with $a, b \in \mathbb{Z}$ and $b \neq 0$, and identified two fractions $\frac{a}{b} = \frac{c}{d}$ whenever $ad = bc$.
- For any nonzero integer $n \in \mathbb{Z}$, we have a unique factorization into prime numbers: $n = \pm p_1^{k_1} \cdots p_\ell^{k_\ell}$.
- We used Euclidean division in $\mathbb{Z}$ repeatedly.

It turns out that we can repeat the first construction for arbitrary integral domains; however, the other two nice features are more restrictive.

4.5.1. Fraction fields.

Definition 4.5.1. Let R be an integral domain. We can define a relation on the set $R \times (R\backslash\{0\})$ as follows:

$$(r, s) \sim (r', s') \Leftrightarrow rs' = sr'.$$

Lemma 4.5.2. The relation above defines an equivalence relation on $R \times (R\backslash\{0\})$.

Proof. Reflexivity: $(r, s) \sim (r, s)$ follows from commutativity $rs = sr$.

Symmetry: suppose $(r, s) \sim (r', s')$, which means that $rs' = sr'$. By commutativity, we get $r's = s'r$, and so $(r', s') \sim (r, s)$.

Transitivity: suppose $(r, s) \sim (r', s')$ and $(r', s') \sim (r'', s'')$. In particular, we get $rs' = sr'$ and $r's'' = s'r''$. Thus

$$\begin{aligned} rs''s' &= s''rs' = s''sr' \\ &= s''r's = s'r''s \\ &= r''ss'. \end{aligned}$$

As R is an integral domain and $s' \neq 0$, we can cancel s' from both sides and get $rs'' = r''s$. Thus $(r, s) \sim (r'', s'')$. □

Notation 4.5.3. We denote the equivalence class of $(r, s) \in R \times (R\backslash\{0\})$ by $\frac{r}{s}$. We shall write $\frac{r}{1}$ simply as r. The equivalence relation guarantees that we can cancel denominators:

$$\frac{r}{s} = \frac{r'}{s'} \Leftrightarrow rs' = sr'.$$

Definition 4.5.4. Given an integral domain R, we define its *fraction field* to be the quotient set

$$\begin{aligned} \mathbb{K}(R) &:= \Big(R \times (R\backslash\{0\})\Big)\Big/\sim \\ &= \left\{ \frac{r}{s} \,\middle|\, r \in R, s \in R\backslash\{0\} \right\}. \end{aligned}$$

Theorem 4.5.5. Let R be an integral domain. There are well-defined binary operations on the fraction field $\mathbb{K}(R)$:

$$\begin{aligned} \mathbb{K}(R) \times \mathbb{K}(R) &\xrightarrow{+} \mathbb{K}(R), \\ \left(\frac{a}{b}, \frac{c}{d}\right) &\longmapsto \frac{ad + bc}{bd}, \end{aligned}$$

$$\mathbb{K}(R) \times \mathbb{K}(R) \xrightarrow{\cdot} \mathbb{K}(R),$$
$$\left(\frac{a}{b}, \frac{c}{d}\right) \longmapsto \frac{ac}{bd}.$$

Moreover, $(\mathbb{K}(R), +, \cdot, 0_R, 1_R)$ is a field and contains R as a subring. The inverse of $\frac{a}{b}$, where $a \neq 0$, is given by $\frac{b}{a}$.

Example 4.5.6. We have $\mathbb{K}(\mathbb{Z}) \cong \mathbb{Q}$.

Example 4.5.7. If $\mathbb{F}$ is a field, then $\mathbb{K}(\mathbb{F}) \cong \mathbb{F}$. This is because $\frac{a}{b} = \frac{ab^{-1}}{1} = ab^{-1}$. In other words, there are no inverses to add in $\mathbb{F}$.

Notation 4.5.8. Suppose R is an integral domain. Then its polynomial ring $R[x]$ is an integral domain. We denote by $R(x)$ its fraction field $\mathbb{K}(R[x])$.

4.5.2. ⋆ Rings of fractions.

We generalize the fraction field approach.

Definition 4.5.9. Let R be a ring. A nonempty subset S of R is said to be multiplicative if $\forall a, b \in S \Rightarrow ab \in S$.

Example 4.5.10. An ideal of a ring R is always multiplicative.

Example 4.5.11. Let R be a commutative ring. Let P be a prime ideal. The set $S = R \backslash P$ is multiplicative. Indeed, recall that if $ab \in P$ then $a \in P$ or $b \in P$. Thus if $a \notin P$ and $b \notin P$ then $ab \notin P$.

Example 4.5.12. The group of units $R^\times$ is multiplicative in R.

Example 4.5.13. If R is an integral domain then $R \backslash \{0\}$ is multiplicative.

Definition 4.5.14. Let R be a commutative ring. Let $S \subseteq R$ be a multiplicative subset. Define a relation on $R \times S$:

$$(r, s) \sim (r', s') \Leftrightarrow \exists t \in S : t(rs' - r's) = 0.$$

Exercise 4.5.15. Let R and S be as above.

(i) Show that $\sim$ is an equivalence relation.

(ii) Show that if R is an integral domain and $0 \notin S$, then:

$$(r, s) \sim (r', s') \Leftrightarrow rs' = sr'.$$

(iii) If $0 \in S$, then there exists a unique equivalence class with respect to $\sim$.

(iv) For all $s, t \in S$ and $r \in R$, show that $(tr, ts) \sim (r, s)$.

Definition 4.5.16. We denote the equivalent class of (r, s) in $R \times S$ by $\frac{r}{s}$ and we denote the quotient set:

$$S^{-1}R := R \times S / \sim .,$$

and refer to it as the *localization of R at S*, or the *ring of fractions of R by S*.

Example 4.5.17. If R is an integral domain and $S = R \backslash \{0\}$, then $S^{-1}R = \mathbb{K}(R)$.

Theorem 4.5.18. Let R be a commutative ring. Let $S \subseteq R$ be a multiplicative subset. Then $S^{-1}R$ is a ring with the following well-defined binary operations:

$$S^{-1}R \times S^{-1}R \xrightarrow{+} S^{-1}R$$
$$\left(\frac{a}{b}, \frac{c}{d}\right) \longmapsto \frac{ad + bc}{bd},$$

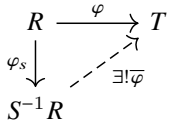
$$\begin{aligned} S^{-1}R \times S^{-1}R &\longrightarrow S^{-1}R \\ \left(\frac{a}{b}, \frac{c}{d}\right) &\longmapsto \frac{ac}{bd}. \end{aligned}$$

Moreover, if R is an integral domain and $0 \notin S$, then $S^{-1}R$ is an integral domain.

Lemma 4.5.19. Let R be a commutative ring, let $S \subseteq R$ be a multiplicative subset, and $0 \notin S$.

- **(i)** Given $s \in S$, define $\varphi_s : R \to S^{-1}R$ by $\varphi_s(r) = \frac{rs}{s}$. Then φ_s is a ring homomorphism and $\varphi_s(t)$ is a unit for all $t \in S$.
- **(ii)** If S contains no zero divisors, then φ_s is injective.
- **(iii)** If S contains only units, then φ_s is an isomorphism of rings.

Theorem 4.5.20 (Universal property of ring of fractions)**.** Let R be a commutative ring. Let $S \subseteq R$ be a multiplicative subset. Let T be a commutative ring. If $\varphi : R \to T$ is a ring homomorphism such that $\varphi(s) \in T^\times$ for all $s \in S$, then there exists a unique ring homomorphism $\overline{\varphi} : S^{-1}R \to T$ such that the following diagram commutes for all $s \in S$:



Theorem 4.5.21. Let R be a commutative ring. Let $S \subseteq R$ be a multiplicative subset.

- **(i)** If I is an ideal of R then:
$$S^{-1}I = \{\frac{x}{s} \mid x \in I, s \in S\},$$
is an ideal of $S^{-1}I$.
- **(ii)** Let I and J be ideals of R. Then:

$$S^{-1}(I+J) = S^{-1}I + S^{-1}J, \quad S^{-1}(IJ) = (S^{-1}I)(S^{-1}J), \quad S^{-1}(I \cap J) = (S^{-1}I) \cap (S^{-1}J).$$

- **(iii)** Let I be an ideal of R. Then $S^{-1}I = S^{-1}R \Leftrightarrow S \cap I \neq \emptyset$.

4.5.3. Local rings.

Definition 4.5.22. A commutative ring is said to be local if there exists a unique maximal ideal.

Theorem 4.5.23. Let R be a commutative ring. Let P be a prime ideal. Let $S = R \backslash P$. Then $S^{-1}R$ is a local ring.

4.6. Unique factorization domains

4.6.1. Irreducible and prime elements. We now turn our attention to the prime factorization property of $\mathbb{Z}$. In order to state it in the general setting, we need to first define what "prime" means in an arbitrary ring. It turns out that there are two natural notions that are in general not equivalent.

Definition 4.6.1. Let R be an integral domain.

(i) Let $r \in R$, $r \neq 0$ and $r \notin R^\times$. We say that r is *irreducible* if

$$r = ab \Rightarrow a \in R^\times \text{ or } b \in R^\times.$$

If r is not irreducible, we say it is *reducible*.

(ii) Let $p \in R$, $p \neq 0$, $p \notin R^\times$. We say that p is *prime* if the ideal (p) is prime, i.e.,

$$p|ab \Rightarrow p|a \text{ or } p|b.$$

⚠ **Warning 4.6.2.** The notions of being prime or irreducible depend on the ambient ring. For instance, $2 \in \mathbb{Z}$ is prime, but $2 \in \mathbb{Q}$ is not prime.

These notions, yet not equivalent, can still be related.

Proposition 4.6.3. Let R be an integral domain, then any prime element is irreducible.

Proof. Suppose r is prime and assume it is written as a product $r = ab$, where $a, b \in R$. Since r is prime and divides ab, we conclude $r|a$ or $r|b$ by definition. Without loss of generality, let us assume $r|a$. Then $\exists s \in R$ such that $a = rs$. Thus $r = ab = rsb$. Since we are in an integral domain, we can cancel r and obtain $1 = sb$. Thus $b \in R^\times$, and we conclude that r is irreducible. □

⚠ **Warning 4.6.4.** The converse is in general not true, as the next two examples show.

Example 4.6.5. Define a ring R as:

$$\begin{aligned} R &= \mathbb{Q} + x\mathbb{R}[x] \\ &= \{a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \mid n \geq 0,\ a_0 \in \mathbb{Q},\ a_i \in \mathbb{R} \text{ for } 0 < i \leq n\}. \end{aligned}$$

One can see that R is a ring. Observe that $x \in R$ is irreducible but not prime:

$$x|\left(\sqrt{2}x\right)^2 = 2x^2,$$

but x does not divide $\sqrt{2}x$ as $\sqrt{2} \notin \mathbb{Q}$.

Example 4.6.6. Consider the ring $\mathbb{Z}\left[\sqrt{-5}\right]$. Then we can write 6 as a product of two irreducible elements in two ways:

$$6 = 2 \cdot 3 = \left(1 + \sqrt{-5}\right)\left(1 - \sqrt{-5}\right).$$

So 2 is irreducible, divides 6, yet does not divide $1 + \sqrt{-5}$, and so it is not prime.

Example 4.6.7. In a field, there are neither irreducible nor prime elements.

Proposition 4.6.8. [See also Proposition 4.6.18] Suppose R is a PID and $p \in R$. Then p is prime if and only if p is irreducible.

PROOF. By Proposition 4.6.3, we only need to check that if p is irreducible, then p is prime. So suppose p is irreducible.

We first show that the ideal (p) is maximal. Suppose $(p) \subseteq I \subseteq R$ is an ideal containing (p), where $I = (m)$. As $p \in I$, then $p = sm$, for some $s \in R$. But as p is irreducible, then $s \in R^\times$ or $m \in R^\times$. If $m \in R^\times$, then $I = (m) = R$. If $s \in R^\times$, then $(p) = (m)$. Thus (p) is maximal.

Since (p) is maximal, it follows from Corollary 4.4.23 that it is prime, and so p is prime. □

Example 4.6.9. In $\mathbb{Z}$, since it is a PID, irreducible elements are prime elements, and they are precisely the prime numbers (and their negatives).

Example 4.6.10. In $\mathbb{R}[x]$, the reader may be familiar with the notion irreducible polynomial, which we recover here. Irreducible elements are prime and can be of two types:

- $u(a + bx + cx^2)$, where $a, b, c \in \mathbb{R}$ such that $b^2 - 4ac < 0$, and $u \in \mathbb{R}^\times$;
- $u(x - a)$, where $a \in \mathbb{R}$, and $u \in \mathbb{R}^\times$.

Example 4.6.11. In $\mathbb{C}[x]$, irreducible elements are prime and are of the form $u(x - a)$ where $a \in \mathbb{C}$ and $u \in \mathbb{C}^\times$.

Exercise 4.6.12. The irreducible elements of $\mathbb{Z}[i]$ are prime, and the following list describes all of them:

- $1 + i$, $1 - i$;
- primes p in $\mathbb{Z}$ such that $p \equiv 3 \mod 4$;
- elements $a + bi$ and $a - bi$ in $\mathbb{Z}[i]$ that appear in the decomposition
$$p = a^2 + b^2 = (a + bi)(a - bi),$$
for some prime integer $p \equiv 1 \mod 4$.

Definition 4.6.13. A *unique factorization domain (UFD)* is an integral domain R such that $\forall r \in R$, $r \neq 0$, $r \notin R^\times$:

(i) $r = p_1 \dots p_n$, where $p_i \in R$ are irreducible elements not necessarily distinct;
(ii) the decompostion is unique: if $r = q_1 \dots q_m$ where $q_j \in R$ are irreducibles, then $m = n$, and $q_i = up_j$ for some $u \in R^\times$.

Example 4.6.14. The ring $\mathbb{Z}$ is a UFD, every number is uniquely decomposed as a product of prime numbers.

Example 4.6.15. We shall see below that every PID is a UFD.

Example 4.6.16. We shall see that $R[x]$ is a UFD if and only if R is a UFD.

Example 4.6.17. The integral domain $\mathbb{Z}[2i]$ is not a UFD as we obtain two distinct irreducible decompositions:
$$4 = 2 \cdot 2 = (-2i)(2i).$$

Proposition 4.6.18. Suppose R is a UFD and $p \in R$. Then p is prime if and only if p is irreducible.

PROOF. Let p be irreducible. Assume $p|ab$. Then $ab = pr$ for some $r \in R$. By uniqueness of decomposition, we see up to a unit, p must appear in the decompostion of a or b. Without loss of generality, let us assume a, so $a = (up)p_2 \dots p_n$ for some $u \in R^\times$, and $p_2, \dots, p_n$ irreducible. Then p divides a. Thus p is prime. □

Exercise 4.6.19. Let R be a UFD, and let $a, b \in R$, $b \neq 0$. Denote the compositions:

$$a = up_1^{e_1} \dots p_n^{e_n}, \qquad b = vp_1^{f_1} \dots p_n^{f_n},$$

where $u, v \in R^\times$, $p_i \in R$ are distinct irreducibles/primes, and $e_i, f_i \geq 0$. Show that:

$$\gcd(a, b) = p_1^{\min(e_1, f_1)} \dots p_n^{\min(e_n, f_n)}.$$

Theorem 4.6.20. A PID is a UFD.

Proof. The proof is similar as in $\mathbb{Z}$. Let R be a PID. Let us first prove uniqueness. Take $r \in R$, $r \neq 0$, $r \notin R^\times$. If r happens to be irreducible, then there is nothing to do. So assume r is reducible. In particular this means $r = r_1 r_2$ where $r_1, r_2 \notin R^\times$. If both r_1, r_2 are irreducible, then we are done, so suppose one is not, say r_1. Then again $r_1 = r_{11} r_{12}$ and so forth. We now must argue this process terminates. We can consider the ascending ideals:

$$(r) \subseteq (r_1) \subseteq (r_{11}) \subseteq (r_{111}) \subseteq \cdots .$$

Consider I the union of all these ideals. It is also an ideal, and since R is a PID then $I = (a)$ for some $a \in R$. Thus $a \in (r_{1\dots1})$ and thus the ascending sequence of ideals eventually becomes stationary, and therefore proves the existence.

Let us prove uniqueness. Suppose:

$$r = p_1 \dots p_n = q_1 \dots q_m$$

where $m \geq n$, and p_i and q_j are primes. Then p_1 divides $q_1 \dots q_m$, thus $p_1 | q_j$ for some j. By irreducibility, we get $p_1 = uq_j$ for some $u \in R^\times$. We can repeat the argument to conclude. □

4.7. Euclidean domains

We have seen that the integers $\mathbb{Z}$ and polynomial rings $\mathbb{F}[x]$ over a field have a Euclidean division. We shall study rings here with similar properties. We begin with a somewhat general definition.

Definition 4.7.1. Let R be an integral domain. A *norm* on R is a function

$$N\colon R \to \mathbb{N}$$

such that $N(0_R) = 0$.

Definition 4.7.2. An integral domain R is said to be a *Euclidean domain* if there is a norm $N\colon R \to \mathbb{N}$ such that, for all $a \in R$, $b \in R\backslash\{0\}$, there exist $q, r \in R$ such that:

$$a = qb + r, \qquad \text{where } r{=}0 \text{ or } N(r) < N(b).$$

We refer to a as the *dividend*, b as the *divisor*, q a *quotient* and r a *remainder*.

⚠ **Warning 4.7.3.** In a general Euclidean domain, we do not require the choice of q and r to be unique. This differs from our intuition originated by $\mathbb{Z}$ and $\mathbb{F}[x]$.

Example 4.7.4. The usual Euclidean division in $\mathbb{Z}$ shows that it is a Euclidean domain with norm $N : \mathbb{Z} \to \mathbb{N}$ defined by absolute values $N(k) = |k|$.

Example 4.7.5. Given a field $\mathbb{F}$, its polynomial ring $\mathbb{F}[x]$ is a Euclidean domain with norm given by the degree of polynomials $\deg\colon \mathbb{F}[x] \to \mathbb{N}$. By Proposition 4.7.8 below and Corollary 4.8.23, if R is not a field, then $R[x]$ is not a Euclidean domain.

Example 4.7.6. Any field $\mathbb{F}$ is a Euclidean domain in an uninteresting way: define $N\colon \mathbb{F} \to \mathbb{N}$ by $N(a) = 0$ for all $a \in \mathbb{F}$. Given $a \in \mathbb{F}$ and $b \in \mathbb{F}\backslash\{0\}$, we have $a = qb + 0$, where $q = ab^{-1}$, as $\mathbb{F}\backslash\{0\} = \mathbb{F}^\times$. The remainder will always be zero.

Example 4.7.7. The Gaussian integers $\mathbb{Z}[i]$ form a Euclidean domain with norm

$$N\colon \mathbb{Z}[i] \longrightarrow \mathbb{N},$$
$$a + bi \longmapsto a^2 + b^2.$$

Given $\alpha, \beta \in \mathbb{Z}[i]$, where $\beta \neq 0$, we can find $q, r \in \mathbb{Z}[i]$ such that $\alpha = q\beta + r$ using plane geometry. Indeed, notice that $\sqrt{N(a+bi)}$ is the distance in the complex 2D-plane of $a + bi$. Pick $\gamma = q\beta$, where $q \in \mathbb{Z}[i]$ such that $|\alpha - \gamma| < |\beta|$. We obtain $N(\alpha - \gamma) < N(\beta)$. Therefore $\alpha = q\beta + r$, with $r = \alpha - \gamma$.

Proposition 4.7.8. A Euclidean domain is a principal ideal domain (PID).

Proof. The proof is similar to $\mathbb{Z}$ and $\mathbb{F}[x]$. Let $I \neq 0$ be an ideal of a Euclidean domain R. Choose $d \in I$, $d \neq 0$ with minimal norm in I: if $x \in I$, then $N(d) \leq N(x)$. Given $x \in I$, we have $\exists q, r \in R$: $x = qd + r$ where $r = 0$ or $N(r) < N(d)$. Since $d \in I$, then $qd \in I$. Hence $r = x - qd \in I$ as $x \in I$. However, by minimality of d, we cannot have $N(r) < N(d)$. Thus we must have $r = 0$. Hence $x = qd$, i.e. $x \in (d)$. Thus $I = (x)$. Hence every ideal is principal. □

Notation 4.7.9. Let R be a commutative ring. Let $x_1, \ldots, x_n \in R$. Then we denote by

$$(x_1, \ldots, x_n) = \{r_1x_1 + \cdots r_nx_n \mid r_i \in R\}$$

the *ideal in R generated by $x_1, \ldots, x_n$*. When $n = 1$, we recover our notion of principal ideal.

Definition 4.7.10. Let R be a commutative ring. Let $a, b \in R$, $b \neq 0$. A *greatest common divisor* of a and b is a non-zero element $d \in R$ such that:

(i) $d|a$ and $d|b$;
(ii) if $k|a$ and $k|b$ for some $k \in R$, then $k|d$.

If we denote by $I = (a, b)$ the ideal generated by a and b, then the axioms above translate to:

(i) $(a, b) \subseteq (d)$;
(ii) if $(a, b) \subseteq (k)$ for some $k \in R$, then $(d) \subseteq (k)$.

Lemma 4.7.11. Let R be an integral domain. Let $a, b \in R$, $b \neq 0$. If a greatest common divisor of a and b exists, then it is unique up to a unit.

Proof. Suppose d and d' are greatest common divisor of a and b. Then $d|d'$ and $d'|d$. Therefore there exists $u, v \in R$ such that $d = ud'$ and $d' = vd$. Combining the equations, we get:

$$\begin{aligned} & d = uvd \\ \Rightarrow & d(1 - uv) = 0. \end{aligned}$$

As $d|b \neq 0$ we get $d \neq 0$, and so $1 - uv = 0$, thus $u, v \in R^\times$. □

⚠ **Warning 4.7.12.** A greatest common divisor may not always exist. For instance, in the ring $\mathbb{Z}[i\sqrt{5}]$, take $a = 6$ and $b = 2 + 2i\sqrt{5}$. Then 2 divides both a and b and $1 + i\sqrt{5}$ divides both a and b, but neither divides the other.

Notation 4.7.13. In an integral domain, we denote the greatest common divisor of a and b by $\gcd(a, b)$ which is well-defined up to a unit.

Proposition 4.7.14. Let R be a commutative ring. Let $a, b \in R$, $b \neq 0$. Suppose there exists $d \in R$ such that $(a, b) = (d)$. Then $\gcd(a, b) = d$ and $\exists x, y \in R$:

$$d = ax + by \quad \textit{(Bézout's identity).}$$

In particular, a greatest common divisor always exists if R is a PID.

If R is a PID, we know that $\gcd(a, b)$ exists but we do not have a method to compute it. If R is a Euclidean domain however, we have the following familiar method to find a greatest common divisor.

Theorem 4.7.15 (Euclidean algorithm)**.** Let R be a Euclidean domain. Let $a, b \in R$, $b \neq 0$. Then iterating Euclidean division with previous remainder and divisor eventually leads to a zero remainder:

- $\exists q_0, r_0 \in R : a = q_0 b + r_0$ and $N(r_0) < N(b)$;
- $\exists q_1, r_1 \in R : b = q_1 r_0 + r_1$ and $N(r_1) < N(r_0)$;
- $\exists q_2, r_2 \in R : r_0 = q_2 r_1 + r_2$ and $N(r_2) < N(r_1)$;
- $\vdots$
- $\exists q_n, r_n \in R : r_{n-2} = q_n r_{n-1} + r_n$ and $N(r_n) < N(r_{n-1})$;
- $\exists q_{n+1} \in R : r_{n-1} = q_{n+1} r_n$.

Denoting $d = r_n$ the last non-zero remainder of the above procedure, we obtain:

(i) $d = \gcd(a, b)$;
(ii) $(d) = (a, b)$, in particular, $\exists x, y \in R$:

$$d = ax + by \quad \textit{(Bézout's identity).}$$

PROOF. Since b is fixed, then $N(b) \in \mathbb{N}$ is a fixed value. The algorithm creates a sequence:

$$N(b) > N(r_0) > N(r_1) > \cdots .$$

Thus there must exists $n \geq 0$ such that $N(r_n) > 0$ but $N(r_{n+1}) = 0$.

Denote $d = r_n$. We first argue $(a, b) \subseteq (d)$. As $r_{n-1} = q_{n+1}d$, we get $d|r_{n-1}$. As $r_{n-2} = q_n r_{n-1} + d$, we obtain $d|r_{n-2}$. By induction, fix $k \geq 0$, if we suppose $d|r_i$ for all $k + 1 \leq i \leq n$, then as $r_k = q_{k+2}r_{k+1} + r_{k+2}$, we get $d|r_k$. Thus we can conclude $d|r_k$ for all $k \geq 0$. Thus as $b = q_1 r_0 + r_1$, we get $d|b$. As $a = q_0 b + r_0$, then $d|a$. Thus $(a, b) \subseteq (d)$.

We now argue $(d) \subseteq (a, b)$, this would show $(d) = (a, b)$, and by previous result, $\gcd(a, b) = d$. From $a = q_0 b + r_0$, we get $r_0 \in (a, b)$. From $b = q_1 r_0 + r_1$, we get $r_1 \in (b, r_0) \subseteq (a, b)$. Inductively, we obtain eventually $d \in (r_{n-2}, r_{n-1}) \subseteq \ldots \subseteq (a, b)$. □

⚠ **Warning 4.7.16.** Euclidean domains and PIDs are very similar notions and most of the PIDs we meet are in fact Euclidean domains. It turns out that it is not easy to find a PID that is not a Euclidean domain. One of the most "simple" examples is $\mathbb{Z}\left[\frac{1+i\sqrt{19}}{2}\right]$.

4.8. Polynomial rings

In Example 4.2.8, we introduced the polynomial ring $R[x]$ for any ring R. Recall that we can view any element in R as a constant polynomial in $R[x]$, and therefore we have a natural injective ring homomorphism $\iota\colon R \hookrightarrow R[X]$. The polynomial ring $R[x]$ (together with the ring homomorphism ι) is entirely determined by the following universal property.

Theorem 4.8.1 (Universal property of polynomial rings)**.** Let R and S be rings. Let $\varphi\colon R \to S$ be a ring homomorphism and fix $s \in S$. Then there exists a unique ring homomorphism $\overline{\varphi}\colon R[x] \to S$ such that $\overline{\varphi}(x) = s$ and $\overline{\varphi} \circ \iota = \varphi$:

$$\begin{array}{ccc} R & \xrightarrow{\varphi} & S. \\ {\scriptstyle\iota}\big\downarrow & \nearrow_{\exists!\overline{\varphi}} & \\ R[x] & & \end{array}$$

In other words, a ring homomorphism $R[x] \to S$ is entirely determined by its value on x and on constants.

Proof. Suppose first that such a ring homomorphism $\overline{\varphi}$ exists. Then we obtain:

$$\overline{\varphi}\left(\sum_{i=0}^{n} a_i x^i\right) = \sum_{i=0}^{n} \overline{\varphi}(a_i x^i) = \sum_{i=0}^{n} \overline{\varphi}(a_i)\overline{\varphi}(x)^i = \sum_{i=0}^{n} \varphi(a_i) s^i.$$

Therefore, if $\overline{\varphi}$ exists, it must be unique and defined as:

$$\begin{aligned} \overline{\varphi}\colon R[x] &\longrightarrow S \\ \sum_{i=0}^{n} a_i x^i &\longmapsto \sum_{i=0}^{n} \varphi(a_i) s^i. \end{aligned}$$

It only remains to check that it is a ring homomorphism. We check first that $\overline{\varphi}$ respects addition:

$$\begin{aligned} \overline{\varphi}\left(\sum_{i=0}^{n} a_i x^i + \sum_{i=0}^{m} b_i x^i\right) &= \overline{\varphi}\left(\sum_{i=0}^{\max(n,m)} (a_i + b_i) x^i\right) \\ &= \sum_{i=0}^{\max(n,m)} \varphi(a_i + b_i) s^i \\ &= \sum_{i=0}^{\max(n,m)} (\varphi(a_i) + \varphi(b_i)) s^i \\ &= \sum_{i=0}^{n} \varphi(a_i) s^i + \sum_{i=0}^{m} \varphi(b_i) s^i \\ &= \overline{\varphi}\left(\sum_{i=0}^{n} a_i x^i\right) + \overline{\varphi}\left(\sum_{i=0}^{m} b_i x^i\right). \end{aligned}$$

Then, we check that $\overline{\varphi}$ respects multiplication as well:

$$\begin{aligned}
\overline{\varphi}\left(\sum_{i=0}^{n} a_i x^i \cdot \sum_{i=0}^{m} b_i x^i\right) &= \overline{\varphi}\left(\sum_{i=0}^{n+m}\left(\sum_{j=0}^{i} a_j b_{i-j}\right)x^i\right) \\
&= \sum_{i=0}^{n+m} \varphi\left(\sum_{j=0}^{i} a_j b_{i-j}\right)s^i \\
&= \sum_{i=0}^{n+m}\left(\sum_{j=0}^{i} \varphi(a_j)\varphi(b_{i-j})\right)s^i \\
&= \sum_{i=0}^{n} \varphi(a_i)s^i \cdot \sum_{i=0}^{m} \varphi(b_i)s^i \\
&= \overline{\varphi}\left(\sum_{i=0}^{n} a_i x^i\right) \cdot \overline{\varphi}\left(\sum_{i=0}^{m} b_i x^i\right).
\end{aligned}$$

Finally, we see $\overline{\varphi}(1_{R[x]}) = \overline{\varphi}(1_R) = \varphi(1_R) = 1_S$, and this finishes the verification of the axioms of a ring homomorphism. □

Example 4.8.2. If R is a subring of S, then it defines a ring homomorphism $R \hookrightarrow S$. If we pick $\alpha \in S$, then the universal property provides precisely the evaluation homomorphsim $ev_\alpha \colon R[\alpha] \to S$ of Example 4.3.17.

Example 4.8.3. If instead we consider any rings R and S and replace S by $S[x]$ in the universal property, then, given a ring homomorphism $\varphi \colon R \to S$, we obtain the ring homomorphism $R \xrightarrow{\varphi} S \hookrightarrow S[x]$. By the universal property, fixing $x \in S[x]$, we obtain a unique ring homomorphism $R[x] \to S[x]$ as in Example 4.3.21, which we again denote by φ.

In fact, the unique ring homomorphism in Example 4.8.2 is the composition of these maps:

$$R[x] \xrightarrow{\varphi} S[x] \xrightarrow{\mathrm{ev}_\alpha} S.$$

Exercise 4.8.4. Suppose we are given a ring A together with a fixed element $a \in A$ and a ring homomorphism $i \colon R \to A$ that satisfies the following property: given any ring homomorphism $\varphi \colon R \to S$ and $s \in S$, there exists a unique ring homomorphism $\overline{\varphi} : A \to S$ such that $\overline{\varphi} \circ i = \varphi$ and $\overline{\varphi}(a) = s$. Show that $A \cong R[x]$ as rings.

Lemma 4.8.5. Let R be an integral domain. Then polynomials with coefficients in R enjoy the following properties.

- **(i)** For $f, g \in R[x]$, we get $\deg(fg) = \deg(f) + \deg(g)$.
- **(ii)** The polynomial ring $R[x]$ is an integral domain.
- **(iii)** We have $(R[x])^\times = R^\times$.

Proof. Suppose $f(x) = \sum_{i=0}^{n} a_i x^i$ and $g(x) = \sum_{i=0}^{m} b_i x^i$ are polynomials of degree n and m, respectively. In particular, we have $a_n \neq 0$ and $b_m \neq 0$. Then their product is

$$f(x)g(x) = a_0 b_0 + (a_0 b_1 + a_1 b_0)x + \cdots + \left(\sum_{i+j=k} a_i b_j\right)x^k + \cdots + a_n b_m x^{n+m}.$$

Since R is an integral domain, we have $a_n b_m \neq 0$. Therefore,

$$\deg(f+g) = n+m = \deg(f) + \deg(g).$$

In particular, the product of two nonconstant polynomials will result in a nonconstant polynomial. Thus if there are any zero divisors, they should be constant polynomials. But R is an integral domain, so has no nonzero zero divisors; hence $R[x]$ is an integral domain too.

Assume $f \in (R[x])^\times$. Then there exists $g \in R[x]$ such that $fg = 1_R$. In particular, $\deg(f) + \deg(g) = \deg(fg) = 0$. Thus $\deg(f) = 0 = \deg(g)$, and so f and g must be constant polynomials. Therefore, $R[x]^\times = R^\times$. □

Example 4.8.6. Let $R = \mathbb{Z}/4\mathbb{Z}$—this is not an integral domain. Let $f(x) = 1 + 2x$ and $g(x) = -1 + 2x$ in $R[x]$. We get

$$(fg)(x) = (1+2x)(-1+2x) = -1 + 4x + 4x^2 = 3.$$

Thus here $\deg(fg) = 0$ and is not equal to $\deg(f) + \deg(g) = 2$. Moreover, we can see that

$$f^2(x) = (1+2x)^2 = 1 + 4x + 4x^2 = 1.$$

So f is its own inverse although not a constant.

Definition 4.8.7. Let R be a ring. Let $f(x)$ be a polynomial in $R[x]$. We say $\alpha \in R$ is a *root* of $f(x)$ in R if $f(\alpha) = 0$.

Recall that a ring homomorphism $\varphi\colon R \to S$ induces a ring homomorphism $\varphi\colon R[x] \to S[x]$, as shown in Example 4.8.3. We will prove that ring homomorphisms preserve roots.

Proposition 4.8.8. Let $\varphi\colon R \to S$ be a ring homomorphism. Let $f \in R[x]$ be a polynomial with a root $\alpha \in R$. Then $\varphi(\alpha)$ is a root of $\varphi(f)$ in S.

Proof. Let $f(x) = \sum_i^n a_i x^n$. We have

$$f(\alpha) = \sum_i^n a_i \alpha^n = 0.$$

Therefore,

$$\varphi(f)(\varphi(\alpha)) = \sum_i^n \varphi(a_i)\varphi(\alpha)^n = \sum_i^n \varphi(a_i\alpha^n) = \varphi\left(\sum_i^n a_i\alpha^n\right) = \varphi(f(\alpha)) = \varphi(0) = 0.$$

Thus $\varphi(\alpha)$ is a root of $\varphi(f)$ in S. □

Example 4.8.9. If R is a subring of S, then any polynomial with coefficients in R can be regarded in particular as a polynomial with coefficients in S. A polynomial may not have roots in R, but can acquire those in S. Famously, $x^2 + 1$ has no roots in $\mathbb{R}$, but has two roots $\pm i$ in $\mathbb{C}$.

Example 4.8.10. Proposition 4.8.8 has a powerful application. Imagine that we are tasked with showing that

$$x^3 - 5x^2 - x - 17 = 0$$

has no solution in $\mathbb{Z}$. Consider the quotient homomorphism $\gamma\colon \mathbb{Z} \to \mathbb{Z}/5\mathbb{Z}$ and the polynomial $f(x) = x^3 - 5x^2 - x - 17$. Then $\gamma(f) = x^3 - x - 2$. However, $\gamma(f)$ has no roots in $\mathbb{Z}/5\mathbb{Z}$ (it suffices to check for every element). Thus by Proposition 4.8.8, f

cannot have any roots in $\mathbb{Z}$ either, as otherwise it would imply the existence of a root for $\gamma(f)$ in $\mathbb{Z}/5\mathbb{Z}$.

Theorem 4.8.11. Let R be an integral domain. Fix $f, g \in R[x]$, where $g \neq 0$, and the leading coefficient of g is invertible in R. Then there exist $q, r \in R[x]$ such that $f = gq + r$ and $\deg(r) < \deg(g)$.

PROOF. To be finished. □

Example 4.8.12. Euclidean division of $1 + x + 3x^3 + x^4$ by $1 + x^2$ in $\mathbb{Z}[x]$.

Definition 4.8.13. Let $f, g \in R[x]$. We say that g *divides* f if there exists $h \in R[x]$ such that $f = gh$. We write $g|f$.

Theorem 4.8.14. Let R be an integral domain. Let $f \in R[x]$ and fix $\alpha \in R$. Then α is a root of f if and only if $(x - \alpha)|f$. More generally, the remainder of the Euclidean division of f by $(x - \alpha)$ is $f(\alpha)$.

PROOF. By Theorem 4.8.11, we know that there exist polynomials $q, r \in R[x]$ such that

$$f(x) = (x - \alpha)q(x) + r(x)$$

and $\deg(r) < \deg(x - \alpha) = 1$. Thus $\deg(r) \leq 0$, i.e., $r(x) = r$ is a constant. Moreover, evaluating at α shows that

$$f(\alpha) = (\alpha - \alpha)q(\alpha) + r = 0 + r = r.$$

Therefore $r = f(\alpha)$. □

Corollary 4.8.15. Let R be an integral domain and assume that $x - \alpha$ divides a polynomial $f \in R[x]$. Assume that f can be written as a product $f = gh$. Then $x - \alpha$ divides at least one of $g(\alpha) = 0$ or $h(\alpha) = 0$.

PROOF. If $x - \alpha$ divides f, then by Theorem 4.8.14, $f(\alpha) = 0$. That means that $g(\alpha)h(\alpha) = f(\alpha) = 0$. Since R is an integral domain and $g(\alpha), h(\alpha) \in R$, we conclude that one of them is zero, i.e., $g(\alpha) = 0$ or $h(\alpha)$. So again by Theorem 4.8.14, g or h is divisible by $x - \alpha$. □

Corollary 4.8.16. Let R be an integral domain. Then a polynomial in $R[x]$ of degree $n \geq 1$ has at most n distinct roots in R.

PROOF. We prove by induction on $n \geq 1$. Suppose f is of degree 1 and α is a root of f. Then by Theorem 4.8.14, we get $f(x) = q(x)(x - \alpha)$ for some $q(x) \in R[x]$. But as $1 = \deg(f) = \deg(u) + \deg(x - \alpha) = \deg(q) + 1$, we obtain $q = q(x)$ is a constant polynomial. Therefore, q has no roots, and f has at most one root.

Suppose we proved that polynomials of degree n have at most n roots. Let f be a polynomial of degree $n+1$ and suppose that α is a root of f. Then $f(x) = q(x)(x - \alpha)$ for some $q(x)$, and $\deg(q) = n$. For any other root $\beta \neq \alpha$ of f, we know by Theorem 4.8.14 and Corollary 4.8.15 that $x - \beta$ must divide q. But by induction hypothesis, q has at most n distinct roots, so we obtain that f has at most $n + 1$ distinct roots. □

⚠ **Warning 4.8.17.** If R is not an integral domain, then the previous results may not be true. For instance, if $R = \mathbb{Z}/6\mathbb{Z}$, then all elements of R are roots of $f(x) = x^3 - x$.

Corollary 4.8.18. Let $\mathbb{F}$ be a field and $G \subseteq \mathbb{F}^\times$ be a finite subgroup. Then G must be a cyclic group. In particular, if $\mathbb{F}$ is a finite field, then $\mathbb{F}^\times$ is a cyclic group.

PROOF. As G is a finite abelian group, we get by the classification theorem that

$$G \cong C_1 \times \cdots \times C_k$$

for some cyclic subgroups $C_i \leq G$ of orders $n_1, \ldots, n_k$ such that each n_i divides the next: $n_1 | n_2 | \ldots | n_k$. Consider the polynomial $x^{n_1} - 1 \in \mathbb{F}[x]$. By Corollary 4.8.16, it has at most n_1 roots, i.e., there are at most n_1 elements g in G for which $g^{n_1} = 1$. However, in a finite group of order n, every element g satisfies $g^n = 1$, so every element of C_1 is a root of $x^{n_1} - 1$. If $k \geq 2$, then each factor C_i for $i \geq 2$ contains elements of order dividing n_1, which are forced to be roots of $x^{n_1} - 1$. But this would mean that $x^{n_1} - 1 \in \mathbb{F}[x]$ contains more $n_1 = |C_1|$ distinct roots in $\mathbb{F}$, which is not possible. Thus $k = 1$, and so $G \cong C_1$ is cyclic. □

Definition 4.8.19. Let R be a ring and $f(x) \in R[x]$ be a polynomial. Let $\alpha \in R$. Suppose $f(x) = (x - \alpha)^m q(x)$ for some $m \geq 1$ and $q(x) \in R[x]$ such that $q(\alpha) \neq 0$. Then we say that m is the *multiplicity* of the root α of f.

Definition 4.8.20. Let R be a ring. We say a polynomial in $R[x]$ is *monic* if its leading coefficient is 1_R.

Lemma 4.8.21. Let R be an integral domain. Suppose f and g are nonzero monic polynomials in $R[x]$. If $f|g$ and $g|f$, then $f = g$.

PROOF. By assumption, there exist polynomials p, q such that:

$$f(x) = q(x)g(x), \qquad g(x) = p(x)f(x).$$

Therefore $f(x) = p(x)q(x)f(x)$. Thus by degree comparison, we must have that p and q are constant polynomials. Since f and g are monic, we obtain from $f(x) = qg(x)$ that $q = 1$ by comparing leading coefficients. Thus $f = g$. □

Theorem 4.8.22. Let $\mathbb{F}$ be a field. Then $\mathbb{F}[x]$ is a PID, i.e., every ideal I of $\mathbb{F}[x]$ is principal. If we require f to be monic, that choice is unique.

PROOF. Suppose $I \neq 0$ is an ideal of $\mathbb{F}[x]$. Suppose $g(x) \in I$ is a non-zero polynomial. As I is an ideal, for any $k \in \mathbb{F}$, we get $kg(x) \in I$. In particular, since $\mathbb{F}$ is a field, we can choose k to be the inverse of the leading coefficient of $g(x)$ and thus kg is monic. Pick $f(x) \in I$ to be a monic polynomial with minimal degree in I, i.e., if $h(x) \in I$ is monic, then $\deg(f) \leq \deg(h)$. By Euclidean division, for any $g(x) \in I$, there exist $q(x), r(x) \in \mathbb{F}[x]$ such that $g(x) = f(x)q(x) + r(x)$ and $\deg(r) < \deg(f)$. As I is an ideal, then $f(x)q(x) \in I$, and therefore $r(x) = g(x) - f(x)q(x)$ is in I. If a is the leading coefficient of $r(x)$, then $a^{-1}r(x) \in I$ is a monic polynomial. By minimality of f, this is not possible, so $r(x) = 0$. Therefore $g(x) = f(x)q(x)$, i.e., $g(x) \in (f(x))$. Thus $I = (f(x))$.

The choice is unique since, if $I = (\widetilde{f}(x))$ where $\widetilde{f} \in R[x]$ is a monic polynomial, then from $(\widetilde{f}(x)) = (f(x))$, we get $\widetilde{f}(x) | f(x)$ and $f(x) | \widetilde{f}(x)$. By Lemma 4.8.21, we get $f(x) = \widetilde{f}(x)$. □

We conclude the section with an observation that the assumption that $\mathbb{F}$ is a field cannot be relaxed.

Corollary 4.8.23. Let R be a commutative ring.

$$R[x] \text{ is a PID} \quad \Leftrightarrow \quad R \text{ is a field.}$$

Proof. We only need to prove the forward direction, since the converse was proved in Theorem 4.8.22. Suppose $R[x]$ is a PID. It is in particular an integral domain. As $R \subseteq R[x]$, we get that R must be an integral domain. Therefore, by Theorem 4.4.19, the ideal (x) is a prime ideal as we have an isomorphism of rings:

$$R[x]/(x) \cong R.$$

But since $R[x]$ is a PID, then (x) is also a maximal ideal by Corollary 4.4.23. Thus $R[x]/(x)$ is a field by Theorem 4.4.16. Hence R is a field. □

4.9. Irreducible polynomials

Definition 4.9.1. Let R be a ring. We can define a polynomial ring $R[x,y]$ with two variables x and y as $R[x,y] = (R[x])[y]$. Therefore elements in this ring are for instance of the form:

$$f(x,y) = a_{00} + a_{10}x + a_{01}y + a_{11}xy + a_{20}x^2 + \cdots + a_{ij}x^i y^j + \cdots + a_{nm}x^n y^m,$$

where $a_{ij} \in R$. Inductively, we can define the polynomial ring on R with n variables as $R[x_1,\ldots,x_n] = (R[x_1,\ldots,x_{n-1}])[x_n]$ and notice $R[x_1,\ldots,x_{n-1}]$ is a subring. We can even define the polynomial ring with countably many variables as:

$$R[x_1,x_2,\ldots] = \bigcup_{n\geq 1} R[x_1,\ldots,x_n].$$

Exercise 4.9.2. Let R and S be rings, and let $\varphi : R \to S$ be a ring homomorphism. Let $\alpha_1,\ldots,\alpha_n \in S$ be fixed elements. Show there exists a unique ring homomorphism $\overline{\varphi} : R[x_1,\ldots,x_n] \to S$ such that $\overline{\varphi}(x_i) = \alpha_i$ and $\overline{\varphi} \circ \iota = \varphi$ where $\iota : R \hookrightarrow R[x_1,\ldots,x_n]$ is the inclusion of the subring. *Hint: use induction on $n \geq 1$.*

Suppose R is a commutative ring. Given an ideal I in $R[x_1,\ldots,x_n]$, we obtain a new ring:

$$R[x_1,\ldots,x_n]/I.$$

One can view each x_i as a generator of the ring, and I as the imposing relations.

Example 4.9.3. The field $\mathbb{C}$ is obtained as the above procedure:

$$\mathbb{R}[x]/(x^2+1) \cong \mathbb{C}.$$

This can be proven by considering $\mathrm{ev}_i : \mathbb{R}[x] \to \mathbb{C}$ and the first isomorphism theorem. Philosophically, we added a generator on $\mathbb{R}$ with the equivalence class $[x]$ of x and we imposed $[x^2+1] = [x]^2 + 1 = 0$, i.e. $[x]^2 = -1$, and renamed $[x]$ by i.

Example 4.9.4. The ring $\mathbb{R}[x]/(x^2)$ is called the dual number ring. It is an infinite commutative ring that is not an integral domain.

One question we can first ask is: if $R = \mathbb{F}$ is a field, when are we guaranteed $\mathbb{F}[x]/I$ is also a field? We actually can already answer this fully. Firstly, recall that $\mathbb{F}[x]$ is a PID (even an Euclidean domain), so we know $I = (f(x))$ for some $f(x) \in \mathbb{F}[x]$.

Theorem 4.9.5. Let $\mathbb{F}$ be a field. Let $f(x) \in \mathbb{F}[x]$ be a non-zero polynomial. Then the following are equivalent:

- **(i)** $f(x)$ is an irreducible polynomial in $\mathbb{F}[x]$;
- **(ii)** $(f(x))$ is a prime ideal in $\mathbb{F}[x]$;
- **(iii)** $(f(x))$ is a maximal ideal in $\mathbb{F}[x]$;
- **(iv)** $\mathbb{F}[x]/(f(x))$ is a field;
- **(v)** $\mathbb{F}[x]/(f(x))$ is an integral domain.

Proof. This is simply combining Theorem **??**, Corollary 4.4.23, Theorem 4.4.16, Theorem 4.4.19, and Proposition 4.6.8. □

We are therefore left to determine when is a polynomial $f(x) \in \mathbb{F}[x]$ irreducible. We begin to first argue that it is enough to consider this question in $\mathbb{F}$ and not a more general ring.

4.9.1. Gauss's Lemma. From the previous section, we know that given an integral domain R, we can view it as a subring of its fraction field $\mathbb{K}(R)$, and thus we can view $R[x]$ as a subring of $\mathbb{K}(R)[x]$. The advantage of that procedure is that we know that $\mathbb{K}(R)[x]$ is an Euclidean domain, and thus we will be able to deduce what irreducible elements are in $R[x]$.

Lemma 4.9.6. Let R be a UFD. Let $f(x) = a_0 + a_1x + \cdots + a_nx^n \in R[x]$. Let p be a prime/irreducible element of R. Let $\gamma : R \to R/(p)$ be the quotient ring homomorphism. Denote $\overline{\gamma} : R[x] \to (R/(p))[x]$ the induced ring homomorphism on polynomials. The following are equivalent:

- **(i)** $p|a_i$ in R for all $i = 0, \ldots, n$;
- **(ii)** $p|f(x)$ in $R[x]$;
- **(iii)** $\overline{\gamma}(f(x)) = 0$.

PROOF. We first argue **(i)** $\Leftrightarrow$ **(ii)**:

$$a_i = pb_i \text{ for some } b_i \in R \Leftrightarrow f(x) = p(b_0 + b_1x + \cdots + b_nx^n).$$

Next we argue **(ii)** $\Leftrightarrow$ **(iii)**:

$$f(x) = pg(x) \text{ for some } g(x) \in R[x] \Leftrightarrow f(x) \in \ker(\overline{\gamma}). \qquad \square$$

Definition 4.9.7. Let R be a commutative ring. Let $f(x) = a_0 + a_1x + \cdots + a_nx^n \in R[x]$. We say f is *primitive* in R if $\deg(f) > 0$ and the ideal $(a_0, \ldots, a_n) = R$, i.e. $\gcd(a_1, \ldots, a_n) = 1$.

We can reformulate the previous lemma using the new definition above.

Lemma 4.9.8. Let R be a UFD. Let $f(x) \in R[x]$ where $\deg(f) > 0$. The following are equivalent:

- **(i)** f is primitive in R;
- **(ii)** for all prime/irreducible $p \in R$, we have p does not divide $f(x)$ in $R[x]$;
- **(iii)** for all prime/irreducible $p \in R$, we have $\overline{\gamma}(f(x)) \neq 0$, where $\gamma : R \to R/(p)$ is the quotient ring homomorphism.

Lemma 4.9.9. Let R be a UFD. Let $p \in R$. Then:

$$p \text{ is prime/irreducible in } R \Leftrightarrow p \text{ is prime/irreducible in } R[x].$$

PROOF. $\Leftarrow$) Suppose p is prime in $R[x]$. Suppose $p|(ab)$ in $R \subseteq R[x]$, then $p|a$ or $p|b$ as $a, b \in R \subseteq R[x]$.

$\Rightarrow$) Suppose p is prime in R. Assume $p|fg$ in $R[x]$. Let $\gamma : R \to R/(p)$ be the quotient ring homomorphism. Then $= \overline{\gamma}(f)\overline{\gamma}(g) = \overline{\gamma}(fg) = 0$ by previous lemma. But as $(R/(p))[x]$ is an integral domain (Theorem 4.4.19 and Lemma 4.8.5), we get $\overline{\gamma}(f) = 0$ or $\overline{\gamma}(g) = 0$, i.e. $p|f$ or $p|g$ in $R[x]$. $\square$

Lemma 4.9.10 (Gauss's Lemma (1st version))**.** Let R be a UFD. If $f, g \in R[x]$ are primitive in R, then fg is primitive in R.

PROOF. Since R is an integral domain and $\deg(f), \deg(g) > 0$, then $\deg(fg) > 0$ by Lemma 4.8.5. Let $p \in R$ be prime. Suppose $p|fg$ in $R[x]$, since p is prime in $R[x]$ also be previous lemma, we get $p|f$ or $p|g$, but then we would get that f or g are not primitive. Thus p cannot divide fg. Thus fg is primitive. $\square$

Theorem 4.9.11 (Gauss's Lemma (2nd version))**.** Let R be a UFD. Let $\mathbb{K}(R)$ be its fraction field. Let $f(x) \in R[x]$ such that $\deg(f) > 0$. Then the following are equivalent:

(i) $f(x)$ is irreducible in $R[x]$;
(ii) $f(x)$ is irreducible in $\mathbb{K}(R)[x]$ and f is primitive in R.

In particular, if $f(x)$ is reducible in $\mathbb{K}(R)[x]$, then it is reducible in $R[x]$.

PROOF. **(i) $\Rightarrow$ (ii)**) Suppose f is irreducible in $R[x]$. Notice f must be primitive as otherwise $f(x) = pg(x)$ for some $g(x) \in R[x]$ and this gives a non-trivial factoring in $R[x]$. Let us show now f is irreducible in $\mathbb{K}(R)[x]$. We prove by contradiction: assume $f = \overline{pq}$ in $\mathbb{K}(R)[x]$. By clearing denominators and factoring out common divisors, we can assume:

$$f(x) = \frac{b}{c}p(x)q(x),$$

where $p, q \in R[x]$ are primitive polynomials, and $b, c \in R$, $c \neq 0$. Therefore, we obtain the equation:

$$cf(x) = bp(x)q(x).$$

The greatest common divisor of the coefficients of $cf(x)$ must be c as f is primitive. Previous lemma says that pq remains primitive. Hence the greatest common divisor of the coefficients of $bp(x)q(x)$ must be b. Since the polynomials are equal, the greatest common divisor of their coefficients must differ only by a unit $u \in R$, i.e. $b = uc$. Hence $\frac{b}{c} = \frac{u}{1} = u$. Thus $f(x) = up(x)q(x)$ in $R[x]$. But this means f is reducible in $R[x]$ which we assume was not the case. Hence we obtained a contradiction. Therefore, f must be irreducible in $\mathbb{K}(R)$.

(ii) $\Rightarrow$ (i)) Suppose $f(x)$ is irreducible in $\mathbb{K}(R)[x]$ and f is primitive in R. Assume $f(x) = g(x)h(x)$ for some $g, h \in R[x] \subseteq \mathbb{K}(R)$. Since f is irreducible in $\mathbb{K}(R)$, then $\deg(g) = 0$ or $\deg(h) = 0$. But as f is primitive, whichever polynomial is of degree zero must actually be a unit. Hence g or h is a unit in $R[x]$, hence f is irreducible in $R[x]$. □

⚠ **Warning 4.9.12.** IF R is not a UFD, then the previous theorem is not true. Suppose $R = \mathbb{Z}[2i]$. Then $\mathbb{K}(R) = \mathbb{Q}[2i]$. Consider $f(x) = 1 + x^2 \in \mathbb{Z}[2i]$. Notice that:

$$1 + x^2 = (x - i)(x + i).$$

Notice $i = \frac{1}{2}2i$, so $i \in \mathbb{Q}[2i]$, but $i \notin \mathbb{Z}[2i]$. So f is irreducible in R but not its fraction field $\mathbb{K}(R)$.

⚠ **Warning 4.9.13.** The previous does *not* state that if R is a subring of a field $\mathbb{F}$, then if a polynomial in R is irreducible in $\mathbb{F}$ then it is irreducible in $\mathbb{R}$ For instance $f(x) = 1 + x^2$ is irreducible in $\mathbb{R}$ but reducible in $\mathbb{C}$.

Theorem 4.9.14. Let R be an integral domain. Then:

$$R \text{ is a UFD} \Leftrightarrow R[x] \text{ is a UFD.}$$

PROOF. $\Leftarrow$) Suppose $R[x]$ is a UFD. Then R is an integral domain. Consider $a \in R \subseteq R[x]$, $a \neq 0$, $a \notin R^\times$. Then $a = p_1(x) \cdots p_n(x)$ where $p_i(x)$ are irreducible in $R[x]$. We obtain that $\deg(p_i) = 0$ for all $i = 1, \ldots, n$. Hence $p_i \in R$, $p_i \neq 0$, $p_i \notin R^\times$. Also each p_i is irreducible in R: suppose $p_i = ab$ for some $a, b \in R \subseteq R[x]$. As p_i is irreducible in $R[x]$, we get $a \in (R[x])^\times = R^\times$ or $b \in (R[x])^\times = R^\times$. Thus a

decomposition for a into irreducibles in R exists and it is unique as it is unique for $R[x]$.

$\Rightarrow$) Suppose R is a UFD. Let $f(x) \in R[x]$, and $f \neq 0$. If $\deg(f) = 0$, then we can factor in R. So let us assume $\deg(f) > 0$. We can factor f as:

$$f(x) = dg(x),$$

where $d \in R\backslash\{0\}$ is the greatest common divisor of the coefficients of f (unique up to units), and $g(x) \in R[x]$ is primitive. Consider $g(x) \in \mathbb{K}(R)[x]$, then as $\mathbb{K}(R)$ is an Euclidean domain, it is in particular a UFD. Then $g(x) = \overline{p_1}(x) \dots \overline{p_n}(x)$ where $\overline{p_i}(x) \in \mathbb{K}(R)[x]$ are irreducible. Just as in the previous proof, we can factor out the denominators of each $\overline{p_i}$, and factor out common divisors in their numerator so that that

$$g(x) = \frac{b}{c} p_1(x) \dots p_n(x),$$

where p_i are primitive in R and irreducible in $\mathbb{K}(R)$, and thus in R. We get that $\frac{b}{c} = u \in R^\times$ just as in last proof, and therefore:

$$f(x) = du p_1(x) \dots p_n(x).$$

As R is a UFD, we can decompose $d = q_1 \dots q_m$ into irreducibles in R, and thus irreducibles in $R[x]$. We have therefore obtained the desired factorization. This decomposition is unique as it is unique up to units in $\mathbb{K}(R)$:

$$g(x) = \overline{p_1}(x) \dots \overline{p_n}(x) = \overline{p_1}'(x) \dots \overline{p_n}'(x)$$

then $\overline{p_i}(x) = \frac{r_i}{s_i} \overline{p_i}'(x)$ and so $s_i \overline{p_i}(x) = r_i \overline{p_i}'(x)$. □

Example 4.9.15. If $\mathbb{F}$ is a field, then $\mathbb{F}[x_1, \dots, x_n]$ is a UFD.

4.9.2. Irreducibility criteria. We have just argued that as long as R is a UFD, then understanding what irreducible polynomials are in $R[x]$ is equivalent to understand irreducible polynomials in $\mathbb{K}(R)[x]$, where $\mathbb{K}(R)$ is the fraction field of R. Therefore, we will now focus on $R = \mathbb{F}$ a field. We have already shown that:

- $\alpha \in \mathbb{F}$ is a root of $f(x) \in \mathbb{F}[x]$ if and only if $f(x) = (x - \alpha)g(x)$ for some $g(x) \in \mathbb{F}[x]$;
- $\alpha_1, \dots, \alpha_n \in \mathbb{F}$ are roots in $\mathbb{F}$, then $f(x) = (x - \alpha_1) \dots (x - \alpha_n)g(x)$ for some $g(x) \in \mathbb{F}[x]$;
- if $\deg(f) = n > 0$ then $f(x) \in \mathbb{F}[x]$ has at most n roots in $\mathbb{F}$ (not necessarily distinct).

Theorem 4.9.16. Let $\mathbb{F}$ be a field. Let $f(x) \in \mathbb{F}[x]$. Suppose $\deg(f) = 2$ or $\deg(f) = 3$. Then:

$$f \text{ is irreducible} \Leftrightarrow f \text{ has no root in } \mathbb{F}.$$

Proof. We will prove the contraprosition.

$$\begin{aligned}
&f \text{ is reducible in } \mathbb{F} \\
\Leftrightarrow &\exists p(x), q(x) \in \mathbb{F}[x] \text{ such that } f(x) = p(x)q(x) \text{ where } \deg(p) \neq 0 \neq \deg(q) \\
\Leftrightarrow &\exists \alpha \in \mathbb{F}, q(x) \in \mathbb{F}[x], \deg(q) = 1 \text{ or } \deg(q) = 2 \text{ such that } f(x) = (x - \alpha)q(x) \\
\Leftrightarrow &\exists \alpha \in \mathbb{F} : f(\alpha) = 0.
\end{aligned}$$

□

⚠ **Warning 4.9.17.** Theorem 4.9.16 is only valid for deg ≤ 3. Indeed, if $f \in \mathbb{F}[x]$ is of degree ≥ 4 then we can have f reducible without roots: $f(x) = g(x)h(x)$ where $\deg(g) = 2 = \deg(h)$, and g, h have no roots, and thus irreducible. For instance, consider $f(x) = 1 + 2x + 3x^2 + 2x^3 + x^4 \in \mathbb{R}[x]$:

$$f(x) = (1 + x + x^2)^2.$$

We see that f has no roots in $\mathbb{R}$ but is not irreducible.

More generally, we only have this result.

Theorem 4.9.18. Let $\mathbb{F}$ be a field. Let $f(x) \in \mathbb{F}[x]$, $\deg(f) \geq 2$. Then if f has a root in $\mathbb{F}$ then f is not irreducible.

The last theorem prompts us to ask when a polynomial has roots. We have the following theorem in $\mathbb{Q}$.

Theorem 4.9.19 (Rational root test)**.** Let R be a UFD. Let $\mathbb{K}(R)$ be its fraction field. Let $f(x) = a_0 + a_1 x + \cdots + a_n x^n \in R[x]$. If $\frac{r}{s} \in \mathbb{K}(R)$, $\gcd(r, s) = 1$, and $f(\frac{r}{s}) = 0$, then $r|a_0$ and $s|a_n$ in R. In particular, given if $f(x)$ is monic such that $f(d) \neq 0$ for all $d|a_0$, then $f(x)$ has no roots in $\mathbb{K}(R)$.

Proof. We have $f(\frac{r}{s}) = 0$ thus:

$$0 = a_0 + a_1 \frac{r}{s} + \cdots + a_{n-1} \left(\frac{r}{s}\right)^{n-1} + a_n \left(\frac{r}{s}\right)^n.$$

We multiply by s^n:

$$0 = a_0 s^n + a_1 r s^{n-1} + \cdots + a_{n-1} r^{n-1} s + a_n r^n.$$

Isolating $a_n r^n$ we obtain:

$$a_n r^n = s\big(-a_0 s^n - \cdots - a_{n-1} r^{n-1}\big).$$

Thus s divides $a_n r^n$. As $\gcd(r, s) = 1$, then s must divide a_n. If we isolate $a_0 s^n$, we obtain:

$$a_0 s^n = r\big(-a_1 s^{n-1} - \cdots - a_n r^{n-1}\big).$$

Thus r divides $a_0 s^n$, and as $\gcd(r, s) = 1$, we get that r must divide a_0. □

Example 4.9.20. Suppose p is a prime in $\mathbb{Z}$. Then the polynomials $x^2 - p$ and $x^3 - p$ are irreducible in $\mathbb{Q}[x]$ as they have no roots by the rational root test.

Theorem 4.9.21. Let $\varphi : R \to S$ be a ring homomorphism between integral domains. Let $\overline{\varphi} : R[x] \to S[x]$ be the induced ring homomorphism on polynomials. Suppose $f(x) \in R[x]$ monic. If $\overline{\varphi}(f(x))$ is irreducible in $S[x]$, then $f(x)$ is irreducible in $R[x]$.

Proof. Suppose f is reducible in R, then $f(x) = g(x)h(x)$ in which we have $\deg(g), \deg(h) > 0$ and their leading coefficients are units. Since φ preserves units, then we obtain $\deg(\varphi(g)), \deg(\varphi(h)) > 0$ with leading coefficients are units. Thus we obtain $\varphi(f(x)) = \varphi(g(x))\varphi(h(x))$ is reducible in $S[x]$. □

⚠ **Warning 4.9.22.** It is important to that f was monic in previous statement. For instace, if we considered $f(x) = (2x + 1)(1 + x + x^2) \in \mathbb{Z}[x]$

Theorem 4.9.23 (Eisenstein's criterion)**.** Let R be an integral domain. Let P be a prime ideal. Let $f(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + x^n \in R[x]$ be a monic polynomial, $n \geq 1$. Suppose $a_i \in P$ for all $0 \leq i \leq n-1$, but $a_0 \notin P^2$. Then f is irreducible in $R[x]$.

Proof. Proof by contradiction: suppose f was reducible. Then there exists:

$$g(x) = b_0 + b_1x + \cdots + b_kx^k \in R[x], \quad h(x) = c_0 + c_1x + \cdots + c_\ell x^\ell \in R[x],$$

such that $f(x) = g(x)h(x)$. In particular $k + \ell = n$, $b_kc_\ell = 1$ and $b_0c_0 = a_0$. As $a_0 \in P$ and P is a prime ideal, then $b_0 \in P$ or $c_0 \in P$. We cannot have both $b_0, c_0 \in P$ as $a_0 \notin P^2$. So let us assume $b_0 \in P$ but $c_0 \notin P$. Notice moreover that $1 \notin P$ as otherwise $P = R$ which is not possible as P is prime. Thus $b_kc_\ell \notin P$. Thus $b_k \notin P$ and $c_\ell \notin P$. Let us denote $0 \leq m \leq k < n$ the minimal integer such that $b_m \notin P$. Then we have:

$$a_m = \sum_{i+j=m} b_ic_j = b_mc_0 + b_{m-1}c_1 + \cdots b_0c_m.$$

As $0 \leq m \leq n-1$, we know $a_m \in P$. We also know that $b_i \in P$ for $i < m$. Thus $b_{m-1}c_1 + \cdots b_0c_m$ is in P. Thus $b_mc_0 \in P$. But this is impossible as $b_m \notin P$ and $c_0 \notin P$. □

A direct application of the previous criterion to $R = \mathbb{Z}$ leads to the following.

Corollary 4.9.24 (Eisenstein's criterion for $\mathbb{Z}$)**.** Let p be a prime in $\mathbb{Z}$. Let $f(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + x^n \in \mathbb{Z}[x]$ be a monic polynomial, $n \geq 1$. Suppose p divides a_i for all $0 \leq i \leq n-1$ but p^2 does not divide a_0. Then $f(x)$ is irreducible in $\mathbb{Z}[x]$ (and thus in $\mathbb{Q}[x]$ by Gauss's lemma).

Example 4.9.25. Let $f(x) = 5 + 10x + x^4 \in \mathbb{Z}[x]$. By Eisenstein's criterion, as 5 divides the non-leading coefficients, and $5^2 = 25$ does not divide the constant term, we get that $f(x)$ is irreducible.

Theorem 4.9.26 (Change of variable)**.** Suppose $\varphi : R \to S$ is an isomorphism. Then any induced ring homomorphism $\overline{\varphi} : R[x] \to S[x]$ where $\deg(\overline{\varphi}(x)) = 1$ preserves and conserves irreducible polynomials. In particular, for any $\alpha \in R$, then $f(x) \in R[x]$ is irreducible if and only if $f(x = \alpha) \in R[x]$ is irreducible.

Example 4.9.27. Let p be a prime number. Consider the polynomial $x^p - 1 \in \mathbb{Z}[x]$. Of course, $x = 1$ is a root, and thus we can factor out $x - 1$. Define the cyclotomic polynomial:

$$\Phi_p(x) = \frac{x^p - 1}{x - 1} = 1 + x + \cdots + x^{p-2} + x^{p-1} \in \mathbb{Z}[x].$$

We argue that $\Phi_p(x)$ is irreducible. It is equivalent to show that $\Phi_p(x + 1)$ is irreducible. We have:

$$\Phi_p(x + 1) = \frac{(x + 1)^p - 1}{x} = x^{p-1} + px^{p-2} + \cdots + \frac{p(p-1)}{2}x + p.$$

By Eisenstein's criterion, we obtain that $\Phi_p(x+1)$ is irreducible in $\mathbb{Z}[x]$. Thus the cyclotomic polynomial $\Phi_p(x)$ is irreducible in $\mathbb{Z}[x]$ and thus in $\mathbb{Q}[x]$ by Gauss's lemma.

CHAPTER 5

Modules

In the previous chapter, we talked about generalizing number systems and trying to do arithmetic in a more abstract setting, or so to say, to recover the minimal conditions which allow our usual intuition to work. Now that we have these number systems, we want to define the analog of linear algebra over them. The objects that replace vector spaces will be called modules.

5.1. Definition

5.1.1. Reminder: vector spaces.

Definition 5.1.1. Let F be a field. Then an *F-vector space* V is an abelian group $(V, +)$ together with *scalar multiplication* $\cdot\colon F \times V \to V$ such that the following conditions are satisfied for any $r, s \in F$ and $u, v \in V$:

(i) $r(u + v) = ru + rv$;
(ii) $(r + s)u = ru + su$;
(iii) $r(sv) = (rs)v$;
(iv) $1 \cdot v = v$.

Elements of V are called *vectors*.

We generally agree on the utility of studying linear algebra, since vector spaces appear naturally in various subjects.

Example 5.1.2. $\mathbb{C}$ can be realized as a two-dimensional $\mathbb{R}$-vector space.

Example 5.1.3. Solutions to homogeneous ODEs and PDEs form vector spaces. We use this structure of a vector space to, first, efficiently classify these solutions, and second, describe solutions of certain inhomogeneous equations.

Example 5.1.4. $\mathbb{C}$ can be realized as a two-dimensional $\mathbb{R}$-vector space.

Definition 5.1.5. Let $S \subset V$ be a subset of vectors, possibly infinite.

- We say that S is *linearly independent* if, whenever we have a finite linear combination
$$r_1 v_1 + \cdots + r_n v_n = 0$$
with $r_i \in F$ and $v_i \in S$, all the coefficients r_i are forced to be 0.
- We say that S *generates* V if any vector $u \in V$ can be written as a finite F-linear combination of vectors from S, i.e., there exist $r_i \in F$ and $v_i \in S$ such that
$$u = r_1 v_1 + \cdots + r_n v_n.$$
- We say that S is a *basis* of V if it is linearly independent and generates V.

Example 5.1.6. Any $S \subseteq V$ that contains the zero vector is not linearly independent, because we always get a nontrivial linear combination that is equal to 0:
$$1 \cdot 0 = 0.$$

Example 5.1.7. The empty set $\varnothing \subset V$ is always linearly independent. If $V = 0$, then it is a basis of V.

Theorem 5.1.8. Let V be a vector space over F. Then any linearly independent set $S \subseteq V$ can be extended to a basis. In particular, every vector space has a basis.

5.1.2. Modules.

Recall that we only consider associative unital rings. We will give the definition of modules, and notice how we only changed a few words and one symbol. This explains why you can think of modules as an attempt to do linear algebra over rings.

Definition 5.1.9. Let R be a ring. Then a *left R-module* V is an abelian group $(V, +)$ together with *scalar multiplication* $\cdot\colon R \times V \to V$ such that the following conditions are satisfied for any $r, s \in R$ and $u, v \in V$:

(i) $r(u+v) = ru + rv$;
(ii) $(r+s)u = ru + su$;
(iii) $r(sv) = (rs)v$;
(iv) $1 \cdot v = v$.

We give an analogous definition for the *right R-module*, where we say that scalars act on the right $\cdot\colon V \times R \to V$.

Unlike vectors in a vector space, elements of V no longer have a specific name.

Example 5.1.10. $\{0\}$ is a module over any ring R. It is called the *trivial module*, or the *zero module*.

Example 5.1.11. For any ring R, its additive group $(R, +)$ is an R-module, with the left action induced by multiplication in R. This module is called the *free module of rank* 1.

Example 5.1.12. Any left ideal $I \trianglelefteq R$ is a left R-module. Any right ideal is a right module.

Example 5.1.13. Given a ring homomorphism $\varphi\colon R \to S$, we can endow the additive group of S with the structure of an R-module as follows:

$$R \times S \longrightarrow S,$$
$$(r, s) \longmapsto \varphi(r) \cdot s.$$

Example 5.1.14. Any abelian group $(A, +)$ can be viewed as a $\mathbb{Z}$-module using the fact that we can take multiples of an element (see Definition 3.2.4 and Theorem 3.2.5):

$$\mathbb{Z} \times A \longrightarrow A,$$
$$(n, a) \longmapsto na.$$

Example 5.1.15. Let F be a field and V be an F-vector space together with a linear operator $\varphi\colon V \to V$. Then V can be endowed with the structure of an $F[x]$-module as follows:

$$F[x] \times V \longrightarrow V,$$
$$(f(x), v) \longmapsto f(\varphi)(v).$$

Notice that we use the fact that we can take powers of a linear operator from V to itself, and we set $\varphi^0 = \mathrm{id}_V$.

Remark 5.1.16. At this point, we can remember that we have a classification of finitely generated abelian groups (Theorem 3.12.2), and recall that over an algebraically closed field F, e.g., $F = \mathbb{C}$, we can write any square matrix in the Jordan normal form. In fact, these are instances of the same phenomenon that we will cover in this chapter—structure of finitely generated modules over PIDs.

5.2. Homomorphisms

5.2.1. Definition and basic properties.

Definition 5.2.1. A *homomorphism of R-modules*, or just a *homomorphism*, is a group homomorphism $\varphi\colon V \to W$ between R-modules V and W such that in addition φ is compatible with scalar multiplication, i.e., for each $r \in R$ and $v \in V$, we require $\varphi(rv) = r \cdot \varphi(v)$.

We denote the set of homomorphisms from V to W by $\mathrm{Hom}_R(V, W)$.

Lemma 5.2.2. If $\varphi\colon V \to W$ is a homomorphism, then $\varphi(0) = 0$.

Example 5.2.3. We have the *trivial*, or *zero*, homomorphism $0\colon V \to W$ defined as $0(v) = 0$ for any $v \in V$.

Example 5.2.4. The identity map $\mathrm{id}_V\colon V \to V$ is a homomorphism, called the *identity* homomorphism.

Definition 5.2.5. A subset $V' \subseteq V$ of a module is called a *submodule* if the natural inclusion map $V' \to V$ is a homomorphism.

Example 5.2.6. If F is a vector space, then the notion of an F-module is equivalent to the notion of an F-vector space, and a map between two vector spaces is a homomorphism if and only if it is a linear transformation.

Lemma 5.2.7. Given two homomorphisms $\varphi, \psi\colon V \to W$, their sum $\varphi + \psi$, defined on elements as

$$(\varphi + \psi)(v) = \varphi(v) + \psi(v),$$

is a homomorphism.

Proposition 5.2.8. $\mathrm{Hom}_R(V, W)$ is an abelian group.

Lemma 5.2.9. If R is commutative and $\varphi\colon V \to W$ is a homomorphism, then for any $r \in R$, the map $r\varphi$, defined on elements as

$$(r\varphi)(v) = r \cdot \varphi(v),$$

is a homomorphism.

Proposition 5.2.10. If R is commutative, then $\mathrm{Hom}_R(V, W)$ is an R-module.

⚠ **Warning 5.2.11.** The above proposition is no longer true when R is non-commutative.

5.2.2. Isomorphisms.

Definition 5.2.12. A homomorphism $\varphi\colon V \to W$ is an *isomorphism* if there is a homomorphism $\psi\colon W \to V$ which is inverse to φ, i.e., $\varphi\psi = \mathrm{id}_W$ and $\psi\varphi = \mathrm{id}_V$.

Proposition 5.2.13. A homomorphism $\varphi\colon V \to W$ is an isomorphism if and only if it is bijective.

5.2.3. Kernels and cokernels.

Definition 5.2.14. Let $\varphi\colon V \to W$ be a homomorphism of R-modules. We define its *kernel* as

$$\mathrm{Ker}\,\varphi = \{v \in V \mid \varphi(v) = 0\}.$$

Lemma 5.2.15. $\mathrm{Ker}\,\varphi$ is a submodule of V.

Proposition 5.2.16. φ is injective if and only if $\mathrm{Ker}\,\varphi = 0$.

Definition 5.2.17. Let $\varphi\colon V \to W$ be a homomorphism of R-modules. We define its *cokernel* as

$$\mathrm{Coker}\,\varphi = W/\mathrm{Im}\,\varphi.$$

Lemma 5.2.18. Consider a submodule $W' \subseteq W$ and the quotient abelian group W/W'. Then there is a natural R-module structure on W/W' that makes the canonical projection map $\gamma\colon W \to W/W'$ into a homomorphism. Further, $\mathrm{Ker}\,\gamma = W'$.

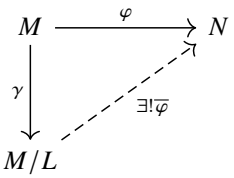
In particular, we see that $\mathrm{Coker}\,\varphi$ is naturally an R-module.

Proposition 5.2.19. φ is surjective if and only if $\mathrm{Coker}\,\varphi = 0$.

5.2.4. Universal properties.

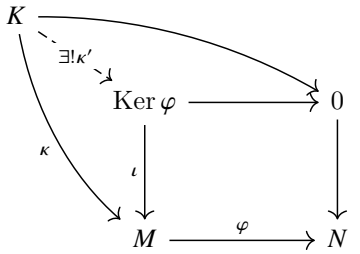
Proposition 5.2.20. Let $L \subseteq M$ be a submodule, and let $\gamma\colon M \to M/L$ denote the canonical projection map.

- **(i)** If $\varphi\colon M \to N$ is a homomorphism such that $L \subseteq \operatorname{Ker}\varphi$, then there exists a unique factoring $\overline{\varphi}\colon M/L \to N$ such that $\varphi = \overline{\varphi} \circ \gamma$.
- **(ii)** If Q is an R-module with a homomorphism $q\colon M \to Q$ such that $L \subseteq \operatorname{Ker} q$, and if it satisfies the above universal property, then there exists a unique isomorphism $Q \cong M/L$.



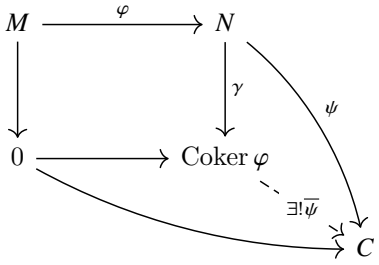
Proposition 5.2.21. Let $\varphi\colon M \to N$ be a homomorphism, and let $\iota\colon \operatorname{Ker}\varphi \to M$ denote the natural inclusion map.

- **(i)** If $\kappa\colon K \to M$ is a homomorphism such that $\varphi\kappa = 0$, then there exists a unique factoring $\kappa'\colon K \to \operatorname{Ker}\varphi$ such that $\kappa = \iota \circ \kappa'$.
- **(ii)** $\iota\colon \operatorname{Ker}\varphi \to M$ is a unique homomorphism, up to a unique isomorphism, that composes with φ to 0 and satisfies the universal property above.



Proposition 5.2.22. Let $\varphi\colon M \to N$ be a homomorphism, and let $\gamma\colon N \to \operatorname{Coker}\varphi$ denote the natural projection map.

- **(i)** If $\psi\colon N \to C$ is a homomorphism such that $\psi\varphi = 0$, then there exists a unique factoring $\overline{\psi}\colon \operatorname{Coker}\varphi \to C$ such that $\psi = \overline{\psi} \circ \gamma$.
- **(ii)** $\gamma\colon N \to \operatorname{Coker}\varphi$ is a unique homomorphism, up to a unique isomorphism, that composes with φ to 0 and satisfies the universal property above.



5.3. Constructions

5.3.1. Quotient modules. We have already proved in Lemma 5.2.18 that quotients of modules by submodules are modules as well.

5.3.2. Direct products. Let us fix an *indexing set* I, and for each $i \in I$, fix an R-module M_i.

Definition 5.3.1. The *direct product* of a set of modules $\{M_i \mid i \in I\}$ is the set of indexed tuples

$$\prod_{i\in I} M_i = \prod \{M_i \mid i \in I\} = \{(m_i)_{i\in I} \mid m_i \in M_i\}\,.$$

Addition and scalar multiplication are coordinate-wise.

Informally, we think of the elements of $\prod_{i\in I} M_i$ as "infinite vectors".

Example 5.3.2. If I is finite, e.g., $I = \{1, 2, 3\}$, then we can write

$$\prod_{i\in I} M_i = M_1 \times M_2 \times M_3 = \{(m_1, m_2, m_3) \mid m_1 \in M_1, m_2 \in M_2, m_3 \in M_3\}.$$

Addition and multiplication in this example look like this:

$$(m_1, m_2, m_3) + (n_1, n_2, n_3) = (m_1 + n_1, m_2 + n_2, m_3 + n_3),$$
$$r(m_1, m_2, m_3) = (rm_1, rm_2, rm_3).$$

Example 5.3.3. For every $j \in I$, we can define the jth *projection*

$$\pi_j \colon \prod_{i\in I} M_i \longrightarrow M_j,$$
$$(m_i)_{i\in I} \longmapsto m_j.$$

Each projection is a homomorphism of R-modules.

Proposition 5.3.4 (Universal property of products)**.** Let I be a set that indexes some R-modules M_i, $i \in I$, let P denote the product of these modules, and let $\pi_i \colon P \to M_i$ denote the ith projection.

- **(i)** For any test module T and any collection of homomorphisms $\tau_i \colon T \to M_i$, there is a unique homomorphism $\tau \colon T \to P$ that factors each τ_i through M_i, i.e., $\pi_i \tau = \tau_i$.
- **(ii)** P is the unique module, up to isomorphism, that satisfies the universal property above.

Slogan 5.3.5. Maps to products are defined by maps to each of the factors.

Remark 5.3.6. Product as defined above is actually a particular instance of a categorical construction. More precisely, product in an arbitrary category is a limit of a discrete (with no arrows) diagram consisting of some objects M_i.

5.3.3. Direct sums. As above, we fix an *indexing set* I, and a set of R-modules $\{M_i \mid i \in I\}$.

Definition 5.3.7. The *direct sum* of a set of modules $\{M_i \mid i \in I\}$ is the set of indexed tuples

$$\bigoplus_{i\in I} M_i = \bigoplus \{M_i \mid i \in I\} = \left\{ (m_i)_{i\in I} \;\middle|\; \begin{array}{c} m_i \in M_i, \text{ and only finitely many} \\ \text{of the } m_i\text{'s are nonzero} \end{array} \right\}.$$

Addition and scalar multiplication are coordinate-wise.

Informally, we think of the elements of $\bigoplus_{i \in I} M_i$ as "infinite vectors" with finitely many nonzero entries. We notice that $\bigoplus_{i \in I} M_i$ is naturally a submodule of $\prod_{i \in I} M_i$.

Example 5.3.8. If I is finite, e.g., $I = \{1, 2, 3\}$, then we can write

$$\bigoplus_{i \in I} M_i = M_1 \oplus M_2 \oplus M_3 = \{(m_1, m_2, m_3) \mid m_1 \in M_1, m_2 \in M_2, m_3 \in M_3\}.$$

Addition and multiplication in this example look like this:

$$(m_1, m_2, m_3) + (n_1, n_2, n_3) = (m_1 + n_1, m_2 + n_2, m_3 + n_3),$$
$$r(m_1, m_2, m_3) = (rm_1, rm_2, rm_3).$$

Lemma 5.3.9. If I is finite, then the natural inclusion map

$$\bigoplus_{i \in I} M_i \to \prod_{i \in I} M_i$$

is an isomorphism.

Example 5.3.10. In general, the natural map above is not an isomorphism. E.g., consider the case of $\mathbb{Z}$-modules:

$$\bigoplus_{i \in \mathbb{N}} \mathbb{Z} \to \prod_{i \in \mathbb{N}} \mathbb{Z}.$$

We can notice that $(1, 1, 1, \dots)$ is not an element of $\bigoplus_{i \in \mathbb{N}} \mathbb{Z}$. Moreover, there does not exist an isomorphism $\bigoplus_{i \in \mathbb{N}} \mathbb{Z} \to \prod_{i \in \mathbb{N}} \mathbb{Z}$ because the underlying set of the direct sum is countable, while the underlying set of the product is uncountable.

Example 5.3.11. For every $j \in I$, we can define the *inclusion of the jth summand*

$$\iota_j \colon M_j \longrightarrow \bigoplus_{i \in I} M_i,$$
$$m \longmapsto (m_i \mid m_j = m,\ m_i = 0 \text{ for } i \neq j)_{i \in I}.$$

Each ι_j is a homomorphism of R-modules.

Proposition 5.3.12 (Universal property of direct sums)**.** Let I be a set that indexes some R-modules M_i, $i \in I$, let S denote the direct sum of these modules, and let $\iota_i \colon M_i \to S$ denote the inclusion of the ith component.

- **(i)** For any test module T and any collection of homomorphisms $\tau_i \colon M_i \to T$, there is a unique homomorphism $\tau \colon S \to T$ that factors each τ_i through M_i, i.e., $\tau \iota_i = \tau_i$.
- **(ii)** T is the unique module, up to isomorphism, that satisfies the universal property above.

Slogan 5.3.13. Maps from direct sums are defined by maps from each of the summands.

Remark 5.3.14. Direct sum as defined above is also a particular instance of a categorical construction called a *coproduct*. More precisely, coproduct in an arbitrary category is a colimit of a discrete (with no arrows) diagram consisting of some objects M_i.

5.3.4. Sums and intersections of submodules. Let M be an R-module and $M_i \subseteq M$ a set of submodules indexed by $i \in I$.

Definition 5.3.15. The *sum* of the submodules $M_i \subseteq M$, $i \in I$, is the set of finite sums of elements from the M_i's:

$$\sum_{i \in I} M_i = \left\{ m_1 + \cdots + m_k \mid m_l \in M_{i_l} \right\}.$$

When I is finite, e.g., $I = \{1, \dots n\}$, we can write it as

$$M_1 + \cdots + M_n = \sum_{i \in I} M_i = \{m_1 + \cdots + m_n \mid m_i \in M_i\}.$$

Example 5.3.16. When $M = R$, you proved that submodules are exactly ideals of R. In this case, the sum defined above is the same as the sum of ideals.

Lemma 5.3.17. Let M be an R-module and $M_i \subseteq M$ a set of submodules indexed by $i \in I$. Then

(i) $\sum_{i \in I} M_i$ is a submodule of M;
(ii) $\cap_{i \in I} M_i$ is a submodule of M.

Lemma 5.3.18. Let M be an R-module and $M_i \subseteq M$ a set of submodules indexed by $i \in I$. Then there is a natural homomorphism

$$\varphi \colon \bigoplus_{i \in I} M_i \to M$$

with $\operatorname{Im} \varphi = \sum_{i \in I} M_i$.

5.4. Isomorphism theorems

Theorem 5.4.1 (First isomorphism theorem)**.** If $\varphi \colon M \to N$ is a homomorphism, then $\operatorname{Im} \varphi \cong M/\operatorname{Ker} \varphi$.

Theorem 5.4.2 (Second isomorphism theorem)**.** If K and L are submodules of a module M, then

$$K + L/L \cong K/K \cap L.$$

Theorem 5.4.3 (Third isomorphism theorem)**.** If $K \subseteq L \subseteq M$ is a sequence of submodules in M, then

$$M/K/L/K \cong M/L.$$

Theorem 5.4.4 (Fourth isomorphism theorem, or Correspondence theorem)**.** If K is a submodule of M, then we have a natural bijection

$$\left\{ \begin{array}{c} \text{submodules of } M \\ \text{containing } K \end{array} \right\} \longleftrightarrow \{\text{submodules of } M/K\}.$$

Now let us see how isomorphism theorems can be applied to compare some of the constructions.

Proposition 5.4.5. Let M be a module and $K, L \subseteq M$ be two submodules. If $K + L = M$ and $K \cap L = 0$, then $M \cong K \oplus L$.

5.5. Free modules

From linear algebra, we know that every vector space admits a basis. However, the same is not going to be true for modules. Modules that admit a basis form an especially well-behaved class, and are called *free modules*.

5.5.1. Linear independence and generation. Let us consider direct sums of very specific R-modules: $M_i = R$ for $i \in I$. We will denote the resulting direct sum by

$$R^{\oplus I} = R^I = \bigoplus_{i \in I} R.$$

Denote the standard basis vectors in R^I by e_i, then any tuple in R^I can be conveniently written as a finite linear combination of e_i's.

Example 5.5.1. Let S be a subset of M. Then we have a natural map

$$\varphi_S \colon R^S \to M$$

which is defined on the component indexed by $s \in S \subseteq M$ as

$$\begin{aligned} \varphi_{S,s} \colon R &\longrightarrow M, \\ r &\longmapsto rs. \end{aligned}$$

By the universal property of direct sums (Proposition 5.3.12), this defines φ_S uniquely.

Definition 5.5.2. Let S be a subset of M. We say that

- S *generates* M if the natural map φ_S is a surjection;
- S is *linearly independent* if φ_S is an injection;
- S is a *basis* if it is linearly independent and generates M, or in other words, if φ_S is an isomorphism.

Notation 5.5.3. We denote the submodule generated by S by any of the following: RS, $\langle S \rangle$, $\operatorname{Span} S$.

Definition 5.5.4. We say that M is *finitely generated* if it admits a finite generating set $S \subseteq M$.

Example 5.5.5. Finitely generated ideals are finitely generated modules.

⚠ **Warning 5.5.6.** A submodule of a finitely generated module does not have to be finitely generated. For example, we can consider the ring $R = F[x_1, x_2, \dots]$ of polynomials in countably many variables over some field F. Then R considered as a module is finitely generated, while the ideal $I = (x_1, x_2, \dots)$ generated by all variables is not finitely generated.

Definition 5.5.7. We say that a module M is *cyclic* if it can be generated by one element.

5.5.2. Free modules and rank.

Definition 5.5.8. We say that an R-module is *free* if it admits a basis $S \subset M$. The *rank* of M is defined as the cardinality of S, and is denoted $\operatorname{rk} M = |S|$.

Theorem 5.5.9. Let R be a commutative ring, then rank of R-modules is well-defined. More precisely, if a module M admits two bases S_1 and S_2, then they have the same cardinality $|S_1| = |S_2|$.

5.5.3. Presentation of a module.

Definition 5.5.10. Let M be an R-module.

- A generating set $A \subset M$ of an R-module M defines a surjection $\pi\colon R^A \to M$. We call R^A a *module of generators* of M.
- The kernel of the surjection π from the previous paragraph is another R-module, so $\operatorname{Ker}\pi$ has a set of generators $B \subset \operatorname{Ker}\pi$, and we get a surjection $R^B \to \operatorname{Ker}\pi$. We call R^B the *module of relations* of M.

Pick a set of generators A and relations B of M, then we have surjections $\pi\colon R^A \to M$ and $\rho\colon R^B \to \operatorname{Ker}\pi$. The homomorphism ρ naturally defines a homomorphism $R^B \to R^A$. We then have that $M \cong \operatorname{Coker}(R^B \to R^A)$, and thus any module can be constructed by generators and relations.

Definition 5.5.11. Identifying a module M with $\operatorname{Coker}(R^B \to R^A)$ is called a *presentation* of M.

5.6. Homological algebra 101

5.6.1. Exactness. Let R be an associative unital ring.

Definition 5.6.1. A sequence of R-modules

$$M' \xrightarrow{\iota} M \xrightarrow{\pi} M''$$

is called *exact in* M if $\operatorname{Im}\iota = \operatorname{Ker}\pi$.

Exercise 5.6.2. $\pi\iota = 0$ if and only if $\operatorname{Im}\iota \subseteq \operatorname{Ker}\pi$.

Example 5.6.3. For any R-module M, choose its presentation $R^B \to R^A \to M$. Then the sequence

$$R^B \to R^A \to M \to 0$$

is exact in R^A and M.

Exercise 5.6.4 (Easy but thought-provoking)**.** Choosing a presentation of M in terms of generators and relations is equivalent to choosing an exact sequence

$$R^B \to R^A \to M \to 0.$$

Definition 5.6.5. We say that

$$0 \to M' \xrightarrow{\iota} M \xrightarrow{\pi} M'' \to 0$$

is a *short exact sequence* if it is exact in each term.

Exercise 5.6.6. The sequence above is exact in M' if and only if ι is injective. It is exact in M'' if and only if π is surjective.

Theorem 5.6.7. Take a short exact sequence of R-modules

$$0 \to M' \xrightarrow{\iota} M \xrightarrow{\pi} M'' \to 0.$$

Then for any module N, we get natural sequences

$$0 \to \operatorname{Hom}_R(N, M') \xrightarrow{\iota_*} \operatorname{Hom}_R(N, M) \xrightarrow{\pi_*} \operatorname{Hom}_R(N, M'')$$

and

$$0 \to \operatorname{Hom}_R(M'', N) \xrightarrow{\pi^*} \operatorname{Hom}_R(M, N) \xrightarrow{\iota^*} \operatorname{Hom}_R(M', N).$$

which are exact (in the left and the middle terms).

Remark 5.6.8. Borrowing terminology from homological algebra, we say that Hom_R is left exact: it only preserves exactness in the left and middle terms, but not necessarily in the right term.

5.6.2. Projective modules.

Definition 5.6.9. A module P is called *projective* if for any short exact sequence

$$0 \to M' \xrightarrow{\iota} M \xrightarrow{\pi} M'' \to 0,$$

the induced sequence

$$0 \to \operatorname{Hom}_R(P, M') \xrightarrow{\iota_*} \operatorname{Hom}_R(P, M) \xrightarrow{\pi_*} \operatorname{Hom}_R(P, M'') \to 0$$

is exact.

Remark 5.6.10. In other words, P is projective if $\operatorname{Hom}_R(P, \_)$ is exact.

Definition 5.6.11. Let $\varphi\colon M \to N$ be a homomorphism. We say that $\sigma\colon N \to M$ is a *section* of φ if $\varphi\sigma = \mathrm{id}_N$.

Exercise 5.6.12. **(i)** Show that if φ admits a section, then φ is surjective.
(ii) Show that if a module P is projective, then any surjective homomorphism $M \to P$ admits a section.

Proposition 5.6.13. Let M be an R-module and P a projective R-module. Then every surjection

$$\pi \colon M \to P$$

looks like a projection out of a direct sum decomposition. More precisely, for any section σ of π, the natural map

$$\operatorname{Ker} \pi \oplus \operatorname{Im} \sigma \to M$$

is an isomorphism.

Lemma 5.6.14. Consider a set of R-modules P_i indexed by $i \in I$. Then $P = \bigoplus_{i \in I} P_i$ is projective if and only if each of the summands P_i is projective.

Lemma 5.6.15. **(i)** The free rank one module R is projective.
(ii) Any free module is projective.

Theorem 5.6.16. An R-module P is projective if and only if it is a direct summand of a free module. In other words, if and only if there exists another R-module Q (necessarily projective) such that $P \oplus Q \cong R^I$ for some indexing set I.

Corollary 5.6.17. Let R be a ring. Then any surjection $\pi \colon M \to F$ onto a free module F implies that there is an isomorphism $M \cong \operatorname{Ker} \pi \oplus F$.

5.7. Structure of finitely generated modules

5.7.1. Integral domains. Recall that an element $m \in M$ of an R-module is called *torsion* if there is a nonzero $r \in R \setminus \{0\}$ such that $rm = 0$. A module M is called *torsion-free* if the only torsion element in M is 0.

Lemma 5.7.1. Let R be an integral domain and M an R-module. Then the set of torsion elements forms a submodule.

Proof. Let $m \in M$ be a torsion element, and let $r \in R \setminus \{0\}$ be such that $rm = 0$. Then for any $s \in R$, we check that sm is torsion because

$$r \cdot sm = s \cdot rm = s \cdot 0 = 0.$$

If $n \in M$ is another torsion element such that $tn = 0$ for $t \in R\{0\}$, then we observe that $rt \neq 0$ because R is a domain, and

$$rt(m + n) = trm + rtn = t0 + r0 = 0.$$

□

Definition 5.7.2.

- Over an integral domain R, we define the *torsion submodule* $\operatorname{Tors}_R M = \operatorname{Tors} M \subseteq M$ as the set of torsion elements.
- We can observe that M is torsion-free if and only if $\operatorname{Tors} M = 0$. We say that M is a *torsion* module if $\operatorname{Tors} M = 0$.

Lemma 5.7.3. Let R be an integral domain and M an R-module. Then $M/\operatorname{Tors} M$ is torsion-free.

Proposition 5.7.4. Let M be a finitely generated module over an integral domain R. Then there exists an embedding $M \to R^n$ into a finite free module R^n for some $n \in \mathbb{N}$.

PROOF. Since M is finitely generated, we can choose d generators

$$M = \langle x_1, \dots, x_d \rangle.$$

These define a surjection $\pi\colon R^d \to M$.

We prove that if $A \subset M$ is a linerly independent set, then $|A| \leq d$. Indeed, a linearly independent set defines an injection $\iota\colon R^A \to M$, and since R^A is projective, it lifts to $\bar{\iota}\colon R^A \to R^d$. Now, $\bar{\iota}$ must be injective too. Let $K = \operatorname{Frac} R$ be the fraction field of R, then $\bar{\iota}$ defines an injection of vector spaces $K^A \to K^d$, hence $|A| \leq d$.

Since there is a bound, we conclude that there is a linearly independent set of M of maximal size, denote it by $A = \{v_1, \dots, v_n\}$. For any $x \in M$, we get that $A \sqcup \{x\}$ is linearly dependent by maximality of A, hence we get a relation

$$ax = b_1 v_1 + \dots b_n v_n$$

with $a \neq 0$, which means that $ax \in \langle A \rangle$.

Applying this to the generators $\{x_1, \dots, x_d\}$, we get that $a_1 x_1, \dots, a_d x_d \in \langle A \rangle$, where all $a_i \neq 0$. Now set $a = a_1 \cdots a_d$, which is nonzero because R is a domain, so we get that $aM = a\langle x_1, \dots, x_d \rangle \subseteq \langle A \rangle \cong R^A$. □

5.7.2. Principal ideal domains.

Theorem 5.7.5. Over a PID, a submodule of a finitely generated free module is free of the same or smaller rank.

Remark 5.7.6. In fact, over a PID, a submodule of any free module is free, without the finite generation assumption. But the proof is more involved.

PROOF. Consider a submodule $M \subseteq R^n$. We prove by induction on n and notice that the case $n = 1$ is trivial because R is a PID.

Assume we know the statement up to n, and now we want to consider $M \subset R^{n+1}$. Denote the embedding by ι, and consider the projection $\pi\colon R^{n+1} \to R^n$ that forgets the first component.

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \operatorname{Ker}\pi = R & \longrightarrow & R^{n+1} = R \oplus R^n & \xrightarrow{\pi} & R^n & \longrightarrow & 0 \\ & & \uparrow & & \uparrow \iota & & \uparrow & & \\ 0 & \longrightarrow & \operatorname{Ker}(\pi\iota) & \longrightarrow & M & \longrightarrow & \pi(M) \cong R^k & \longrightarrow & 0 \end{array}$$

Now $\pi(M)$ is a submodule of R^n, so by induction hypothesis, it is free of rank $k \leq n$. Further, since ι is injective, we get that $\operatorname{Ker}(\pi\iota) \subset \operatorname{Ker}\pi = R$, hence free of rank 1 or 0. Finally, by Corollary 5.6.17, we know that any surjection onto a free module yields a direct sum decomposition

$$M \cong \operatorname{Ker}(\pi\iota) \oplus \pi(M) \cong \operatorname{Ker}(\pi\iota) \oplus R^k.$$

Therefore M is free of rank at most $k + 1 \leq n + 1$. □

Corollary 5.7.7. Let R be a PID and M a finitely generated module.

(i) If M is torsion-free, then M is free of finite rank.

(ii) If we have a sequence of injective homomorphisms

$$R^n \to M \to R^n,$$

then $M \cong R^n$.

Example 5.7.8. When R is not a PID, the last corollary does not hold. For example, we can take $R = K[x, y]$, where K is a field. Then $R \cong (x) \subset (x, y) \subset R$, but (x, y) is not free.

Corollary 5.7.9. Let M be a finitely generated module over a PID. Then $M \cong T \oplus F$, where T is a torsion module and F is a finite rank free module.

Proof. Set $T = \operatorname{Tors} M$ and $F = M/T$. By Corollary 5.7.7**(i)**, F is free of finite rank. Since we have a surjection $M \to F$ with kernel T, we apply Corollary 5.6.17 and get $M \cong T \oplus F$. □

5.7.3. Classification of finitely generated modules over PIDs. Given any module M over a PID, its free part is defined canonically as $M/\operatorname{Tors} M$. It turns out that we can also explicitly describe $\operatorname{Tors} M$ and even decompose it into a direct sum of cyclic modules uniquely up to permutation.

Theorem 5.7.10 (Structure of finitely generated modules over PIDs)**.** Let R be a PID and M an R-module. Then

$$M \cong R^n \oplus \bigoplus_{i=1}^{N} R/\left(p_i^{k_i}\right)$$

for some $n, N, k_i \in \mathbb{N}$, $k_i \geq 1$, and prime elements $p_i \in R$. Moreover, this decomposition is unique up to multiplying the p_i's by units and relabelling.

Proof. By Corollary 5.7.9, $M \cong T \oplus F$ for a torsion module T and a free module F. Moreover, $\operatorname{rk} F = \operatorname{rk}(M/\operatorname{Tors} M)$, so the number n in the decomposition in the statement is defined uniquely.

This reduces the theorem to the case when $M = T$ is a finitely generated torsion module. In this case, pick a set of m generators for M, thus obtaining a surjection $\pi\colon R^m \to M$. By Theorem 5.7.5, $\operatorname{Ker}\pi$ is free of rank $k \leq m$. Since we assumed that M is torsion, we conclude $m = k$. It means that we have a short exact sequence

$$0 \to \operatorname{Ker}\pi \xrightarrow{\iota} R^m \xrightarrow{\pi} M \to 0.$$

By the Aligned Bases theorem Theorem 5.7.11, whose proof we delay, there exists a basis $\{v_1, \ldots, v_m\}$ of R^m and constants $a_1, \ldots, a_m \in R$ such that $\{a_1 v_1, \ldots, a_m v_m\}$ is a basis for the submodule $R^m \cong \operatorname{Ker}\pi \subset R^m$. Choosing this basis, we can assume without loss of generality that ι is represented by a diagonal matrix with nonzero entries $a_1, \cdots, a_k$ on the diagonal.

We then get

$$\begin{aligned} M \cong R^m/\operatorname{Ker}\pi &\cong \frac{Rv_1 \oplus \cdots \oplus Rv_m}{Ra_1v_1 \oplus \cdots \oplus Ra_mv_m} \\ &\cong \frac{Rv_1}{Ra_1v_1} \oplus \cdots \oplus \frac{Rv_m}{Ra_mv_m} \cong \frac{R}{(a_1)} \oplus \cdots \oplus \frac{R}{(a_m)}. \end{aligned}$$

□

Theorem 5.7.11 (Aligned bases theorem)**.**

CHAPTER 6

Algebras

6.1. Definition

CHAPTER 7

Galois theory

It is perhaps not said often, but the approach of Galois theory is extremely categorical philosophically. The main goal is to understand fields. As a motivation, the more we know about a field $\mathbb{F}$, the more we know about its polynomial ring $\mathbb{F}[x]$. Therefore, we are aiming at looking at how fields interact with each other: given two fields $\mathbb{F}$ and $\mathbb{E}$, what are all the ring homomorphisms $\mathbb{F} \to \mathbb{E}$?

7.1. Field extensions

Recall that we say that a ring $\mathbb{F}$ is a field when $\mathbb{F}^\times$ is a group.

Example 7.1.1. Recall our common examples of fields: $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$.

Example 7.1.2. For any prime number p, $\mathbb{Z}/p$ is a field, and we denote it by $\mathbb{F}_p$.

Example 7.1.3. Recall that for any integral domain R, we defined its field of fractions $\operatorname{Frac} R$. Given a field $\mathbb{F}$, we used this construction to define the *field of rational functions* $\mathbb{F}(t) = \operatorname{Frac} \mathbb{F}[t]$.

Lemma 7.1.4. A ring homomorphism $\mathbb{F} \to \mathbb{E}$ between fields is injective.

Proof. Given a ring homomorphism $\varphi\colon \mathbb{F} \to \mathbb{E}$, we must have

$$\varphi(0_\mathbb{F}) = 0_\mathbb{E}, \qquad \varphi(1_\mathbb{F}) = 1_\mathbb{E}.$$

Moreover, $0_\mathbb{F} \neq 1_\mathbb{F}$ and $0_\mathbb{E} \neq 1_\mathbb{F}$, hence φ is not the trivial homomorphism. In particular, $\ker(f) \neq \mathbb{F}$. But any field $\mathbb{F}$ is a simple ring, i.e., it has exactly two ideals: $\{0\}$ and $\mathbb{F}$. Since $\ker(\varphi)$ is an ideal, we get $\ker(\varphi) = \{0\}$ and conclude that φ is injective. □

Therefore, a ring homomorphism $\mathbb{F} \to \mathbb{E}$ between fields is the same as a choice of a subfield $\mathbb{F}$ of $\mathbb{E}$. We shall denote it $\mathbb{F} \hookrightarrow \mathbb{E}$.

Definition 7.1.5. Given a field $\mathbb{F}$, an *extension field of* $\mathbb{F}$ is a field $\mathbb{E}$ such that $\mathbb{F}$ is a subfield of $\mathbb{E}$. We call $\mathbb{F} \hookrightarrow \mathbb{E}$ a *field extension*.

Recall that $\mathbb{Z}$ is an initial ring, so there exists a unique ring homomorphism $\varphi\colon \mathbb{Z} \to \mathbb{F}$. Let $\ker \varphi = (n)$ for a nonnegative $n \in \mathbb{N}$. We call $\operatorname{Char} \mathbb{F} = n$ the *characteristic* of the field $\mathbb{F}$. Recall that we proved that the characteristic of any integral domain is a prime number.

Lemma 7.1.6. Let $\mathbb{F}$ be a field, then we have two distinct possibilities:

(i) $\operatorname{Char}(\mathbb{F}) = 0$ and $\mathbb{Q}$ is a subfield of $\mathbb{F}$;
(ii) $\operatorname{Char}(\mathbb{F}) = p$, where p is prime, and $\mathbb{F}_p$ is a subfield of $\mathbb{F}$.

Proof. Consider the unique ring homomorphism $\varphi\colon \mathbb{Z} \to \mathbb{F}$. If $\ker(\varphi) = 0$, then $\operatorname{Char}(\mathbb{F}) = 0$ and $\mathbb{Z}/\ker(\varphi) = \mathbb{Z}$ is (isomorphic to) a subring of $\mathbb{F}$ by the first isomorphism theorem (Theorem 4.3.38). Since $\mathbb{F}$ is a field, it must also contains all inverses of $\mathbb{Z}$, hence $\mathbb{Q}$ is a subfield of $\mathbb{F}$. If $\ker(\varphi) = (n) \neq 0$, then $\mathbb{Z}/n$ is (isomorphic to) a subring of $\mathbb{F}$ and hence is an integral domain. Thus (n) is a prime ideal, so $(n) = (p)$ for a prime number p, and $\operatorname{Char}(\mathbb{F}) = p$. Further, $\mathbb{F}_p = \mathbb{Z}/p \cong \operatorname{Im} \varphi$ is a subfield of $\mathbb{F}$, again by the first isomorphism theorem (Theorem 4.3.38). □

Exercise 7.1.7. Show there are no homomorphisms $\mathbb{Z}/p\mathbb{Z} \to \mathbb{Z}/q\mathbb{Z}$ where p and q are distinct prime numbers. Similarly, show that there are no homomorphisms $\mathbb{Z}/p\mathbb{Z} \to \mathbb{Z}$.

Exercise 7.1.8. Even though there exists a homomorphism $\mathbb{Z} \to \mathbb{F}_p$, show that there are no ring homomorphisms $\mathbb{Q} \to \mathbb{F}_p$.

Exercise 7.1.9. Use the previous two exercises to conclude that if $\mathbb{F}$ and $\mathbb{E}$ are fields with different characteristics, then there are no homomorphisms $\mathbb{F} \to \mathbb{E}$.

Therefore, in our study of fields and their interactions, it is enough to study fields of one characteristic at a time.

Lemma 7.1.10. Given a field extension $\varphi\colon \mathbb{F} \hookrightarrow \mathbb{E}$, we have that $\mathbb{E}$ is an $\mathbb{F}$-vector space.

PROOF. Since $\mathbb{E}$ is a field, it is in particular an abelian group. Multiplication by $\mathbb{F}$-scalars is induced by φ:

$$\begin{aligned} \mathbb{F} \times \mathbb{E} &\longrightarrow \mathbb{E} \\ (k, a) &\longmapsto \varphi(k)a. \end{aligned}$$

Since $\mathbb{E}$ is a field, the associativity and distributivity axioms prove the respective axioms of $\mathbb{F}$-modules. □

Definition 7.1.11. Given a field extension $\mathbb{F} \hookrightarrow \mathbb{E}$, we denote by $[\mathbb{E} : \mathbb{F}] = \dim_{\mathbb{F}} \mathbb{E}$ the dimension of $\mathbb{E}$ as an $\mathbb{F}$-vector space. We refer to it as the *degree* of the extension. If $[\mathbb{E} : \mathbb{F}]$ is finite, we say the field extension is *finite*.

Example 7.1.12. From the usual inclusion $\mathbb{R} \subseteq \mathbb{C}$, we have $[\mathbb{C} : \mathbb{R}] = 2$.

Example 7.1.13. From the usual inclusion $\mathbb{Q} \subseteq \mathbb{R}$, we have $[\mathbb{R} : \mathbb{Q}] = \infty$.

Example 7.1.14. Let $\mathbb{F}$ be a field and consider its polynomial ring $\mathbb{F}[x]$. Its fraction field $\mathbb{F}(x)$ defines an infinite field extension $\mathbb{F} \hookrightarrow \mathbb{F}(x)$, i.e., $[\mathbb{F}(x) : \mathbb{F}] = \infty$.

Notation 7.1.15. Given a field extension $\mathbb{F} \hookrightarrow \mathbb{E}$, let $\alpha \in \mathbb{E}$. We have the evaluation ring homomorphism $\mathrm{ev}_\alpha\colon \mathbb{F}[x] \to \mathbb{E}$. We denote by $\mathbb{F}[\alpha]$ the image of ev_α: it is the smallest ring in $\mathbb{E}$ containing both $\mathbb{F}$ and α. Denote by $\mathbb{F}(\alpha)$ the fraction field of $\mathbb{F}[\alpha]$. By the universal property of fraction fields, the inclusion $\mathbb{F}[\alpha] \hookrightarrow \mathbb{E}$ uniquely factors through $\mathbb{F}(\alpha)$, and we identify it with a subfield of $\mathbb{E}$. Then $\mathbb{F}(\alpha)$ is the smallest field containing $\mathbb{F}$ and α.

Example 7.1.16. We have $\mathbb{R}[i] = \mathbb{R}(i) = \mathbb{C}$. We shall see below more examples where $\mathbb{F}[\alpha] = \mathbb{F}(\alpha)$.

Notation 7.1.17. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be a field extension. Let $\alpha_1, \ldots, \alpha_n \in \mathbb{E}$. Define inductively

$$\mathbb{F}[\alpha_1, \ldots, \alpha_n] = \big(\mathbb{F}[\alpha_1, \ldots, \alpha_{n-1}]\big)[\alpha_n],$$

and let $\mathbb{F}(\alpha_1, \ldots, \alpha_n)$ be its fraction field.

Definition 7.1.18. A field extension $\mathbb{F} \hookrightarrow \mathbb{E}$ is called *simple* if there exists $\alpha \in \mathbb{E}$ such that $\mathbb{E} = \mathbb{F}(\alpha)$.

7.1.1. Algebraic and transcendental elements.

Definition 7.1.19. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be a field extension and fix an element $\alpha \in \mathbb{E}$.

- If $\mathrm{ev}_\alpha\colon \mathbb{F}[x] \to \mathbb{F}(\alpha)$ is injective, we say that α is *transcendental over* $\mathbb{F}$. In other words, we have $\ker(\mathrm{ev}_\alpha) = 0$, hence $\alpha \in \mathbb{E}$ is not the root of any polynomial in $\mathbb{F}[x]$. Moreover, by the first isomorphism theorem, we get $\mathbb{F}[x] \cong \mathbb{F}[\alpha]$ as rings. Thus $\mathbb{F}[\alpha]$ cannot be a field and $[\mathbb{F}(\alpha) : \mathbb{F}] = \infty$.
- If $\mathrm{ev}_\alpha\colon \mathbb{F}[x] \to \mathbb{F}(\alpha)$ is not injective, we say that α *is algebraic over* $\mathbb{F}$. In this case, $\ker(\mathrm{ev}_\alpha)$ is a non-trival ideal of $\mathbb{F}[x]$. As it is a PID, we get $\ker(\mathrm{ev}_\alpha) = (f(x))$ for some $f(x) \in \mathbb{F}[x]$. If we require $f(x)$ to be monic, the choice is unique. We call it the *minimal polynomial of* α *over* $\mathbb{F}$, and we denote it $m_{\alpha,\mathbb{F}}(x) \in \mathbb{F}[x]$.

Example 7.1.20. Numbers $\pi, e \in \mathbb{R}$ are transcendental over $\mathbb{Q}$. This is a non-trivial fact.

Example 7.1.21. Let us give examples of algebraic numbers over $\mathbb{Q}$.

- The element $i \in \mathbb{C}$ is algebraic over $\mathbb{Q}$ with $m_{i,\mathbb{Q}}(x) = x^2 + 1$.
- The element $\sqrt{2} \in \mathbb{R}$ is algebraic over $\mathbb{Q}$ with $m_{\sqrt{2},\mathbb{Q}}(x) = x^2 - 2$.
- The element $\sqrt[7]{181}$ is algebraic over $\mathbb{Q}$ with $m_{\sqrt[7]{181},\mathbb{Q}}(x) = x^7 - 181$.

Theorem 7.1.22. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be a field extension, and suppose $\alpha \in \mathbb{E}$ is algebraic over $\mathbb{F}$. We then have the following.

(i) $\mathbb{F}[\alpha] = \mathbb{F}(\alpha)$.
(ii) The minimal polynomial $m_{\alpha,\mathbb{F}}(x)$ is irreducible over $\mathbb{F}$.
(iii) Let $n = \deg m_{\alpha,\mathbb{F}}$, then $\{1, \alpha, \alpha^2, \ldots, \alpha^{n-1}\}$ is an $\mathbb{F}$-basis of $\mathbb{F}(\alpha)$, and the degree of the field extension equals the degree of the minimal polynomial: $[\mathbb{F}(\alpha) : \mathbb{F}] = n$.

Proof. First, notice that $\mathbb{F}[x]/(m_{\alpha,\mathbb{F}})$ is an integral domain as it is isomorphic to a subring of $\mathbb{F}(\alpha)$ by the first isomorphism theorem:

$$\mathbb{F}[x]/(m_{\alpha,\mathbb{F}}) \cong \mathbb{F}[\alpha] \subseteq \mathbb{F}(\alpha).$$

Thus $m_{\alpha,\mathbb{F}}$ is irreducible, and since $\mathbb{F}[x]$ is a PID, $m_{\alpha,\mathbb{F}}$ spans a maximal ideal (Corollary 4.4.23). This implies that $\mathbb{F}[x]/(m_{\alpha,\mathbb{F}})$ is a field, hence $\mathbb{F}[\alpha] = \mathbb{F}(\alpha)$. A basis of $\mathbb{F}[x]/(m_{\alpha,\mathbb{F}})$ is given by:

$$\{1, \overline{x}, \ldots, \overline{x}^{n-1}\}.$$

As $\mathrm{ev}_\alpha(\overline{x}) = \alpha$, we obtain the desired result. □

Definition 7.1.23. Let $\alpha \in \mathbb{E}$ be algebraic over $\mathbb{F}$. By Theorem 7.1.22, the following quantities are all equal: $\deg m_{\alpha,\mathbb{F}} = [\mathbb{F}(\alpha) : \mathbb{F}] = \dim_{\mathbb{F}} \mathbb{F}(\alpha)$. We will call this number the *degree* of α and denote it by $\deg_{\mathbb{F}} \alpha$, or $\deg \alpha$ when the ground field $\mathbb{F}$ is implicitly understood.

Suppose α is algebraic over $\mathbb{F}$ of degree n. Then by Theorem 7.1.22, every element of the field $\mathbb{F}(\alpha)$ can be written as a unique $\mathbb{F}$-linear combination of powers of α:

$$a_0 + a_1\alpha + \cdots + a_{n-1}\alpha^{n-1}.$$

To understand multiplication in $\mathbb{F}(\alpha)$, we use the fact that $m_{\alpha,\mathbb{F}}(\alpha) = 0$.

Example 7.1.24. For instance, for $\alpha = i \in \mathbb{C}$ over $\mathbb{Q}$, the fact that $m_{i,\mathbb{Q}}(x) = x^2 + 1$ means that $i^2 + 1 = 0$ and so $i^2 = -1$.

Conversly, given a field $\mathbb{F}$ and an irreducible monic polynomial $f(x) \in \mathbb{F}[x]$, we get a field extension by considering $\mathbb{F}[x]/(f(x))$, in which the equivalence class $[x]$ can be denoted by α and is a root of $f(x)$. Thus $\mathbb{F}[x]/(f(x)) = \mathbb{F}(\alpha)$ and $f(x) = m_{\alpha,\mathbb{F}}(x)$.

Example 7.1.25. Suppose $\mathbb{F} = \mathbb{Z}/2\mathbb{Z}$. Then $x^2 + x + 1 \in \mathbb{F}[x]$ is irreducible as it has no roots. Thus $\mathbb{E} = \mathbb{F}[x]/(x^2 + x + 1)$ is a field. Denote by α the equivalence class of x in $\mathbb{E}$. We get $m_{\alpha,\mathbb{F}}(x) = x^2 + x + 1 \in \mathbb{F}[x]$ and

$$\mathbb{F}(\alpha) = \{a_0 + a_1\alpha | a_i \in \mathbb{F}\} = \{0, 1, \alpha, 1 + \alpha\}.$$

Since $\alpha^2 + \alpha + 1 = 0$, we get $\alpha^2 = \alpha + 1$ and $\alpha(\alpha + 1) = 1$.

7.1.2. Algebraic extensions.

Definition 7.1.26. A field extension $\mathbb{F} \hookrightarrow \mathbb{E}$ is called *algebraic* if any element $\alpha \in \mathbb{E}$ is algebraic over $\mathbb{F}$.

Lemma 7.1.27. A finite field extension is algebraic.

PROOF. Suppose $\mathbb{F} \hookrightarrow \mathbb{E}$ is a field extension. We argue by contradiction and assume that there exists a transcendental element $\alpha \in \mathbb{E}$. But then $[\mathbb{F}(\alpha) : \mathbb{F}] = \infty$, so if the extension is not algebraic, it cannot be finite. □

⚠ **Warning 7.1.28.** The converse is not true. There exist algebraic extensions that are not finite.

We can improve the above result as follows.

Exercise 7.1.29. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be a field extension. Show that $[\mathbb{E} : \mathbb{F}] < \infty$ if and only if $\mathbb{E} = \mathbb{F}(\alpha_1, \ldots, \alpha_n)$ for $\alpha_i \in \mathbb{E}$ algebraic over $\mathbb{F}$.

Exercise 7.1.30. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ and $\mathbb{E} \hookrightarrow \mathbb{K}$ be finite field extensions. Then

$$[\mathbb{K} : \mathbb{F}] = [\mathbb{K} : \mathbb{E}][\mathbb{E} : \mathbb{F}].$$

Lemma 7.1.31. Let $\mathbb{F} \hookrightarrow \mathbb{E} \hookrightarrow \mathbb{K}$ be field extensions. If $\mathbb{K}$ is algebraic over $\mathbb{E}$ and $\mathbb{E}$ is algebraic over $\mathbb{F}$, then $\mathbb{K}$ is algebraic over $\mathbb{F}$.

PROOF. Let $\beta \in \mathbb{K}$. Then β is algebraic over $\mathbb{E}$, and so there exists an irreducible monic polynomial $f(x) \in \mathbb{E}[x]$ such that $f(\beta) = 0$. In particular, there exist coefficients $\alpha_i \in \mathbb{E}$ such that

$$f(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_{n-1} x^{n-1} + x^n \in \mathbb{E}[x].$$

As each α_i is algebraic over $\mathbb{F}$ we get that $\mathbb{F}(\alpha_0, \ldots, \alpha_{n-1})$ is a finite extension of $\mathbb{F}$. Thus we obtain:

$$\begin{array}{c} \mathbb{F}(\alpha_0, \ldots, \alpha_{n-1})(\beta) \\ \uparrow \\ \mathbb{F}(\alpha_0, \ldots, \alpha_{n-1}) \\ \uparrow \\ \mathbb{F} \end{array}$$

Since each extension is finite, we get that $\mathbb{F}(\alpha_0, \ldots, \alpha_{n-1})(\beta)$ is a finite extension over $\mathbb{F}$. So β must be algebraic over $\mathbb{F}$. □

Corollary 7.1.32. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be a field extension. If $\alpha, \beta \in \mathbb{E}$ are algebraic over $\mathbb{F}$, then $\alpha + \beta, \alpha - \beta, \alpha\beta, \alpha\beta^{-1}$ are also algebraic over $\mathbb{F}$.

PROOF. Since β is algebraic over $\mathbb{F}$, then β is algebraic over $\mathbb{F}(\alpha)$ and $\mathbb{F}(\alpha, \beta)$ is an algebraic extension of $\mathbb{F}(\alpha)$. Thus by previous lemma $\mathbb{F}(\alpha, \beta)$ is an algebraic extension over $\mathbb{F}$. Thus every element of $\mathbb{F}(\alpha, \beta)$ is algebraic over $\mathbb{F}$. □

Exercise 7.1.33. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be an algebraic extension. Suppose R is a ring such that $\mathbb{F} \subseteq R \subseteq \mathbb{E}$. Show that R must be a field.

7.2. Splitting Fields

Simple algebraic extensions $\mathbb{F} \hookrightarrow \mathbb{F}(\alpha)$ are constructed by adjoining a root α of an irreducible polynomial in $\mathbb{F}[x]$. We know that if $f(x) \in \mathbb{F}[x]$ is irreducible, then the class of x in the field $\mathbb{F}[x]/(f(x))$ is a root of f, but do we necessarily obtain other roots? The answer is no, but there is a procedure to add all roots of a polynomial, and the result is called the *splitting field* of this polynomial.

Definition 7.2.1. Let $\mathbb{F}$ be a field and let $f(x) \in \mathbb{F}[x]$. A *splitting field of* f is a field extension $\mathbb{F} \hookrightarrow \mathbb{E}$ such that:

- **(i)** the polynomial $f(x)$ splits in $\mathbb{E}[x]$, i.e. $f(x)$ factors completely into linear factors:
$$f(x) = u(x - \alpha_1) \cdots (x - \alpha_n),$$
where $n = \deg(f)$, $\alpha_i \in \mathbb{E}$ and $u \in \mathbb{E}$, $u \neq 0$;
- **(ii)** the field $\mathbb{E} = \mathbb{F}(\alpha_1, \ldots, \alpha_n)$ where $\alpha_1, \ldots, \alpha_n$ are the roots of f.

Theorem 7.2.2. A splitting field always exists.

We shall prove in next section that it is also unique up to isomorphism.

Proof. Proof by induction on $n = \deg(f)$. For $n = 1$, the polynomial alreadys splits in $\mathbb{F}$ and there is nothing to prove. More generally, choose an irreducible factor $g(x)$ of $f(x)$. Define α_1 as the class of x in $\mathbb{F}[x]/(g(x))$. Then over $\mathbb{F}(\alpha_1)$, we get $f(x) = (x - \alpha_1)h(x)$, where $\deg(h) = n - 1$. Repeat the process inductively to conclude. □

Example 7.2.3. Consider $x^n - 1 \in \mathbb{Q}[x]$. Roots are called the n-th roots of unity: $e^{\frac{2\pi k i}{n}}, k = 0, \ldots, n - 1$. We see $\zeta_n = e^{\frac{2\pi i}{n}}$ generates the other roots. The set of all n-th roots form thus a cyclic group. The splitting field of $x^n - 1$ is $\mathbb{Q}(\zeta_n)$, and it is called the *cyclotomic field of the n-th roots of unity.*

Lemma 7.2.4. Let $\mathbb{L}$ be a field. The following are equivalent.

- **(i)** All $f(x) \in \mathbb{L}[x]$ of degree ≥ 1 splits.
- **(ii)** All $f(x) \in \mathbb{L}[x]$ of degree ≥ 1 has a root in $\mathbb{L}$.
- **(iii)** If $f(x) \in \mathbb{L}[x]$ is irreducible then $\deg(f) = 1$.
- **(iv)** If $\mathbb{L} \hookrightarrow \mathbb{E}$ is an algebraic extension, then $\mathbb{L} = \mathbb{E}$.

Proof. **(i) ⇒ (ii))** Let $f(x) \in \mathbb{F}[x]$ be of degree ≥ 1. By assumption:
$$f(x) = u(x - \alpha_i) \ldots (x - \alpha_n)$$
where $\alpha_i \in \mathbb{L}$. Thus $f(\alpha_1) = 0$ and so $\alpha_1 \in \mathbb{L}$ is a root of f.

(ii) ⇒ (iii)) Suppose $f(x)$ where $\deg(f) \geq 1$ is irreducible. We know there exists a root $\alpha \in \mathbb{L}$ of f, so $f(x) = (x - \alpha)g(x)$ for some $g(x) \in \mathbb{L}[x]$. As f is irreducible, we must have that g is a unit, and so $\deg(g) = 0$. Hence $\deg(f) = 1$.

(iii) ⇒ (iv)) Let $\alpha \in \mathbb{E}$. Then $\mathbb{L}(\alpha) \cong \mathbb{L}[x]/(m_{\alpha,\mathbb{L}}(x))$. As $m_{\alpha,\mathbb{L}}$ is irreducible, we get $\deg(m_{\alpha,\mathbb{L}}) = 1$. So $[\mathbb{L}(\alpha) : \mathbb{L}] = 1$. Thus $\alpha \in \mathbb{L}$. So $\mathbb{E} = \mathbb{L}$.

(iv) ⇒ (i)) Let $f(x) \in \mathbb{L}[x]$ be of degree ≥ 1. We can decompose f as:
$$f(x) = ug_1(x) \ldots g_n(x),$$
where $u \in \mathbb{L}^\times$ and $g_i(x) \in \mathbb{L}[x]$ are irreducible. Then each $\mathbb{L}[x]/(g_i(x))$ is an algebraic extension of $\mathbb{L}$, and thus their degree of extension is 1. Thus $\deg(g_i) = 1$, hence f splits. □

Definition 7.2.5. A field $\mathbb{L}$ is *algebraically closed* if all polynomials of degree greater or equal to 1 split.

Theorem 7.2.6. Let $\mathbb{F}$ be a field. Then there exists a field extension $\mathbb{F} \hookrightarrow \mathbb{L}$ in which $\mathbb{L}$ is algebraically closed.

We shall see that the choice of $\mathbb{L}$ is unique up to isomorphism in the next section.

Proof. We first build $\mathbb{F} \hookrightarrow \mathbb{L}_1$ in which all $f(x) \in \mathbb{F}[x]$ have a root. Define:

$$\mathcal{P} = \{f(x) \in \mathbb{F}[x] \mid \deg(f) \geq 1\}$$

Define $R = \mathbb{F}[x_f \mid f \in \mathcal{P}]$ – the polynomial ring with a variable x_f for each $f \in \mathcal{P}$. Define I to be the ideal generated by all $f \in \mathcal{P}$.

Notice first that $I \neq R$. Indeed, by contradiction, suppose $I = R$, then $1 \in I$, so:

$$1 = \sum_{j=1}^{n} h_j f_j(x_j),$$

where $h_j \in R$ and $f_j \in \mathcal{P}$. Let $\mathbb{K}$ be a field extension of $\mathbb{F}$ in which f_j has a root α_j for each $j = 1, \ldots, n$. By universal property of polynomial rings, define a ring homomorphism $\varphi : R \to \mathbb{K}$ where $\varphi(x_{f_j}) = \alpha_j$, $\varphi(x_f) = 0$ if $f \notin \mathcal{P}$ and $\varphi(k) = k$ for all $k \in \mathbb{F} \subseteq \mathbb{K}$. We obtain:

$$1 = \varphi(1) = \sum_{j=1}^{n} \varphi(h_j)\varphi(f_j(x_j)) = \sum_{j=1}^{n} \varphi(h_j) \underbrace{f_j(\alpha_j)}_{=0} = 0$$

This is a contradiction. Thus $I \neq R$.

As $I \neq R$, there exists a maximal ideal $M \subseteq R$ containing I. Construct $\mathbb{L}_1 = R/M$, it is a field and we obtain $\mathbb{F} \subseteq \mathbb{L}_1$. Each $f(x) \in \mathbb{F}[x]$ has a root in $\mathbb{L}_1$ given by the class of x_f in $\mathbb{L}_1$.

Define similarly $\mathbb{L}_2$ as an extension of $\mathbb{L}_1$ in which all polynomials over $\mathbb{L}_1$ have a root, as we did above. Inductively, we obtain field extensions:

$$\mathbb{F} \hookrightarrow \mathbb{L}_1 \hookrightarrow \mathbb{L}_2 \hookrightarrow \ldots \hookrightarrow \mathbb{L}_i \hookrightarrow \ldots.$$

Define $\mathbb{L} = \bigcup_{i=1}^{\infty} \mathbb{L}_i$. It remains a field. If $g(x) \in \mathbb{L}[x]$ where $\deg(g) \geq 1$, then the coefficients of g belong to some $\mathbb{L}_j$ where j is big enough. Thus g has a root in $\mathbb{L}_j \subseteq \mathbb{L}$. Thus $\mathbb{L}$ is algebraically closed. □

Definition 7.2.7. Let $\mathbb{F} \hookrightarrow \mathbb{L}$ be an extension. We say $\mathbb{L}$ is an *algebraic closure* of $\mathbb{F}$ if $\mathbb{L}$ is algebraically closed and $\mathbb{F} \hookrightarrow \mathbb{L}$ is an algebraic extension.

Corollary 7.2.8. An algebraic closure always exists for any field.

Again, we prove in next section that the algebraic closure is unique up to isomorphism.

Proof. By previous theorem, given a field $\mathbb{F}$, there exists $\mathbb{F} \hookrightarrow \mathbb{L}$ in which $\mathbb{L}$ is algebraically closed. Define a subfield:

$$\overline{\mathbb{F}} = \{\alpha \in \mathbb{L} \mid \alpha \text{ is algebraic over } \mathbb{F}\} \subseteq \mathbb{L}.$$

Then by Corollary 7.1.32, we have that $\overline{\mathbb{F}}$ is indeed a subfield. Moreover, by construction $\mathbb{F} \hookrightarrow \overline{\mathbb{F}}$ is an algebraic extension. It remains to show that $\overline{\mathbb{F}}$ is algebraically

closed. Suppose $\beta \in \mathbb{L}$ is algebraic over $\overline{\mathbb{F}}$. Then as $\overline{\mathbb{F}}$ is algebraic over $\mathbb{F}$, we get that β must be algebraic over $\mathbb{F}$ by Lemma 7.1.31. Thus $\beta \in \overline{\mathbb{F}}$. So $\overline{\mathbb{F}}$ is algebraically closed. □

Definition 7.2.9. If $\mathbb{F}$ is a field, we denote by $\overline{\mathbb{F}}$ its algebraic closure (unique up to isomorphism, see next section).

Example 7.2.10. We shall see later that $\overline{\mathbb{R}} = \mathbb{C}$. This is referred to as the fundamental theorem of algebra.

Suppose $f(x) = a_0 + a_1 x + \cdots + a_n x^n \in \mathbb{F}[x]$ is some polynomial, and suppose $\alpha_1, \ldots, \alpha_n \in \overline{\mathbb{F}}$ are its roots (possibly not distinct), then we obtain:

$$f(x) = a_n(x - \alpha_1) \ldots (x - \alpha_n).$$

Can we recover the coefficients a_i from knowing the roots of f? For instance, the reader is most likely familiar with the fact that given a quadratic polynomial $f(x) = c + bx + x^2 \in \mathbb{R}[x]$, and given its roots $\alpha, \beta \in \mathbb{C}$, we have:

$$b = -(\alpha + \beta), \qquad c = \alpha\beta.$$

This generalizes nicely into the following formula.

Theorem 7.2.11 (Viète's formula)**.** Let $\mathbb{F}$ be a field of characteristic zero. Let $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ be a polynomial of degree $n \geq 1$ in $\mathbb{F}$. Let $\alpha_1, \ldots, \alpha_n \in \overline{\mathbb{F}}$ be its roots (possibly not distinct). Then we have, for all $k = 1, 2, \ldots, n$:

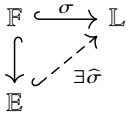
$$a_{n-k} = (-1)^{n-k} a_n \sum_{1 \leq i_1 \leq \cdots \leq i_k \leq n} \left(\prod_{j=1}^{k} \alpha_{i_j} \right).$$

In particular, if f is monic (i.e. $a_n = 1$), we obtain:

$$a_{n-1} = -(\alpha_1 + \cdots + \alpha_n), \quad \text{and,} \quad a_0 = (-1)^n \alpha_1 \ldots \alpha_n.$$

7.3. Lifting extensions

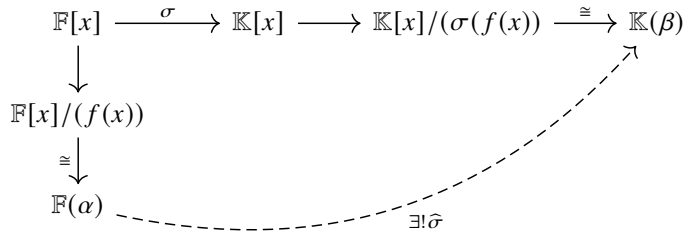
Definition 7.3.1. Let $\sigma : \mathbb{F} \hookrightarrow \mathbb{L}$ be a field extension. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be another field extension. A field extension $\widehat{\sigma} : \mathbb{E} \hookrightarrow \mathbb{L}$ is said to be a *lift of* σ *over* $\mathbb{E}$ if $\widehat{\sigma}(k) = \sigma(k)$ for all $k \in \mathbb{F} \subseteq \mathbb{E}$:



Lemma 7.3.2. Let $\sigma : \mathbb{F} \hookrightarrow \mathbb{L}$ be a field extension in which $\mathbb{L}$ is algebraically closed. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be a simple algebraic extension, i.e. $\exists \alpha \in \mathbb{E}$ such that $\mathbb{E} = \mathbb{F}(\alpha)$. Denote $f(x) = m_{\alpha,\mathbb{F}}(x) \in \mathbb{F}[x]$ the minimal polynomial. We obtain the following.

- **(i)** There exists a lift $\widehat{\sigma} : \mathbb{E} \hookrightarrow \mathbb{L}$ of σ over $\mathbb{E}$. Moreover, if $\beta \in \mathbb{L}$ is a root of $\sigma(f(x))$, then $\widehat{\sigma}(\alpha) = \beta$. Fixing β, the choice of $\widehat{\sigma}$ is unique.
- **(ii)** If $\widetilde{\sigma} : \mathbb{E} \to \mathbb{L}$ is any lift of σ over $\mathbb{E}$, then $\widetilde{\sigma}(\alpha)$ is a root of $\sigma(f(x))$.
- **(iii)** The number of possible lifts $\widehat{\sigma} : \mathbb{E} \hookrightarrow \mathbb{L}$ over $\mathbb{E}$ equals the number of roots of $f(x)$. Moreover, that number is less than or equal to $[\mathbb{E} : \mathbb{F}]$.

Proof. **(i)** Let $\beta \in \mathbb{L}$ be a root of $\sigma(f(x))$ which exists as $\mathbb{L}$ is algebraically closed. Fixing $\mathbb{K} = \mathrm{Im}(\sigma) \subseteq \mathbb{L}$ as a subfield of $\mathbb{L}$. We obtain by the universal property a unique ring homomorphism $\widehat{\sigma} : \mathbb{F}(\alpha) \to \mathbb{K}(\beta) \subseteq \mathbb{L}$ such that $\widehat{\sigma}(\alpha) = \beta$:



(ii) Let $f(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + x^n \in \mathbb{F}[x]$. We have $\sigma(f)(x) = \sigma(a_0) + \sigma(a_1)x + \cdots + \sigma(a_{n-1})x^{n-1} + x^n \in \mathbb{L}[x]$. Suppose $\widetilde{\sigma} : \mathbb{E} \to \mathbb{L}$ is a lift of σ. Then:

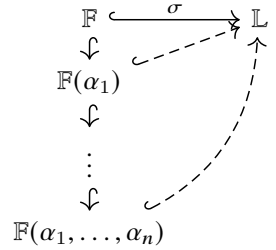
$$\begin{aligned}
\sigma(f)(\widetilde{\sigma}(\alpha)) &= \sigma(a_0) + \sigma(a_1)\widetilde{\sigma}(\alpha) + \cdots + \sigma(a_{n-1})\widetilde{\sigma}(\alpha)^{n-1} + \widetilde{\sigma}(\alpha)^n \\
&= \widetilde{\sigma}(a_0 + a_1\alpha + \cdots + a_{n-1}\alpha^{n-1} + \alpha^n) \\
&= \widetilde{\sigma}(f(\alpha)) \\
&= \widetilde{\sigma}(0) \\
&= 0.
\end{aligned}$$

(iii) We have seen that the number of lifts is precisely equal to the number of roots in $\mathbb{L}$ of $\sigma(f)(x) \in \mathbb{L}[x]$, which is equal to the number of roots of $f(x) \in \mathbb{F}[x]$, which there are at most $\deg(f) = [\mathbb{E} : \mathbb{F}]$. □

Theorem 7.3.3 (Lifting theorem)**.** Let $\sigma : \mathbb{F} \hookrightarrow \mathbb{L}$ be a field extension in which $\mathbb{L}$ is algebraically closed.

- **(i)** If $\mathbb{F} \hookrightarrow \mathbb{E}$ is a finite extension, then there exits a lift of σ over $\mathbb{E}$, and there are at most $[\mathbb{E} : \mathbb{F}]$ such lifts.
- **(ii)** If $\mathbb{F} \hookrightarrow \mathbb{E}$ is an algebraic extension of $\mathbb{F}$, then there exists a lift of σ over $\mathbb{E}$.

PROOF. **(i)** Since $\mathbb{E}$ is a finite extension, there exist $\alpha_1, \dots, \alpha_n \in \mathbb{E}$ algebraic over $\mathbb{E}$ such that $\mathbb{E} = \mathbb{F}(\alpha_1, \dots, \alpha_n)$. We can therefore apply inductively previous lemma:



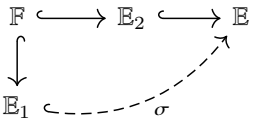
as each α_{i+1} is algebraic over $\mathbb{F}(\alpha_1, \dots, \alpha_i)$.

(ii) Let $\mathcal{P}$ be the set of pairs $(\mathbb{K}, \tau)$ where $\mathbb{K}$ is a field such that $\mathbb{F} \subseteq \mathbb{K} \subseteq \mathbb{E}$ and $\tau : \mathbb{K} \to \mathbb{L}$ is a lift of σ over $\mathbb{K}$. We say $(\mathbb{K}, \tau) \leq (\mathbb{K}', \tau')$ in $\mathcal{P}$ if $\mathbb{K} \subseteq \mathbb{K}'$ and $\tau'(k) = \tau(k)$ for all $k \in \mathbb{K}$ We obtain that $(\mathcal{P}, \leq)$ is a poset. Given a totally ordered subset $\mathcal{S}$ of $\mathcal{P}$, there exists an upper bound given $\bigcup_{\mathbb{K} \in \mathcal{P}} \mathbb{K}$. Thus by Zorn's lemma, there exists a maximal element of $\mathcal{P}$, call it $(\mathbb{G}, \eta)$. We argue that $\mathbb{G} = \mathbb{E}$. Indeed, if $\alpha \in \mathbb{E}$, then α is algebraic over $\mathbb{F}$ and thus over $\mathbb{G}$. Thus there exists a lift $\widehat{\eta} : \mathbb{G}(\alpha) \to \mathbb{L}$ of η by previous lemma. But then $(\mathbb{G}, \eta) \leq (\mathbb{G}(\alpha), \widehat{\eta})$. Therefore by maximality, we get $\mathbb{G}(\alpha) = \mathbb{G}$, in particular $\alpha \in \mathbb{G}$. Thus $\mathbb{E} = \mathbb{G}$ and we obtained the desired lift. □

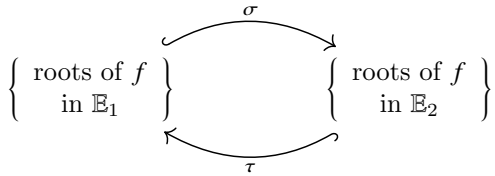
Corollary 7.3.4. Let $\mathbb{F}$ be a field.

- **(i)** Let $f(x) \in \mathbb{F}[x]$ and let $\mathbb{E}_1$ and $\mathbb{E}_2$ be splitting fields of f. Then there exists an isomorphism $\mathbb{E}_1 \xrightarrow{\cong} \mathbb{E}_2$ of $\mathbb{F}$-algebras.
- **(ii)** Let $\mathbb{L}_1$ and $\mathbb{L}_2$ be two algebraic closures of $\mathbb{F}$. Then there exists an isomorphism $\mathbb{L}_1 \xrightarrow{\cong} \mathbb{L}_2$ of $\mathbb{F}$-algebras.

PROOF. **(i)** Let $\mathbb{E}$ be an algebraic closure of $\mathbb{E}_2$, it also an algebraic closure of $\mathbb{F}$. By the lifting theorem, we obtain a lift:



If $\alpha \in \mathbb{E}_1$ is a root of f, then $\sigma(\alpha) \in \mathbb{E}$ is a root of f by the previous lemma. By definition of a splliting field, we must have $\sigma(\alpha) \in \mathbb{E}_2$. Thus we can corestrict $\sigma : \mathbb{E}_1 \to \mathbb{E}_2$. Switching the roles of $\mathbb{E}_1$ and $\mathbb{E}_2$, we obtain another lift $\tau : \mathbb{E}_2 \to \mathbb{E}_1$ that is also $\mathbb{F}$-linear. We obtain two injections:



Since the sets are finite, these injections must be bijections. Since the roots of f are the generators of $\mathbb{E}_1$ and $\mathbb{E}_2$ we can conclude.

(ii) By the lifting theorem, there is a lift $\sigma : \mathbb{L}_2 \hookrightarrow \mathbb{L}_1$ in which $\sigma(k) = k$ for all $k \in \mathbb{F}$. Notice that we must have $\sigma(\mathbb{L}_2) \cong \mathbb{L}_2$ since σ must send roots of

a polynomial over $\mathbb{F}$ to another root of that polynomial. So $\sigma(\mathbb{L}_2)$ is algebraically closed, and thus $\sigma(\mathbb{L}_2) = \mathbb{L}_1$, and thus σ is an isomorphism as desired. □

Definition 7.3.5. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be a field extension. We denote by $\mathrm{Aut}_{\mathbb{F}}(\mathbb{E})$ the set of all $\mathbb{F}$-algebra isomorphisms $\sigma\colon \mathbb{E} \to \mathbb{E}$.

Exercise 7.3.6. Verify that $(\mathrm{Aut}_{\mathbb{F}}(\mathbb{E}), \circ, \mathrm{id}_{\mathbb{E}})$ is a group.

Theorem 7.3.7. Suppose $\mathbb{F} \hookrightarrow \mathbb{E}$ is a field extension, let $\alpha \in \mathbb{E}$ be algebraic over $\mathbb{F}$. If $\sigma \in \mathrm{Aut}_{\mathbb{F}}(\mathbb{E})$, then $\sigma(\alpha)$ is a root $m_{\alpha,\mathbb{F}}$. In other words, $\mathrm{Aut}_{\mathbb{F}}(\mathbb{E})$ permutes the roots of the minimal polynomial. More generally, if $f(x) \in \mathbb{F}[x]$ has a root $\alpha \in \mathbb{E}$, then $\sigma(\alpha)$ is also a root of f.

Proof. This is a reformulation of Lemma 7.3.2. □

Corollary 7.3.8. Given a simple algebraic extension $\mathbb{F} \hookrightarrow \mathbb{E} = \mathbb{F}(\alpha)$, an element $\sigma \in \mathrm{Aut}_{\mathbb{F}}(\mathbb{E})$ is entirely determined by its value $\sigma(\alpha)$, which must be a root of $m_{\alpha,\mathbb{F}}$.

Example 7.3.9. Let $\mathbb{E} = \mathbb{Q}(\sqrt{2})$. Then the group $\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\sqrt{2}))$ permutes the roots of the minimal polynomial of $\sqrt{2}$ which is $x^2 - 2$. The only other root is $-\sqrt{2}$. Thus given $\sigma \in \mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\sqrt{2}))$, we $\sigma(\sqrt{2}) = \pm\sqrt{2}$. As $\sigma(k) = k$ for all $k \in \mathbb{Q}$, we must have $\sigma(a + b\sqrt{2}) = a \pm b\sqrt{2}$ for all $a + b\sqrt{2} \in \mathbb{Q}(\sqrt{2})$. Thus $\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\sqrt{2})) = \{id, \tau\} \cong C_2$, where $\tau(\sqrt{2}) = -\sqrt{2}$.

Example 7.3.10. Suppose $\mathbb{F} = \mathbb{Q}(\sqrt[3]{2})$. The minimal polynomial is $f(x) = x^3 - 2$. Notice that the other roots of $x^3 - 2$ are not in $\mathbb{F}$ (they are complex numbers). Thus there is only one choice for $\sigma \in \mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\sqrt[3]{2}))$ namely the identity. Thus $\mathrm{Aut}_{\mathbb{Q}}(\mathbb{Q}(\sqrt[3]{2}))$ is the trivial group.

The issue we faced is that we did not have enough roots in our field extension.

Corollary 7.3.11. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be a finite extension. Then $|\mathrm{Aut}_{\mathbb{F}}(\mathbb{E})| \leq [\mathbb{E} : \mathbb{F}]$.

Proof. This follows from Lemma 7.3.2. □

7.4. Galois extensions

7.4.1. Normal extensions.

Definition 7.4.1. A field extension $\mathbb{F} \hookrightarrow \mathbb{E}$ is *normal* if for any irreducible polynomial $p(x) \in \mathbb{F}[x]$, either $\mathbb{E}$ contains all roots of p, or none. We shall also say that $\mathbb{E}$ *is normal over* $\mathbb{F}$.

Theorem 7.4.2. A field extension $\mathbb{F} \hookrightarrow \mathbb{E}$ is finite and normal if and only if $\mathbb{E}$ is the splitting field of some polynomial in $\mathbb{F}[x]$.

Proof. We will start with the forward implication, which is easier. Let $\mathbb{E} = \mathbb{F}(\alpha_1, \ldots, \alpha_n)$. Let $p_i(x) = m_{\alpha_i,\mathbb{F}}(x) \in \mathbb{F}[x]$. As the extension is normal, all the roots of each p_i are in $\mathbb{E}$. Thus $\mathbb{E}$ is the splitting field of $f(x) = p_1(x) \ldots p_n(x)$.

For the converse, suppose $\mathbb{E}$ is the splitting field of a polynomial $f(x) \in \mathbb{F}[x]$. Then it must be a finite extension. Let $p(x)$ be an irreducible polynomial such that $\mathbb{E}$ contains a root of p, say $\alpha \in \mathbb{E}$. Choose an algebraic closure $\overline{\mathbb{E}}$; by since it is algebraic over $\mathbb{F}$ and algebraically closed, we observe $\overline{\mathbb{F}} = \overline{\mathbb{E}}$. Let $\beta \in \overline{\mathbb{F}}$ be another root of p. By the lifting theorem, there exists a unique $\mathbb{F}$-linear isomorphism $\sigma\colon \mathbb{F}(\alpha) \to \mathbb{F}(\beta)$ with $\sigma(\alpha) = \beta$. Now since $\mathbb{E}$ is the splitting field of $f(x) \in \mathbb{F}[x]$, it is also the splitting field of $f(x)$ regarded as a polynomial over $\mathbb{F}(\alpha)$. Similarly, $\mathbb{E}(\beta)$ is the splitting field of $\sigma(f)(x) \in \mathbb{F}(\beta)[x]$. By uniqueness of the splitting fields, we obtain an isomorphism $\widehat{\sigma}\colon \mathbb{E} \to \mathbb{E}(\beta)$ that is $\mathbb{F}$-linear. Therefore $[\mathbb{E}(\beta) : \mathbb{F}] = [\mathbb{E} : \mathbb{F}]$, and $[\mathbb{E}(\beta) : \mathbb{E}] = 1$, thus $\beta \in \mathbb{E}$ as desired. □

Corollary 7.4.3. Let $\mathbb{E} = \mathbb{F}(\alpha)$ be a simple algebraic extension of $\mathbb{F}$. Then $\mathbb{E}$ is a normal extension if and only if $m_{\alpha,\mathbb{F}}$ splits in $\mathbb{E}$.

Example 7.4.4. The field $\mathbb{Q}\left(\sqrt[3]{2}\right)$ is not a normal extension of $\mathbb{Q}$. Indeed, $\sqrt[3]{2}$ is a root of an irreducible polynomial $x^3 - 2 \in \mathbb{Q}[x]$. However, the other roots $\omega\sqrt[3]{2}$ and $\omega^2\sqrt[3]{2}$ are not in the extension, where $\omega = e^{\frac{2\pi i}{3}} = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$ is the third root of unity.

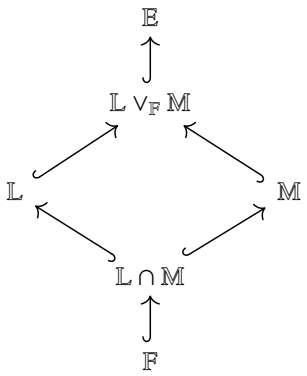
Example 7.4.5. The field $\mathbb{Q}(\sqrt[4]{2}, i)$ is a normal extension over $\mathbb{Q}$, as it is the splitting field of $x^4 + 2 \in \mathbb{Q}[x]$.

Example 7.4.6. Suppose $\mathbb{F} \hookrightarrow \mathbb{E}$ is a degree 2 extension. Then it must be normal. Indeed, consider $\alpha \in \mathbb{E} \setminus \mathbb{F}$. Then $m_{\alpha,\mathbb{F}}(x) = x^2 + bx + c$ for some $b, c \in \mathbb{F}$. From Viète's formula, we can guess another root $\beta = -\alpha - b \in \mathbb{F}(\alpha)$.

Proposition 7.4.7. Suppose we have field extensions $\mathbb{F} \hookrightarrow \mathbb{E} \hookrightarrow \mathbb{K}$. If $\mathbb{K}$ is normal over $\mathbb{F}$, then $\mathbb{K}$ is normal over $\mathbb{E}$.

7.4.2. Composites of subfields.

Definition 7.4.8. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be a field extension. Let $\mathbb{L}$ and $\mathbb{M}$ be field extensions of $\mathbb{F}$ contained in $\mathbb{E}$. The *composite* of the fields $\mathbb{L}$ and $\mathbb{M}$ in $\mathbb{E}$ is the smallest subfield of $\mathbb{E}$ containing both $\mathbb{L}$ and $\mathbb{M}$. We denote it $\mathbb{L} \vee_{\mathbb{F}} \mathbb{M}$.



Such a field always exists, as we can obtain it as the intersection

$$\mathbb{L} \vee_{\mathbb{F}} \mathbb{M} = \bigcap_{\mathbb{K} \in \mathcal{I}} \mathbb{K}$$

over the set $\mathcal{I}$ of all subfields of $\mathbb{E}$ that contain $\mathbb{L}$ and $\mathbb{M}$.

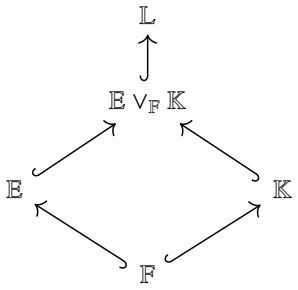
Proposition 7.4.9. Suppose $\mathbb{M} = \mathbb{F}(\beta_1, \ldots, \beta_m)$, in which β_i is algebraic over $\mathbb{F}$ for all $i = 1, \ldots, n$. Then $\mathbb{L} \vee_{\mathbb{F}} \mathbb{M} = \mathbb{L}(\beta_1, \ldots, \beta_m)$. If moreover we have $\mathbb{L} = \mathbb{F}(\alpha_1, \ldots, \alpha_n)$, then $\mathbb{L} \vee_{\mathbb{F}} \mathbb{M} = \mathbb{F}(\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_m)$.

Proposition 7.4.10. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ and $\mathbb{F} \hookrightarrow \mathbb{K}$ be two finite field extensions that are both contained in some larger field $\mathbb{L}$. Set $e = [\mathbb{E} : \mathbb{F}]$ and $k = [\mathbb{K} : \mathbb{F}]$, and set $c = [\mathbb{E} \vee_{\mathbb{F}} \mathbb{K}; \mathbb{F}]$. Then $e|c$, $k|c$, and $c \leq ek$.

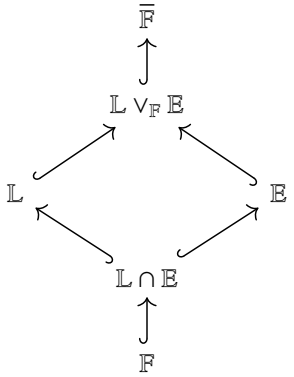
Proof. By multiplicativity of degree, we know that

$$c = e[\mathbb{E} \vee_{\mathbb{F}} \mathbb{K} : \mathbb{K}] = k[\mathbb{E} \vee_{\mathbb{F}} \mathbb{K} : \mathbb{E}].$$

We will show that $[\mathbb{E} \vee_{\mathbb{F}} \mathbb{K} : \mathbb{K}] \leq k$ by induction on the number n of elements that generate $\mathbb{K} = \mathbb{F}(\alpha_1, \ldots, \alpha_n)$ as a field over $\mathbb{F}$. □



Proposition 7.4.11. Suppose $\mathbb{L}$ and $\mathbb{E}$ are field extensions of a field $\mathbb{F}$, contained in $\overline{\mathbb{F}}$.



(i) If $\mathbb{E}$ is normal over $\mathbb{F}$, then their composite $\mathbb{L} \vee_{\mathbb{F}} \mathbb{E}$ is normal over $\mathbb{L}$.
(ii) If $\mathbb{E}$ and $\mathbb{L}$ are normal over $\mathbb{F}$, then their composite $\mathbb{L} \vee_{\mathbb{F}} \mathbb{E}$ and their intersection $\mathbb{L} \cap \mathbb{E}$ are normal over $\mathbb{F}$

7.4.3. Normal closure.

Definition 7.4.12. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be an algebraic extension contained in $\overline{\mathbb{F}}$. The *normal closure* of $\mathbb{E}$ in $\overline{\mathbb{F}}$ is the field

$$\mathbb{E}^{\mathrm{nc}} = \bigcap_{\mathbb{F} \in \mathcal{E}} \mathbb{F},$$

where $\mathcal{E}$ is the set of all normal field extensions of $\mathbb{E}$ contained in $\overline{\mathbb{F}}$.

Exercise 7.4.13. Show that $\mathbb{E}^{\mathrm{nc}}$ is the field spanned by all the images $\sigma(\mathbb{E})$ where $\sigma \colon \mathbb{E} \hookrightarrow \overline{\mathbb{F}}$ are $\mathbb{F}$-linear extensions.

Exercise 7.4.14. Show that $\mathbb{E}^{\mathrm{nc}}$ is the splitting field of the family of all minimal polynomials of elements in $\mathbb{E}$:

$$\{m_{\alpha,\mathbb{F}} \mid \alpha \in \mathbb{E}\}.$$

7.4.4. Separable extensions. Another pathological example of a field extension $\mathbb{F} \subset \mathbb{E}$ that we observed was where $\mathbb{F} = \mathbb{F}_p(t)$ and $\mathbb{E} = \mathbb{F}(\sqrt[p]{t})$. In this case, $\mathrm{Aut}_{\mathbb{F}}(\mathbb{E}) = \{e\}$ still has smaller order than $[\mathbb{E} : \mathbb{F}]$, even though the minimal polynomial of $\sqrt[p]{t}$ splits over $\mathbb{E}$:

$$x^p - t = \left(x - \sqrt[p]{t}\right)^p.$$

Example 7.4.15. Let $\mathbb{E} = \mathbb{F}_p(t)$ the field of fractions for the polynomial ring $\mathbb{F}_p[t]$, where we have let t be a variable. Consider $f(x) = x^p - t \in \mathbb{E}[x]$ – it is an irreducible polynomial. If α is a root of f, then:

$$0 = f(\alpha) = \alpha^p - t.$$

Thus $t = \alpha^p$. Thus $f(x) = x^p - t = x^p - \alpha^p = (x - \alpha)^p$. So α is the unique root of f. Thus $\mathbb{E}(\alpha)$ is a splitting field of $f(x) \in \mathbb{E}[x]$, and therefore is a normal extension.

Definition 7.4.16. Let $\mathbb{F}$ be a field. An irreducible polynomial $f(x) \in \mathbb{F}[x]$ is said to be *separable* if it has no multiple roots over its splitting field, i.e., all roots are distinct. If f is not separable, we say it is *inseparable*.

Definition 7.4.17. An algebraic extension $\mathbb{F} \hookrightarrow \mathbb{E}$ is *separable* if $m_{\alpha,\mathbb{F}}(x) \in \mathbb{F}[x]$ is separable, for all $\alpha \in \mathbb{E}$.

A useful way to gauge if a polynomial is separable or not is to use the derivative. Unlike in calculus, we can define it algebraically for any field $\mathbb{F}$:

$$\mathbb{F}[x] \longrightarrow \mathbb{F}[x],$$

$$f(x) = \sum_{i=0}^{n} a_i x^i \longmapsto f'(x) := \sum_{i=1}^{n} i a_i x^{i-1}.$$

It is a group homomorphism:

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

however, it is not a ring homomorphism. Rather, the *Leibniz rule* holds:

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x).$$

Lemma 7.4.18. Let $f(x) \in \mathbb{F}[x]$ and $\alpha \in \overline{\mathbb{F}}$ be a root of f. Then α is a multiple root of f if and only if α is a root of f'.

Proof. Suppose in $\overline{\mathbb{F}}$:

$$f(x) = (x-\alpha)^m g(x),$$

for $m \geq 1$ and $g(\alpha) \neq 0$. Then we obtain:

$$f'(x) = m(x-\alpha)^{m-1} g(x) + (x-\alpha)^m g'(x).$$

Thus α is a root of f' if and only if α is a root of $(x-\alpha)^{m-1}$, i.e. $m \geq 2$. □

Proposition 7.4.19. Let $\mathbb{F}$ be a field. Let $f(x) \in \mathbb{F}[x]$ be irreducible. Then f is separable over $\mathbb{F}$ if and only if $f' \neq 0$.

Proof. Let $\alpha \in \overline{\mathbb{F}}$ be a root of f. Assume without loss of generality that f is monic, so $f = m_{\alpha,\mathbb{F}}$. If α is a repeated root of f, then $f'(\alpha) = 0$ by last lemma. But my minimality of the degree of $m_{\alpha,\mathbb{F}}$, since $\deg(f') < \deg(m_{\alpha,\mathbb{F}})$, we must get $f' = 0$.

Conversely, if $f' = 0$, then α is a repeated root by previous lemma. □

Now in characteristic zero, if a polynomial is of degree ≥ 1, then its derivative cannot be zero. For instance if $f(x) = a + xb$, then $f'(x) = b$, but $b \neq 0$ as otherwise $\deg(f) = 0$. Thus we obtain this result.

Corollary 7.4.20. Suppose $\mathbb{F}$ is a field of characteristic zero. Then every irreducible polynomial over $\mathbb{F}$ is separable, and every algebraic extension of $\mathbb{F}$ is separable.

In characteristic p, what can happen is that if $f(x) = a + x^p$ for instance, then $f'(x) = px^{p-1} = 0$. Hence separability is less immediate in this case. In fact inseparable must be of the form $a_0 + a_1 x^p + a_2 x^{2p} + \cdots + a_n x^{np}$.

Corollary 7.4.21. Let $\mathbb{F}$ be a field of characteristic $p > 0$. Let $f(x) \in \mathbb{F}[x]$ be an inseparable irreducible polynomial. Then there exists an integer $d \geq 1$ and a irreducible separable polynomial $g(x) \in \mathbb{F}[x]$ such that $f(x) = g(x^{p^d})$.

Theorem 7.4.22. Let $\mathbb{F}$ be a finite field (and thus of characteristic $p > 0$). Then every irreducible polynomial over $\mathbb{F}$ is separable, and every algebraic extension of $\mathbb{F}$ is separable.

Proof. We know that if $f(x) \in \mathbb{F}[x]$ is such that $f' \neq 0$, then f is separable. So suppose $f' = 0$, we are going to show that f cannot be irreducible. Consider the so-called Frobenius homomorphism:

$$\begin{aligned} \mathbb{F} &\longrightarrow \mathbb{F} \\ a &\longmapsto a^p. \end{aligned}$$

It is an homomorphism with respect to addition (only true in characteristic p). It is injective as $a^p = 0 \Rightarrow a = 0$. As $\mathbb{F}$ is finite, it must also be a bijection. Therefore, for all $a \in \mathbb{F}$, $\exists b \in \mathbb{F}$ such that $b^p = a$. Threfore:

$$f(x) = a_0 + a_1 x^p + \cdots + a_{np} x^{np} = (b_0 + b_1 x + \cdots b_n x^n)^p,$$

where $a_i = b_i^p$. Thus f is reducible. □

7.4.5. Galois extensions.

Definition 7.4.23. An algebraic extension $\mathbb{F} \hookrightarrow \mathbb{E}$ is said to be a *Galois extension* if it is normal and separable. In this case, we write the group $\mathrm{Aut}_{\mathbb{F}}(\mathbb{E})$ as $\mathrm{Gal}(\mathbb{E}, \mathbb{F})$ and refer to it as the *Galois group* of the extension $\mathbb{F} \hookrightarrow \mathbb{E}$.

Definition 7.4.24. Given a separable polynomial $f(x) \in \mathbb{F}[x]$, its *Galois group*, denoted $\mathrm{Gal}(f)$, is the Galois group of the Galois extension $\mathbb{F} \hookrightarrow \mathbb{E}$, where $\mathbb{E}$ is the splitting field of f.

Theorem 7.4.25. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be a finite Galois extension. Then the order of $\mathrm{Gal}(\mathbb{E}, \mathbb{F})$ equals $[\mathbb{E} : \mathbb{F}]$.

Proof. This follows from Lemma 7.3.2. □

Example 7.4.26. Let $f(x) = x^3 - 2 \in \mathbb{Q}[x]$. Its splitting field is $\mathbb{E} = \mathbb{Q}(\sqrt[3]{2}, \zeta)$, where $\zeta = e^{\frac{2\pi i}{3}}$, the generating 3-rd root of unity. We needed to add ζ as the other roots of $x^3 - 2$ are $\zeta\sqrt[3]{2}$ and $\zeta^2\sqrt[3]{2}$. Let us show that $\mathrm{Gal}(f) \cong D_3$. First of all, we see that it must be a group of order 6 as:

$$[\mathbb{Q}(\sqrt[3]{2}, \zeta) : \mathbb{Q}] = [\mathbb{Q}(\sqrt[3]{2}, \zeta), \mathbb{Q}(\sqrt[3]{2})][\mathbb{Q}(\sqrt[3]{2}), \mathbb{Q}] = 2 \cdot 3.$$

If $\phi \in \mathrm{Gal}(f)$ then $\phi : \mathbb{E} \to \mathbb{E}$ is a $\mathbb{Q}$-automorphisms that permutes the roots of the minimal polynomial of $\sqrt[3]{2}$ (i.e. $x^3 - 2$) and the roots of the minimal polynomial of ζ (i.e. $1 + x + x^2$). In fact ϕ is entirely determined by where $\sqrt[3]{2}$ and ζ are sent to. Thus we can check that $\mathrm{Gal}(f)$ is generated by:

$$\begin{aligned} \sigma : \mathbb{Q}(\sqrt[3]{2}, \zeta) &\longrightarrow \mathbb{Q}(\sqrt[3]{2}, \zeta) \\ \sqrt[3]{2} &\longmapsto \zeta\sqrt[3]{2} \\ \zeta &\longmapsto \zeta \end{aligned}$$

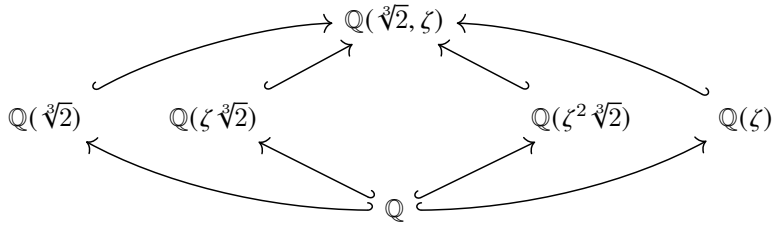
and:

$$\begin{aligned} \tau : \mathbb{Q}(\sqrt[3]{2}, \zeta) &\longrightarrow \mathbb{Q}(\sqrt[3]{2}, \zeta) \\ \sqrt[3]{2} &\longmapsto \sqrt[3]{2} \\ \zeta &\longmapsto \zeta^2 \end{aligned}$$

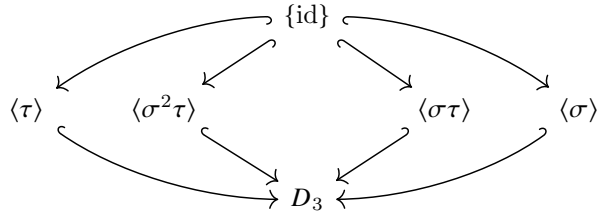
We also verify that we have the following relations:

$$\sigma^4 = \mathrm{id}, \quad \tau^2 = \mathrm{id}, \quad \tau\sigma\tau = \sigma^3.$$

The only possible group of order 6 with these relations is the dihedral group D_3.



All possible subfield extensions above seem to be related to the subgroups of D_3:



For instance, $\mathrm{Gal}\left(\mathbb{Q}(\sqrt[3]{2},\zeta),\mathbb{Q}(\zeta)\right) = \langle\sigma\rangle$. We shall explore this relationship in the following sections.

7.5. Primitive elements

We begin with by the following motivating example.

Example 7.5.1. Consider the algebraic extension:

$$\mathbb{Q} \hookrightarrow \mathbb{Q}(\sqrt{2}, \sqrt{3}).$$

The extension turns out to be a simple extension $\mathbb{Q}(\sqrt{2}+\sqrt{3}) = \mathbb{Q}(\sqrt{2}, \sqrt{3})$. Let us show this by hand. We have $\sqrt{2}+\sqrt{3} \in \mathbb{Q}(\sqrt{2}, \sqrt{3})$, and thus $\mathbb{Q}(\sqrt{2}+\sqrt{3}) \subseteq \mathbb{Q}(\sqrt{2}, \sqrt{3})$. So let us show $\mathbb{Q}(\sqrt{2}, \sqrt{3}) \subseteq \mathbb{Q}(\sqrt{2}+\sqrt{3})$, i.e. let us show that $\sqrt{2}, \sqrt{3} \in \mathbb{Q}(\sqrt{2}+\sqrt{3})$. We have that:

$$(\sqrt{2}+\sqrt{3})^2 = 5 + 2\sqrt{6},$$

and thus $\sqrt{6} \in \mathbb{Q}(\sqrt{2}+\sqrt{3})$. Now notice:

$$\sqrt{6}(\sqrt{2}+\sqrt{3}) = 2\sqrt{3} + 3\sqrt{2} \in \mathbb{Q}(\sqrt{2}+\sqrt{3}).$$

We can conclude since we have:

$$\sqrt{2} = \left(2\sqrt{3} + 3\sqrt{2}\right) - 2(\sqrt{2}+\sqrt{3}), \quad \sqrt{3} = -\left(2\sqrt{3} + 3\sqrt{2}\right) + 3(\sqrt{2}+\sqrt{3}).$$

This finishes the proof. Let us find its minimal polynomial of $\sqrt{2}+\sqrt{3}$:

$$\begin{aligned} x &= \sqrt{2}+\sqrt{3} \\ x^2 &= 5 + 2\sqrt{6} \\ x^4 - 10x^2 + 25 &= (x^2-5)^2 = 24, \end{aligned}$$

therefore its minimal polynomial is $x^4 - 10x^2 + 1 \in \mathbb{Q}[x]$.

We would like to generalize the phenomenon above and find a better method to arrive to its conclusion.

Definition 7.5.2. Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be a finite extension. We say $\gamma \in \mathbb{E}$ is a *primitive element* if $\mathbb{E} = \mathbb{F}(\gamma)$.

Theorem 7.5.3 (Primitive element theorem)**.** Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be a finite separable extension. Then there exists a primitive element.

In order to prove this theorem we first need the following lemma from linear algebra.

Lemma 7.5.4. Let $\mathbb{F}$ be an infinite field. Let V be a finite dimensional $\mathbb{F}$-vector space. Let $W_1, \ldots, W_n$ be subspaces of V, in which $W_i \neq V$ for all $i = 1, \ldots, n$. Then $\bigcup_{i=1}^{n} W_i \neq V$.

Proof. We prove by induction on n. It is vacuous for $n = 1$. Suppose $\bigcup_{i=1}^{n-1} W_i \neq V$. Let $w \in V$ but $w \notin \bigcup_{i=1}^{n-1} W_n$. If $w \notin W_n$, we are done, so let us assume $w \in W_n$. Choose $v \in V$ but $v \notin W_n$. Consider $v + \lambda w$ for $\lambda \in \mathbb{F}$. We have that $v + \lambda w \notin W_n$ as otherwise $v = v + \lambda w - \lambda w \in W_n$, and this is true for any $\lambda \in \mathbb{F}$. Suppose there were two distinct scalars $\lambda \neq \mu \in \mathbb{F}$ such that $v + \lambda w$ and $v + \mu w$ are in W_i for some $i = 1, \ldots, n-1$. Then $(\lambda - \mu)w \in W_i$, and so $w \in W_i$ which is impossible. Thus there is at most one scalar, say $\lambda_i \in \mathbb{F}$, such that $v + \lambda_i w \in W_i$. As $\mathbb{F}$ is infinite, we are guaranteed that there exists a scalar λ disitinct from each λ_i where $v + \lambda w \notin W_i$ for all $i = 1, \ldots, n-1$ and $v + \lambda w \notin W_n$. □

PROOF OF THEOREM 7.5.3. If $\mathbb{F}$ is a finite field, then as $\mathbb{E}$ is a finite dimensional $\mathbb{F}$-vector space, we obtain that that $\mathbb{E}$ must be a finite field as well. Thus $\mathbb{E}^\times$ is a cyclic group (see Corollary 4.8.18), and thus has a generator say γ. Then we obtain $\mathbb{E} = \mathbb{F}(\gamma)$.

Suppose thus $\mathbb{F}$ is an infinite field. Let $n = [\mathbb{E}, \mathbb{F}]$. By separability, there exist n $\mathbb{F}$-linear embeddings $\mathbb{E} \hookrightarrow \overline{\mathbb{F}}$, denoted $\sigma_1, \ldots, \sigma_n$. Consider $\sigma_i - \sigma_j : \mathbb{E} \to \overline{\mathbb{F}}$, for $i \neq j$. It is $\mathbb{F}$-linear but not necessarily a ring homomorphism. Define $H_{ij} = \ker(\sigma_i - \sigma_j)$. It is a $\mathbb{F}$-subspace of $\mathbb{E}$. As $\sigma_i \neq \sigma_j$, we get $H_{ij} \neq \mathbb{E}$. By previous lemma, there exists $\gamma \in \mathbb{E}$ but $\gamma \notin \bigcup_{i \neq j} H_{ij}$.

By construction, we have $\sigma_i(\gamma) \neq \sigma_j(\gamma)$, for all $i \neq j$. Thus $\sigma_1(\gamma), \ldots, \sigma_n(\gamma)$ are all distinct. Thus we obtain n-lifts $\sigma_i : \mathbb{F}(\gamma) \hookrightarrow \overline{\mathbb{F}}$. Thus by Lemma 7.3.2, we get $[\mathbb{F}(\gamma) : \mathbb{F}] \geq n$. Since $[\mathbb{F}(\gamma) : \mathbb{F}]$ divides n as $\mathbb{F}(\gamma) \subseteq \mathbb{E}$, we get that $[\mathbb{F}(\gamma) : \mathbb{F}] = n$. Therefore $\mathbb{F}(\gamma) = \mathbb{E}$. □

The above proof gives a strategy to find a primitive element γ for a separable extension $\mathbb{F} \hookrightarrow \mathbb{E}$: determine first the group $\mathrm{Aut}_{\mathbb{F}}(\mathbb{E})$ and then find an element $\gamma \in \mathbb{E}$ such that $\sigma(\gamma) \neq \tau(\gamma)$ for all $\sigma \neq \tau$ in $\mathrm{Aut}_{\mathbb{F}}(\mathbb{E})$.

⚠ **Warning 7.5.5.** The reader might be tempted to think the primitive element of an extension of the form $\mathbb{F}(\alpha, \beta)$ is simply $\alpha + \beta$. This is wrong. Consider $\mathbb{Q} \hookrightarrow \mathbb{Q}(\sqrt{2} + i, \sqrt{3} - i)$. Then as $\sqrt{2} + i + \sqrt{3} - i = \sqrt{2} + \sqrt{3}$, it is clear that $i \notin \mathbb{Q}(\sqrt{2} + \sqrt{3})$ and thus $\sqrt{2} + \sqrt{3}$ is not the primitive element. Indeed, we can see that $\mathbb{Q}(\sqrt{2} + i, \sqrt{3} - i) = \mathbb{Q}(\sqrt{2}, \sqrt{3}, i)$, and thus the Galois group of the extension is the product $C_2 \times C_2 \times C_2$ in which the three degree 2 generators are induced by $\sqrt{2} \mapsto -\sqrt{2}$, $\sqrt{3} \mapsto -\sqrt{3}$ and $i \mapsto -i$. Therefore the primitive element is $\gamma = \sqrt{2}+\sqrt{3}+i$.

7.5.1. ⋆ A generalization. There is a nice generalization of the theorem.

Theorem 7.5.6 (Steinitz)**.** Let $\mathbb{F} \hookrightarrow \mathbb{E}$ be a finite extension. There exists a primitive element in $\mathbb{E}$ if and only if there exist finitely many intermediate field extensions between $\mathbb{F}$ and $\mathbb{E}$.

PROOF. To be finished. □

Example 7.5.7. It might be surprising, but there are finite extensions which have infinitely many intermediate subfields. Let $\mathbb{F}$ a field of characteristic p. Let $\mathbb{L} = \mathbb{F}(x, y)$, the fraction field of the polynomial ring with two variables x, y. Let α be a root of $X^p - x \in \mathbb{L}[X]$ and β be a root of $X^p - y \in \mathbb{L}[X]$. Then $[\mathbb{L}(\alpha, \beta) : \mathbb{L}] = p^2$ but there are infinitely many intermediate subfields between $\mathbb{L}(\alpha, \beta)$ and $\mathbb{L}$. Thus there exist no primitive element for this extension.

7.6. Fixed field theorem

Proposition 7.6.1. Let $\mathbb{E}$ be a field and $H < \operatorname{Aut}\mathbb{E}$ be a finite subgroup of the group of all ring automorphisms of $\mathbb{E}$. Set $\mathbb{F} = \mathbb{E}^H$.

7.7. Finite fields and cyclotomic extensions

Definition 7.7.1. Let $\mathbb{F}$ be a field and let $n \geq 2$ be an integer such that the characteristic of $\mathbb{F}$ does not divide n. A root of $x^n - 1 \in \mathbb{F}[x]$ is called an *n-th root of unity*. We say that an n-th root of unity ζ is *primitive* if it is not the root of any polynomial $x^d - 1 \in \mathbb{F}[x]$ for $d|n$, $d \neq n$, i.e. $\zeta^n = 1$ but $\zeta^d \neq 1$.

Lemma 7.7.2. Let $\zeta \in \overline{\mathbb{F}}$ be a primitive n-th root of unity.

- **(i)** The set of all n-th roots of unity are $\{1, \zeta, \ldots, \zeta^{n-1}\}$.
- **(ii)** The root ζ^i is also a primitive n-th root of unity if and only if $\gcd(i, n) = 1$.

A primitive n-th root of unity ζ in $\mathbb{F}$ allows us to generate all radical roots. More precisely, suppose $f(x) = x^n - a \in \mathbb{F}[x]$, and suppose α is a root of f, typically denoted $\alpha = \sqrt[n]{a}$, if $\mathbb{F} = \mathbb{Q}$. Then the roots of f are α, $\zeta\alpha$, $\zeta^2\alpha, \ldots, \zeta^{n-1}\alpha$.

Definition 7.7.3. Let ζ be a primitive n-th root of unity of a field $\mathbb{F}$. The extension $\mathbb{F} \hookrightarrow \mathbb{F}(\zeta)$ is a called a *cyclotomic extension* of $\mathbb{F}$

Proposition 7.7.4. Let $\mathbb{F}$ be a field and let $n \geq 2$ be an integer such that the characteristic of $\mathbb{F}$ does not divide n. Let ζ be a primitive n-th root of unity of a field $\mathbb{F}$. The cyclotomic extension $\mathbb{F}(\zeta)$ is a splitting field of $x^n - 1 \in \mathbb{F}[x]$ and is thus a finite Galois extension for which its Galois group is isomorphic to a subgroup of $(\mathbb{Z}/n\mathbb{Z})^\times$.

Proof. By previous lemma, we know all the other roots ζ^i are in $\mathbb{F}(\zeta)$. Since the characteristic of $\mathbb{F}$ does not divide n, the roots are all distinct, and thus the extension is separable, and thus is Galois. If $\sigma \in \mathrm{Gal}(\mathbb{F}(\zeta), \mathbb{F})$, then $\sigma(\zeta)$ is a root of $m_{\zeta,\mathbb{F}}(x) = x^n - 1$. Thus $\sigma(\zeta) = \zeta^i$ for some i. We must have ζ^i is also primitive, as if $(\zeta^i)^d = 1$ for some $d|n$, then if we apply σ^{-1} we get $\zeta^d = 1$, which is not possible as ζ is primitive. Therefore we have that $\gcd(i, n) = 1$, and thus the class of i modulo n is invertible in $\mathbb{Z}/n\mathbb{Z}$. Thus, by Theorem 2.3.18, we can define:

$$\begin{aligned} \Psi : \mathrm{Gal}(\mathbb{F}(\zeta), \mathbb{F}) &\longrightarrow (\mathbb{Z}/n\mathbb{Z})^\times \\ \sigma &\longmapsto i, \end{aligned}$$

for which $\sigma(\zeta) = \zeta^{\Psi(\zeta)}$.

The map Ψ is a group homomorphism: if $\sigma, \tau \in \mathrm{Gal}(\mathbb{F}(\zeta), \mathbb{F})$ are such that $\sigma(\zeta) = \zeta^i$ and $\tau(\zeta) = \zeta^j$, then $\sigma(\tau(\zeta)) = \zeta^{ij}$, and so $\Psi(\sigma\tau) = ij = \Psi(\sigma)\Psi(\tau)$.

Moreover, Ψ is injective as if $\Psi(\sigma) = 1$, then $\sigma(\zeta) = \zeta$ and thus $\sigma = \mathrm{id}$. □

7.7.1. Cyclotomic extensions over $\mathbb{Z}/p\mathbb{Z}$.

Theorem 7.7.5. Let $\mathbb{F}$ be a finite field of characteristic $p > 0$.

- **(i)** The field $\mathbb{F}$ must be a cyclotomic Galois extension of $\mathbb{Z}/p\mathbb{Z}$.
- **(ii)** If q is the number of elements in $\mathbb{F}$, then $q = p^n$ for some $n \geq 1$.
- **(iii)** The field $\mathbb{F}$ is the splitting field of $x^q - x \in \mathbb{Z}/p\mathbb{Z}[x]$.

Conversely, for any $q = p^n$, $n \geq 1$, there exists a field with q-elements which is the splitting field of $x^q - x \in \mathbb{Z}/p\mathbb{Z}[x]$ and is unique in a fixed algebraic closure $\overline{\mathbb{Z}/p\mathbb{Z}}$.

Proof. We know $\mathbb{F}$ must be an extension of $\mathbb{Z}/p\mathbb{Z}$ from Lemma 7.1.6, and since $\mathbb{F}$ is finite, it is also a finite dimensional $\mathbb{Z}/p\mathbb{Z}$-vector space. Say the dimension is n, then $\mathbb{F}$ has p^n-elements. Let $q = p^n$. As $\mathbb{F}^\times$ is a cyclic group of order $q - 1$ (see Corollary 4.8.18), we have $\alpha^{q-1} = 1$ for all $\alpha \in \mathbb{F}^\times$. Hence $\mathbb{F}^\times$ contains all $(q-1)$-th

root of unity and thus if ζ is a primitive element of the extension (which exists by Theorem 7.5.3), it is a primitive $(q-1)$-root as well, and so $\mathbb{F}$ is a cyclotomic extension. Thus $\alpha^q = \alpha$, for all $\alpha \in \mathbb{F}$. Hence $\mathbb{F}$ is the splitting field of $f(x) = x^q - x \in \mathbb{Z}/p\mathbb{Z}[x]$. Notice that $f'(x) = -1 \neq 0$, thus f is separable, and so f has precisely q-distinct roots, which are precisely the elements of $\mathbb{F}$. Thus we obtain that $\mathbb{F}$ is a Galois extension.

Conversely, let $\mathbb{E}$ be a splitting field of $f(x) = x^q - x \in \mathbb{Z}/p\mathbb{Z}[x]$ for some $q = p^n$, $n \geq 1$. Let $\mathcal{R} \subseteq \mathbb{E}$ be the set of roots f. Notice that if $\alpha, \beta \in \mathcal{R}$, then $\alpha + \beta, \alpha\beta$ and α^{-1} remain in $\mathcal{R}$. Thus $\mathcal{R}$ is a subfield of $\mathbb{E}$. Since $\mathbb{E}$ is generated by the elements of $\mathcal{R}$, we obtain $\mathbb{E} = \mathcal{R}$, thus $\mathbb{E}$ must have q-elements, as $f' \neq 0$. We conclude by unicity of splitting fields. □

Definition 7.7.6. Let p be a prime, we denote by $\mathbb{F}_q$ the unique field up to isomorphism with $q = p^n$ elements, for $n \geq 1$.

Example 7.7.7. We have $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$.

Example 7.7.8. We have $\mathbb{F}_4 \cong \mathbb{F}_2[x]/(1 + x + x^2)$.

Example 7.7.9. There are no field of order 6.

We have that $\mathbb{F}_q^\times$ is precisely the field made of the $(q-1)$-th roots of unity. A primitive $(q-1)$-th root of unity is a primitive element of the extension $\mathbb{F}_p \hookrightarrow \mathbb{F}_q$.

Theorem 7.7.10. Let p be a prime. Let $q = p^r$ and $\ell = p^s$. Then $\mathbb{F}_\ell$ is an extension of $\mathbb{F}_q$ if and only if $s = nr$ for some $n \geq 1$. Moreover, $\mathbb{F}_q \hookrightarrow \mathbb{F}_\ell$ is a Galois extension in this case, with Galois group isomorphic to $\mathbb{Z}/n\mathbb{Z}$, induced by the Frobenius automorphism:

$$\begin{aligned} \mathbb{F}_\ell &\longrightarrow \mathbb{F}_\ell \\ \alpha &\longmapsto \alpha^q. \end{aligned}$$

7.7.2. Cyclotomic extensions over $\mathbb{Q}$.

Definition 7.7.11. Let $\mathbb{F}$ be a field and let $n \geq 2$ be an integer such that the characteristic of $\mathbb{F}$ does not divide n. Let ζ be a primitive n-th root of unity. The *n-cyclotomic polynomial* $\Phi_n(x) \in \mathbb{F}(\zeta)[x]$ is defined as:

$$\Phi_n(x) = \prod_{\substack{1 \leq i \leq n \\ \gcd(i,n)=1}} (x - \zeta^i).$$

We denote by $\varphi(n) = \deg(\Phi_n(x))$.

Lemma 7.7.12.

Lemma 7.7.13. We have:

$$\sum_{d|n} \varphi(d) = n.$$

Lemma 7.7.14. We have:

$$x^n - 1 = \prod_{d|n} \Phi_d(x)$$

PROOF. Let ζ be a primitive n-th root of unity. Then we have:

$$\begin{aligned} x^n - 1 &= \prod_{1 \le i \le n} (x - \zeta^i) \\ &= \prod_{d \mid n} \prod_{\substack{1 \le i \le n \\ \gcd(i,n) = d}} (x - \zeta^i) \\ &= \prod_{d \mid n} \Phi_{\frac{n}{d}}(x) \\ &= \prod_{d \mid n} \Phi_d(x). \end{aligned}$$

□

CHAPTER 8

Categories

Philosophy: understand a mathematical structure via its interactions with others.

8.1. Categories

8.1.1. A first definition. The definition of a category is versatile. It generalizes monoids, groups, graphs, and posets, to name a few. At the same time, the definition captures also the systematic approach in mathematics in which we introduce a mathematical structure (e.g. groups) and study a classification of all possible structures (e.g. determine all groups up to group isomorphisms).

Definition 8.1.1. A *category* $\mathcal{C}$ consists of the following data.

(i) A class $\mathrm{Ob}(\mathcal{C})$. Its elements are called *objects* of the category $\mathcal{C}$. We write $X \in \mathcal{C}$ instead of $X \in \mathrm{Ob}(\mathcal{C})$.

(ii) For each ordered pair (X, Y) of objects in $\mathcal{C}$, we have a set $\mathrm{Hom}_{\mathcal{C}}(X, Y)$. Its elements are called *morphisms*, *arrows* or *maps from X to Y*. The set $\mathrm{Hom}_{\mathcal{C}}(X, Y)$ is referred as the *hom-set* of X and Y in $\mathcal{C}$. An element $f \in \mathrm{Hom}_{\mathcal{C}}(X, Y)$ is written as $f : X \to Y$ or as $X \xrightarrow{f} Y$ and we say $f : X \to Y$ is a map in $\mathcal{C}$. Given a map $f : X \to Y$ in $\mathcal{C}$, then X is said to be the *domain* of f and Y the *codomain* of f. We denote $\mathrm{Mor}(\mathcal{C}) = \coprod_{X,Y \in \mathcal{C}} \mathrm{Hom}_{\mathcal{C}}(X, Y)$ the class of all morphisms in $\mathcal{C}$.

(iii) Given objects X, Y and Z in $\mathcal{C}$, we have a function on the hom-sets:

$$\begin{array}{rcl} \mathrm{Hom}_{\mathcal{C}}(Y, Z) \times \mathrm{Hom}_{\mathcal{C}}(X, Y) & \longrightarrow & \mathrm{Hom}_{\mathcal{C}}(X, Z) \\ (g, f) & \longmapsto & g \circ f \end{array}$$

The map $g \circ f : X \to Z$ is called the *composition* or *composite of f and g*.

The above data is subject to the following axioms.

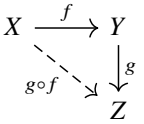
Associativity: Given $f : X \to Y$, $g : Y \to Z$, and $h : Z \to W$ maps in $\mathcal{C}$, then:

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

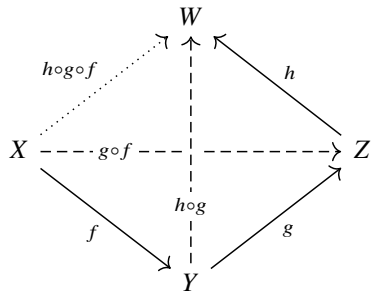
Unitality: For each object $X \in \mathcal{C}$, there exists a map $\mathrm{id}_X : X \to X$ in $\mathcal{C}$, called the *identity map on X*, such that:

- $\mathrm{id}_X \circ f = f$, for all maps $f : Y \to X$ in $\mathcal{C}$;
- $g \circ \mathrm{id}_X = g$, for all maps $g : X \to Y$ in $\mathcal{C}$.

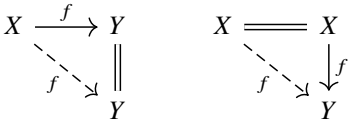
The composition allows us to uniquely fill in the diagram:



Associativity allows to unambiguously fill the diagram below with the dotted line:



The unitality allows us to extend any map $f : X \to Y$:



Remark 8.1.2. What we have defined above is what is usually called a *locally small* category. If we require $\mathrm{Hom}_C(X, Y)$ to only be a class instead of set, for each pair of objects X and Y, then we say we have a *large* category.

Exercise 8.1.3. Show that the identity morphism id_X is unique in every category. In other words, given an object X, if there exists a map $\alpha : X \to X$ such that $\alpha \circ f = f$ for all maps $f : Y \to X$ in C and $g \circ \alpha = g$ for all maps $g : X \to Y$ in C, then $\alpha = \mathrm{id}_X$.

Definition 8.1.4. Given a category C, a map $f : X \to Y$ in C is called an *isomorphism* if there exists a map $g : Y \to X$ in C such that: $f \circ g = \mathrm{id}_Y$ and $g \circ f = \mathrm{id}_X$. If such g exists, it is denoted f^{-1} and is called the *inverse* of f. We say two objects X and Y in C are *isomorphic* if there exists an isomorphism $f : X \to Y$. In this case, we write $X \cong Y$.

Exercise 8.1.5. Show that the inverse of an isomorphism is necessarily unique.

Exercise 8.1.6. Show that composition preserves isomorphisms: given isomorphisms $f : X \to Y$ and $g : Y \to Z$ in a category C, then $g \circ f : X \to Z$ is an isomorphism in C. Conclude that the relation $X \cong Y$ defines an equivalence relation on $\mathrm{Ob}(C)$.

Example 8.1.7. The category of sets, denoted Set, is defined as follows.

- **(i)** Its class of objects are all sets. Notice here the necessity of $\mathrm{Ob}(\mathsf{Set})$ to be a class: there is no set of all sets.
- **(ii)** Given sets X and Y, then $\mathrm{Hom}_{\mathsf{Set}}(X, Y)$ is the set of all functions $X \to Y$.
- **(iii)** Composition in Set is defined as follows. Given $f : X \to Y$ and $g : Y \to Z$, define $g \circ f : X \to Z$ by $(g \circ f)(x) = g(f(x))$, $\forall x \in X$.

For any set X, define $\mathrm{id}_X : X \to X$ by $\mathrm{id}_X(x) = x$, $\forall x \in X$. One can check that the composition is indeed associative and unital and thus Set is a category. An isomorphism in Set is precisely a bijection.

The next examples are categories in which objects are sets with extra structure. Composition, associativity and unitality are induced by the composition in Set.

Example 8.1.8. The category of groups, denoted Grp, is defined as follows.

- **(i)** Its class of objects are groups.
- **(ii)** Given groups G and H, define $\mathrm{Hom}_{\mathsf{Grp}}(G, H)$ to be the set of all group homomorphisms $G \to H$. The name hom-set originates from this category.
- **(iii)** Composition is defined as in Set.

Isomorphisms in Grp are precisely group isomorphisms.

Example 8.1.9. The category of Abelian groups, denoted Ab, is defined as follows.

- **(i)** Its class of objects are Abelian groups.
- **(ii)** Given Abelian groups A and B, define $\mathrm{Hom}_{\mathsf{Ab}}(A, B)$ to be the set of all group homomorphisms $A \to B$.

(iii) Composition is defined as in Set.

Isomorphisms in Ab are precisely group isomorphisms (between Abelian groups).

Example 8.1.10. The category of monoids, denoted Mon, is defined as follows.

(i) Its class of objects are monoids.
(ii) Given monoids M and N, define $\mathrm{Hom}_{\mathsf{Mon}}(M,N)$ to be the set of all monoid homomorphisms $M \to N$.
(iii) Composition is defined as in Set.

Example 8.1.11. The category of rings, denoted Ring, is defined as follows.

(i) Its class of objects are rings (with unity).
(ii) Given rings R and S, define $\mathrm{Hom}_{\mathsf{Ring}}(R,S)$ to be the set of all ring homomorphisms $R \to S$ (that preserves unities).
(iii) Composition is defined as in Set.

Isomorphisms in Ring are precisely ring isomorphisms.

Example 8.1.12. Let G be a group. Define the category of left G-sets ${}_G\mathsf{Set}$ as follows.

(i) Its class of objects are left G-sets.
(ii) Given left G-sets X and Y, define $\mathrm{Hom}_{{}_G\mathsf{Set}}(X,Y)$ to be the set of G-equivariant maps.
(iii) Composition is defined as in Set.

Define similarly the category of right G-sets Set_G.

Example 8.1.13. Let $\mathbb{F}$ be a field. The category $\mathsf{Vect}_{\mathbb{F}}$ of vector spaces over $\mathbb{F}$ is defined as follows.

(i) Its class of objects are vector spaces over $\mathbb{F}$.
(ii) Given vector spaces V and W over $\mathbb{F}$, define $\mathrm{Hom}_{\mathsf{Vect}_{\mathbb{F}}}(V,W)$ to be the set of linear transformation $V \to W$ over $\mathbb{F}$.
(iii) Composition is defined as in Set.

Isomorphisms in $\mathsf{Vect}_{\mathbb{F}}$ are precisely isomorphisms of vector spaces.

⚠ **Warning 8.1.14.** The names of the categories in the examples above can be misleading. They seem to suggest that a category is defined by its objects, but it really is not case. A category is defined by its morphisms. A better name for Set would be the "the category of set functions on all sets", and a better name for Grp would be "the category of group homomorphisms on all groups", for Ab would be "the category of group homomorphisms restricted on Abelian group", and so on. In practice, these names are too long, and it is natural to consider these mathematical objects with the appropriate morphisms.

The next examples shall emphasize how morphisms are the main actors in a category, and not objects.

Example 8.1.15. Given $(\mathbb{P},\leq)$ a poset, it defines a category, also denoted $\mathbb{P}$, as follows.

(i) $\mathrm{Ob}(\mathbb{P}) = \mathbb{P}$.
(ii) $\forall x,y \in \mathbb{P}$, $\mathrm{Hom}_{\mathbb{P}}(x,y) = \begin{cases} * & \text{if } x \leq y \\ \emptyset & \text{otherwise.} \end{cases}$ In other words, we write $x \to y$ if and only if $x \leq y$.

(iii) Composition is defined by transitivity: $\forall x, y, z \in \mathbb{P}$, if $x \leq y$ and $y \leq z$, then $x \leq z$.

We now need to check associativity and unitality.

- Associativity follows from the unambiguity of transivity: given $x \leq y$, $y \leq z$, $z \leq w$ in $\mathbb{P}$, then we can first deduce from $y \leq z \leq w$ that $y \leq w$ and thus from $x \leq y \leq w$ we obtain $x \leq w$; or we could have started from $x \leq y \leq z$ to deduce $x \leq z$, and thus from $x \leq z \leq w$ we obtain $x \leq w$.
- Unitality follows from the fact that $\forall x \in \mathbb{P}$ we have $x \leq x$. Thus we get that if $x \leq x \leq y$ then $x \leq y$. Similarly if $x \leq y \leq y$ then $x \leq y$.

Example 8.1.16. The empty set $\varnothing$ can be regarded as a poset. This defines a category $\mathbb{0}$ with no objects and no morphisms.

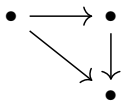
Example 8.1.17. Any singleton $\{\star\}$ is uniquely endowed with a poset structure. This defines a category $\mathbb{1}$ with one object and one morphism:

Since every object in a category associates an identity morphism, we often omit in the picture. Thus the category $\mathbb{1}$ is depicted as:

Example 8.1.18. Given the poset $\{1 \leq 2\}$, we obtain a category $\mathbb{2}$ with two objects and three morphisms (two of them are identity morphisms that we omit in the picture):

Example 8.1.19. Given the poset $\{1 \leq 2 \leq 3\}$, we obtain a category $\mathbb{3}$ with three objects and six morphisms depicted as (three of them are the identity morphisms and are omitted):

or:



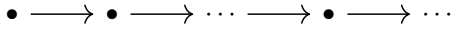
Since the third morphism depicted is the composition of the other two, we often also omit composition, and thus we depict $\mathbb{3}$ as:

Exercise 8.1.20. From the poset $\{1 \leq 2 \leq 3 \leq 4\}$, depict the category $\mathbb{4}$:

with all its compositions.

Remark 8.1.21. Recollecting from the examples above: when we picture a category, we omit the identity and the compositions.

Example 8.1.22. The natural numbers $(\mathbb{N}, \leq)$ form a poset and thus a category $\mathbb{N}$:



Example 8.1.23. The real numbers $(\mathbb{R}, \leq)$ form a poset and thus a category $\mathbb{R}$, but it is harder to depict than $\mathbb{N}$.

Definition 8.1.24. A category C is said to be *small* if $\mathrm{Ob}(C)$ is a set.

Example 8.1.25. If $\mathbb{P}$ is a poset, then its associated category is small. Thus $\mathbb{0}$, $\mathbb{1}$, $\mathbb{2}$, $\mathbb{3}$, $\mathbb{N}$ and $\mathbb{R}$ are small categories.

Example 8.1.26. The categories Set, Grp etc are *not* small.

Exercise 8.1.27. Show that any small categories can be regarded as a directed graph. Is every directed graph a category? What fails?

8.1.2. Monoids with many objects. The composition in a category must be associative and unital. These axioms are very similar to associativity and unitality of the binary operation of a monoid. Here we show precisely that a category is in fact a generalization of a monoid.

Example 8.1.28. Let $(M, *, e)$ be a monoid. Define a category BM as follows.

(i) The category BM has a unique object, labelled arbitrarily $\star$. In other words: $\mathrm{Ob}(BM) = \{\star\}$.

(ii) Given that there is only one object in the category, we only need to define one hom-set. Define $\mathrm{Hom}_{BM}(\star, \star) = M$.

(iii) We define composition in BM via the binary operation on the monoid M:

$$\begin{array}{rcl} M \times M = \mathrm{Hom}_{BM}(\star,\star) \times \mathrm{Hom}_{BM}(\star,\star) & \longrightarrow & \mathrm{Hom}_{BM}(\star,\star) = M \\ (x, y) & \longmapsto & x * y \end{array}$$

Since the monoid M is associative, then so is the composition on BM. The identity map $\mathrm{id}_\star : \star \to \star$ is equal to $e \in M$, the neutral element of M. Since a monoid is unital, then we obtain the unitality axiom on the category BM. Isomorphisms in BM are precisely units of M.

Remark 8.1.29. It is important to not confuse the morphisms in BM with actual set functions. Indeed, in BM, we replace entirely formally an element $x \in M$ by a map $x : \star \to \star$. But it is important to keep in mind that $\star$ is not a set and $x : \star \to \star$ is not a function of sets, it is just a different notation to express $x \in M$.

Exercise 8.1.30. Show that if C is a category with a unique object denoted $\star$, then $(\mathrm{Hom}_C(\star,\star), \circ, \mathrm{id}_\star)$ is a monoid. Conclude there is a correspondence between monoids and categories with one object.

Therefore a category with many objects can be regarded as a "monoid with many objects". Perhaps the next exercise can be enlightening in that regard.

Exercise 8.1.31. Let $\{M_1, \ldots, M_n\}$ be a collection of monoids. Define a category C as follows.

(i) $\mathrm{Ob}(C) = \{1, \ldots, n\}$.

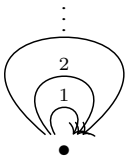
(ii) $\mathrm{Hom}_C(i, j) = \begin{cases} M_i & \text{if } i = j \\ \varnothing & \text{otherwise.} \end{cases}$

(iii) Composition is induced by the binary operations on each M_i.

Verify that C is a category. Can you generalize this example to any collection $\{M_i\}_{i \in I}$ of monoids, where I is any set?

Example 8.1.32. Let $\{1\}$ be the trivial monoid. Then its associated category $B\{1\}$ is the small category $\mathbb{1}$.

Example 8.1.33. We can view $\mathbb{N} = (\mathbb{N}, +, 0)$ as a monoid. This defines a category $B\mathbb{N}$ depicted as:



Given two morphisms $\bullet \xrightarrow{n} \bullet$ and $\bullet \xrightarrow{m} \bullet$, then their composition is the morphism $\bullet \xrightarrow{n+m} \bullet$. Notice the difference with the category of Example 8.1.22 when $\mathbb{N}$ was regarded as a poset.

Exercise 8.1.34. Given a directed graph, show how to define a category in which you freely add all possible compositions of its edges and identities on vertices? What is the category obtained from the directed graph with a single loop?

8.1.3. Groupoids. Since a group G is a monoid in which every element is a unit, we see that the associated category with one object BG, as in Example 8.1.28, is a category with one object in which every morphism is an isomorphism.

Definition 8.1.35. A *groupoid* is a category in which every morphism is an isomorphism.

Example 8.1.36. A groupoid with one object defines precisely a group. Conversely, given a group G, it defines uniquely a groupoid BG.

Example 8.1.37. Define $\mathbb{I}$ to be the category with two objects and exactly one morphism from one object to another. It can be depicted as (we omitted the identity morphisms):

Since the composition of the maps depicted above must be a map with a domain equalling its codomain, then it must be the identity by unicity. Hence $\mathbb{I}$ is a groupoid.

Exercise 8.1.38. Let R be a ring. Define $\mathsf{GL}(R)$ to be the following category.

(i) $\mathrm{Ob}(\mathsf{GL}(R)) = \{1, 2, 3, \ldots\}$.

(ii) $\mathrm{Hom}_{\mathsf{GL}(R)}(n, m) = \begin{cases} \mathrm{GL}_n(R) & \text{if } n = m \\ \emptyset & \text{otherwise.} \end{cases}$

(iii) Composition induced by product of matrices.

Verify that $\mathsf{GL}(R)$ is a groupoid.

Definition 8.1.39. A category C is *discrete* if the only morphisms are the identities. In other words, for all objects X and Y in C, we have:

$$\mathrm{Hom}_C(X, Y) = \begin{cases} \{\mathrm{id}_X\} & \text{if } X = Y \\ \emptyset & \text{otherwise.} \end{cases}$$

Example 8.1.40. Any discrete category is a groupoid.

Example 8.1.41. Given a set $\mathcal{S}$ (or more generally a class), define $\mathcal{S}_{\text{disc}}$ to be the discrete category with objects $\mathcal{S}$. Given a category $\mathcal{C}$, denote $\mathcal{C}_{\text{disc}}$ the discrete category $\text{Ob}(\mathcal{C})_{\text{disc}}$.

Definition 8.1.42. Given a category $\mathcal{C}$, its *maximal groupoid* $\mathcal{C}^{\cong}$, is the category defined as follows.

(i) Define $\text{Ob}(\mathcal{C}^{\cong}) = \text{Ob}(\mathcal{C})$.
(ii) Given objects X and Y, the set $\text{Hom}_{\mathcal{C}^{\cong}}(X,Y) \subseteq \text{Hom}_{\mathcal{C}}(X,Y)$ is defined to be the set of all isomorphisms from X to Y.
(iii) Since the composition of isomorphisms is an isomorphism, the composition $\text{Hom}_{\mathcal{C}}(Y,Z) \times \text{Hom}_{\mathcal{C}}(X,Y) \to \text{Hom}_{\mathcal{C}}(X,Z)$ restricts to a composition $\text{Hom}_{\mathcal{C}^{\cong}}(Y,Z) \times \text{Hom}_{\mathcal{C}^{\cong}}(X,Y) \to \text{Hom}_{\mathcal{C}^{\cong}}(X,Z)$.

We shall see in Exercise 8.1.53 that $\mathcal{C}^{\cong}$ is the largest groupoid contained in $\mathcal{C}$, hence the name "maximal".

Exercise 8.1.43. Let M be a monoid. Recall we can regard it as a category BM with one object. Recall also that we denote by $M^{\times}$ its units. Show that $(BM)^{\cong} = B(M^{\times})$

Exercise 8.1.44. Let R be a ring. Define $\mathsf{Mat}(R)$ to be the following category.

- $\text{Ob}(\mathsf{Mat}(R)) = \mathbb{N}$.
- $\text{Hom}_{\mathsf{Mat}(R)}(n,m) = \mathcal{M}_{m\times n}(R)$, the set of matrices with m-rows and n-columns.
- Composition is induced by matrix multiplication: given matrices $A : n \to m$ and $B : m \to k$ then $B \circ A := BA \in \mathcal{M}_{k\times n}(R)$.

Verify that $\mathsf{Mat}(R)$ is a category and its maximal groupoid is $\mathsf{GL}(R)$.

Exercise 8.1.45. What is the maximal groupoid of the categories obtained in Exercise 8.1.31?

Exercise 8.1.46. Explain why $\mathcal{C}^{\cong}$ can never equal $\mathbb{0}$ (unless $\mathcal{C} = \mathbb{0}$).

8.1.4. Subcategories. We have seen that from a category $\mathcal{C}$, we could restrict either its class of objects, or its sets of morphisms, or both. If one regards a category as a monoid with many objects, since there are submonoids, this leads to the notion of a subcategory.

Definition 8.1.47. Given a category $\mathcal{C}$, a *subcategory* $\mathcal{D}$ consists of:

(i) a subcollection $\text{Ob}(\mathcal{D}) \subseteq \text{Ob}(\mathcal{C})$
(ii) for each X and Y in $\mathcal{D}$, a subset $\text{Hom}_{\mathcal{D}}(X,Y) \subseteq \text{Hom}_{\mathcal{C}}(X,Y)$

such that $\mathcal{D}$ becomes itself a category with the composition induced from $\mathcal{C}$, i.e.:

- if $X \in \mathcal{D}$, then id_X is a morphism $\mathcal{D}$;
- if $f : X \to Y$ is a morphism in $\mathcal{D}$, then X and Y are objects in $\mathcal{D}$;
- if $f : X \to Y$ and $g : Y \to Z$ are morphisms in $\mathcal{D}$, then $g \circ f : X \to Z$ is a morphism in $\mathcal{D}$.

A subcategory $\mathcal{D}$ of $\mathcal{C}$ is *full* if $\text{Hom}_{\mathcal{D}}(X,Y) = \text{Hom}_{\mathcal{C}}(X,Y)$ for all $X, Y \in \mathcal{D}$.

Example 8.1.48. The category Grp is a subcategory of Set but it is not full: not ever set map between groups is a homomorphism. Similarly, we get that Ring is a non-full subcategory of Ab, Mon is a non-full subcategory of Set etc.

Example 8.1.49. The category Ab of Abelian groups is a full subcategory of Grp.

Example 8.1.50. We can define a category $\mathsf{Set}_{\mathrm{inj}}$ with same objects as in Set, but the morphisms are only injective functions of sets. Composition is the same as the composite of injective functions is an injective function and the identity function is always injective. An isomorphism is precisely an injective function that is surjective, i.e. a bijection. In many aspects, the category $\mathsf{Set}_{\mathrm{inj}}$ is similar to Set, but we shall see that these categories are not the same. The category $\mathsf{Set}_{\mathrm{inj}}$ is a subcategory that is not full.

Exercise 8.1.51. Let M be a monoid. Let $N \subseteq M$ be a subset. Show that N is a submonoid if and only if BN is a subcategory of BM. Conclude that a subset H of a group G is a subgroup if and only if BH is a groupoid and a subcategory of the groupoid BG.

Example 8.1.52. Given a category $\mathcal{C}$, then $\mathcal{C}_{\mathrm{disc}}$ and $\mathcal{C}^{\cong}$ are subcategories of $\mathcal{C}$ that are not full.

Exercise 8.1.53. Given a category $\mathcal{C}$, show that if $\mathcal{D}$ is a subcategory $\mathcal{C}$ and a groupoid, then $\mathcal{D}$ is a subcategory of the maximal groupoid $\mathcal{C}^{\cong}$.

Exercise 8.1.54. Show that given a category $\mathcal{C}$, a subclass of objects in $\mathcal{C}$ defines uniquely a full subcategory of $\mathcal{C}$. Apply this to define the category CMon of commutative monoids as a full subcategory of Mon.

8.2. Construction on categories

Given a category, we can construct new categories.

8.2.1. Product of categories.

Definition 8.2.1. Let $\mathcal{C}$ and $\mathcal{D}$ be categories. Define their *Cartesian product* $\mathcal{C} \times \mathcal{D}$ to be the following category.

(i) Its class of objects is $\mathrm{Ob}(\mathcal{C} \times \mathcal{D}) = \mathrm{Ob}(\mathcal{C}) \times \mathrm{Ob}(\mathcal{D})$.

(ii) Given objects $C, C' \in \mathcal{C}$ and $D, D' \in \mathcal{D}$, define

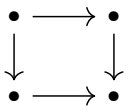
$$\mathrm{Hom}_{\mathcal{C}\times\mathcal{D}}\Big((C,D),(C',D')\Big) = \mathrm{Hom}_{\mathcal{C}}(C,C') \times \mathrm{Hom}_{\mathcal{D}}(D,D')$$

In other terms, a morphism $(C,D) \to (C',D')$ is denoted (f,g) and consists of a morphism $f : C \to C'$ in $\mathcal{C}$ and a morphism $g : D \to D'$ in $\mathcal{D}$.

(iii) Composition is induced by the compositions in $\mathcal{C}$ and $\mathcal{D}$. Given morphisms $(C,D) \overset{(f,g)}{\to} (C',D')$ and $(C',D') \overset{(f',g')}{\to} (C'',D'')$ in $\mathcal{C} \times \mathcal{D}$, define $(g',f') \circ (g,f)$ to be $(g' \circ g, f' \circ f) : (C,D) \to (C'',D'')$.

Exercise 8.2.2. Verify that $\mathcal{C} \times \mathcal{D}$ is indeed a category.

Example 8.2.3. The product $\mathbb{2} \times \mathbb{2}$ can be depicted as:



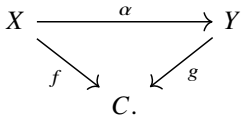
Exercise 8.2.4. Let G and H be monoids or groups. Show that $B(G \times H)$ can be regarded as $BG \times BH$.

8.2.2. Slice categories.

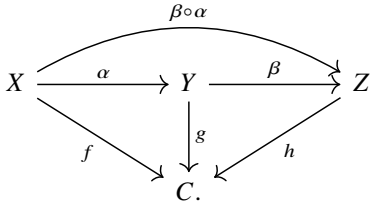
Definition 8.2.5. Let $\mathcal{C}$ be a category. Fix C an object in $\mathcal{C}$. The *slice category of $\mathcal{C}$ over C*, denoted $\mathcal{C}_{/C}$, is defined as follows.

(i) Its class of objects $\mathrm{Ob}(\mathcal{C}_{/C})$ consists of pairs (X,f) in which $C \in \mathcal{C}$ and $f : X \to C$ is a morphism in $\mathcal{C}$.

(ii) Given objects (X,f) and (Y,g) in $\mathcal{C}_{/C}$, a morphism $\alpha : (X,f) \to (Y,g)$ consists of a morphism $\alpha : X \to Y$ in $\mathcal{C}$ such that $g \circ \alpha = f$, i.e. the following diagram commutes in $\mathcal{C}$:



(iii) Composition in $\mathcal{C}_{/C}$ is determined by composition in $\mathcal{C}$. Given morphisms $\alpha : (X,f) \to (Y,g)$ and $\beta : (Y,g) \to (Z,h)$ in $\mathcal{C}_{/C}$, then the composition $\beta \circ \alpha : X \to Z$ remains over C:

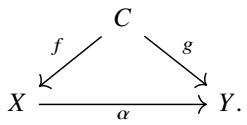


We can construct a similar definition and be under an object C instead of over.

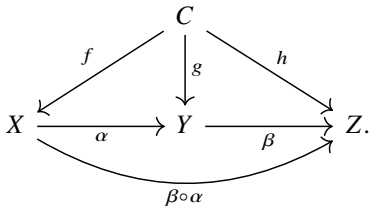
Definition 8.2.6. Let $\mathcal{C}$ be a category. Fix C an object in $\mathcal{C}$. The *slice category of $\mathcal{C}$ under C*, denoted $\mathcal{C}^{\backslash C}$, is defined as follows.

(i) Its class of objects $\mathrm{Ob}(\mathcal{C}^{\backslash C})$ consists of pairs (X, f) in which $C \in \mathcal{C}$ and $f : C \to X$ is a morphism in $\mathcal{C}$.

(ii) Given objects (X, f) and (Y, g) in $\mathcal{C}^{\backslash C}$, a morphism $\alpha : (X, f) \to (Y, g)$ consists of a morphism $\alpha : X \to Y$ in $\mathcal{C}$ such that $\alpha \circ f = g$, i.e. the following diagram commutes in $\mathcal{C}$:



(iii) Composition in $\mathcal{C}^{\backslash C}$ is determined by composition in $\mathcal{C}$. Given morphisms $\alpha : (X, f) \to (Y, g)$ and $\beta : (Y, g) \to (Z, h)$ in $\mathcal{C}^{\backslash C}$, then the composition $\beta \circ \alpha : X \to Z$ remains under C:



Example 8.2.7. Consider the category Set. Denote $\{*\}$ a singleton. We often write $\mathsf{Set}^{\backslash\{*\}}$ by Set_* and refer to it as the *category of pointed sets*. A function $\{*\} \to X$ is picking up an element $x_0 \in X$. Therefore an object in Set_* consists of a pair (X, x_0) in which X is a set and x_0 is a fixed element in X. A morphism $(X, x_0) \to (Y, y_0)$ in Set_* is determined by a function $f : X \to Y$ such that $f(x_0) = y_0$.

Exercise 8.2.8. Did the choice of the singleton $\{*\}$ mattered in the example above?

Example 8.2.9. Given any set X, there is a unique function $X \to \{*\}$. Therefore we see that $\mathsf{Set}_{/\{*\}}$ is similar to Set.

8.2.3. skeleton.

Example 8.2.10. Given any category $\mathcal{C}$, we see that isomorphisms define an equivalence relation on objects. We can therefore define a category $\mathcal{C}/\cong$ for which objects are $\mathrm{Ob}(\mathcal{C})/\cong$ and $\mathrm{Hom}_{\mathcal{C}/\cong}(X, Y) = \mathrm{Hom}_{\mathcal{C}}(X, Y)/\cong$. We see that $\mathcal{C}/\cong$ is not isomorphic to $\mathcal{C}$ but for a category theorist they should be.

8.2.4. Duality principle.

Definition 8.2.11. Given a category $\mathcal{C}$, its *opposite category*, denoted $\mathcal{C}^{\mathsf{op}}$, is a category with same object as in $\mathcal{C}$ in which we abstractly reverse the directions of the arrows. Formally:

(i) Its class of objects is $\mathrm{Ob}(\mathcal{C}^{\mathsf{op}}) = \mathrm{Ob}(\mathcal{C})$.

(ii) Given objects X and Y, define $\mathrm{Hom}_{\mathcal{C}^{\mathsf{op}}}(X, Y) := \mathrm{Hom}_{\mathcal{C}}(Y, X)$. We rewrite a map $f : Y \to X$ in $\mathcal{C}$ by $f^{\mathsf{op}} : X \to Y$ in $\mathcal{C}^{\mathsf{op}}$.

(iii) Given objects X, Y and Z, define:

$$\begin{array}{rcl} \mathrm{Hom}_{C^{\mathsf{op}}}(Y,Z) \times \mathrm{Hom}_{C^{\mathsf{op}}}(X,Y) & \longrightarrow & \mathrm{Hom}_{C^{\mathsf{op}}}(X,Z) \\ (g^{\mathsf{op}}, f^{\mathsf{op}}) & \longmapsto & g^{\mathsf{op}} \circ_{\mathsf{op}} f^{\mathsf{op}} := (f \circ g)^{\mathsf{op}}. \end{array}$$

Exercise 8.2.12. Check that C^{op} defines indeed a category.

Exercise 8.2.13. Prove that $(C^{\mathsf{op}})^{\mathsf{op}} = C$.

Exercise 8.2.14. Recall that if M is a monoid, we can define M^{op} is opposite monoid. Show that $B(M^{\mathsf{op}}) = (BM)^{\mathsf{op}}$.

Example 8.2.15. If $\mathbb{P}$ is a poset, we can obtain $\mathbb{P}^{\mathsf{op}}$ by reversing the order. For instance, as in Example 8.1.22, if we view the poset $(\mathbb{N}, \leq)$ as a category $\mathbb{N}$:

$$\bullet \longrightarrow \bullet \longrightarrow \cdots \longrightarrow \bullet \longrightarrow \cdots,$$

then its opposite category $\mathbb{N}^{\mathsf{op}}$ can be depicted as:

$$\bullet \longleftarrow \bullet \longleftarrow \cdots \longleftarrow \bullet \longleftarrow \cdots.$$

Essentially the above category is the one associated to the poset $(\mathbb{N}, \geq)$.

Exercise 8.2.16. Let C be a category and let C be a fixed object in C. Show that $\left(C_{/C}\right)^{\mathsf{op}} = C^{\backslash C}$.

8.3. Functors

Since categories are mathematical structures, there must be a notion of morphisms between them. If we regard categories as monoids with many objects, these morphisms should be thought as a generalization of homomorphisms of monoids. These morphisms go from one category to another. Herein lies the core motivation of category theory: we can travel between the world of groups, or rings, or topological spaces etc, and observe the interactions between these mathematical worlds. A morphism between categories is called a functor.

Definition 8.3.1. Let $\mathcal{C}$ and $\mathcal{D}$ be categories. A *functor F from $\mathcal{C}$ to $\mathcal{D}$*, denoted $F : \mathcal{C} \to \mathcal{D}$, is the following data.

(i) A function on the objects, also denoted F:

$$\begin{array}{rcl} \mathrm{Ob}(\mathcal{C}) & \longrightarrow & \mathrm{Ob}(\mathcal{D}) \\ X & \longmapsto & F(X). \end{array}$$

(ii) For each pair of objects $X, Y \in \mathcal{C}$, a function on the hom-sets, also denoted F:

$$\begin{array}{rcl} \mathrm{Hom}_{\mathcal{C}}(X,Y) & \longrightarrow & \mathrm{Hom}_{\mathcal{D}}(F(X),F(Y)) \\ \left(X \overset{f}{\to} Y\right) & \longmapsto & \left(F(X) \overset{F(f)}{\to} F(Y)\right) \end{array}$$

This data must follow the following two axioms.

Composition preserving: Denote $\circ_{\mathcal{C}}$ and $\circ_{\mathcal{D}}$ the compositions of morphisms in $\mathcal{C}$ and $\mathcal{D}$ respectively. Given $f : X \to Y$ and $g : Y \to Z$ in $\mathcal{C}$, we have the following equality of morphisms in $\mathcal{D}$:

$$F(g \circ_{\mathcal{C}} f) = F(g) \circ_{\mathcal{D}} F(f).$$

Identities preserving: For any object $X \in \mathcal{C}$, we have the equality of morphisms in $\mathcal{D}$:

$$F(\mathrm{id}_X) = \mathrm{id}_{F(X)}.$$

Subsequently, when defining a functor $F : \mathcal{C} \to \mathcal{D}$ we shall specify the functions on objects and morphisms at the same time as follows:

$$\begin{array}{rcl} \mathcal{C} & \longrightarrow & \mathcal{D} \\ X & \longmapsto & F(X) \\ \left(X \overset{f}{\to} Y\right) & \longmapsto & \left(F(X) \overset{F(f)}{\to} F(Y)\right). \end{array}$$

Once we specify the two functions, we need to verify it is composition and identities preserving.

Example 8.3.2. Let $\mathcal{C}$ be a category. Define the identity functor $\mathrm{id}_{\mathcal{C}} : \mathcal{C} \to \mathcal{C}$ as:

$$\begin{array}{rcl} \mathrm{id}_{\mathcal{C}} : \mathcal{C} & \longrightarrow & \mathcal{C} \\ X & \longmapsto & X \\ \left(X \overset{f}{\to} Y\right) & \longmapsto & \left(X \overset{f}{\to} Y\right). \end{array}$$

We see immediately that $\mathrm{id}_{\mathcal{C}}(g \circ f) = g \circ f = \mathrm{id}_{\mathcal{C}}(g) \circ \mathrm{id}_{\mathcal{C}}(f)$, for any composable morphisms f and g, and $\mathrm{id}_{\mathcal{C}}(\mathrm{id}_X) = \mathrm{id}_X = \mathrm{id}_{\mathrm{id}_{\mathcal{C}}(X)}$, for all object X in $\mathcal{C}$.

Example 8.3.3. Recall from Example 8.1.28 that a monoid can be regarded as a category with one object. Let $f : M \to N$ be a homomorphism of monoids. This defines a functor $Bf : BM \to BN$:

$$\begin{array}{rcl} Bf : BM & \longrightarrow & BN \\ \star & \longmapsto & \star \\ \left(\star \xrightarrow{m} \star\right) & \longmapsto & \left(\star \xrightarrow{f(m)} \star\right). \end{array}$$

Let us verify it preserves composition and identities. Given $m, n \in M$, since f is a homomorphism, we have $f(mn) = f(m)f(n)$. This precisely translates to $Bf(m \circ n) = Bf(m) \circ Bf(n)$. Recall that $\mathrm{id}_\star$ in BM is the neutral element e_M of M. Since f is a homomorphism, then $f(e_M) = e_N$, the neutral element of N. Thus $Bf(\mathrm{id}_\star) = \mathrm{id}_\star$. Therefore, Bf is indeed a functor.

Exercise 8.3.4. Suppose M and N are monoids and let $F : BM \to BN$ be any functor. Show that there is a unique homomorphism of monoids $f : M \to N$ such that $Bf = F$. Conclude there is a correspondence between homomorphisms of monoids and functors between categories with one object.

Example 8.3.5. If G and H are groups, then a group homomorphism $f : G \to H$ defines a functor $Bf : BG \to BH$. Moreover, any functor $F : BG \to BH$ defines a group homomorphisms $f : G \to H$ such that $F = Bf$.

Example 8.3.6. A mathematical construction can be viewed as functorial in several ways. Let R be a ring. Recall that we can define the polynomial ring $R[x_1, \ldots, x_n]$ with n-variables. This is functorial if we vary the ring: i.e. there is a functor:

$$\begin{array}{rcl} \mathsf{Ring} & \longrightarrow & \mathsf{Ring} \\ R & \longmapsto & R[x_1, \ldots, x_n] \\ \left(R \xrightarrow{f} S\right) & \longmapsto & \left(\begin{array}{c} R[x_1, \ldots, x_n] \xrightarrow{f} S[x_1, \ldots, x_n] \\ \displaystyle\sum_{i,j \geq 0}^{n} a_{ij} x_i^j \mapsto \sum_{i,j \geq 0}^{n} f(a_{ij}) x_i^j \end{array}\right) \end{array}$$

On the other hand, we could have also varied the amount of variables. If we denote Fin the full subcategory of Set spanned by finite sets, then there is a functor:

$$\begin{array}{rcl} \mathsf{Fin} & \longrightarrow & \mathsf{Ring} \\ \{1, \ldots, n\} & \longmapsto & R[x_1, \ldots, x_n] \\ \left(\{1, \ldots, n\} \xrightarrow{\sigma} \{1, \ldots, m\}\right) & \longmapsto & \left(\begin{array}{c} R[x_1, \ldots, x_n] \xrightarrow{\sigma} R[x_1, \ldots, x_m] \\ \displaystyle\sum_{i,j \geq 0}^{n} a_{ij} x_i^j \mapsto \sum_{i,j \geq 0}^{n} a_{\sigma(i)j} x_{\sigma(i)}^j \end{array}\right) \end{array}$$

We can record this fact by capturing the two variances into one functor:

$$\begin{aligned} \mathsf{Ring} \times \mathsf{Fin} &\longrightarrow \mathsf{Ring} \\ (R, \{1, \dots, n\}) &\longmapsto R[x_1, \dots, x_n] \\ (f, \sigma) &\longmapsto \begin{pmatrix} R[x_1, \dots, x_n] \overset{(f,\sigma)}{\to} S[x_1, \dots, x_m] \\ \sum_{i,j \geq 0}^{n} a_{ij} x_i^j \mapsto \sum_{i,j \geq 0}^{n} f(a_{\sigma(i)j}) x_{\sigma(i)}^j \end{pmatrix} \end{aligned}$$

Exercise 8.3.7. Let R be a ring. Recall that for a (possibly infinite) set, we can define $R[X]$. Show this leads to functors $\mathsf{Set} \to \mathsf{Ring}$ and $\mathsf{Ring} \times \mathsf{Set} \to \mathsf{Ring}$.

Exercise 8.3.8. Let $\mathcal{C}$, $\mathcal{D}$ and $\mathcal{E}$ be categories. Show that the data of a functor $\mathcal{E} \to \mathcal{C} \times \mathcal{D}$ is equivalent to the data of two functors $\mathcal{E} \to \mathcal{C}$ and $\mathcal{E} \to \mathcal{D}$.

Example 8.3.9. Recall that for a ring R we denote by $R^\times$ its group of units. If $f : R \to S$ is a ring homomorphism, then if $r \in R$ is a unit, then so is $f(r)$. Thus we can restrict and corestrict $f^\times : R^\times \to S^\times$. This defines a functor:

$$\begin{aligned} (-)^\times : \mathsf{Ring} &\longrightarrow \mathsf{Grp} \\ R &\longmapsto R^\times \\ \left(R \overset{f}{\to} S\right) &\longmapsto \left(R^\times \overset{f^\times}{\to} S^\times\right) \end{aligned}$$

Example 8.3.10. A lot of mathematical structures are sets with additional structures. For instance, a monoid $(M, *, e_M)$ is a set M with the extra data of a multiplication and unity. Moreover, any monoid homomorphism $M \to N$ is a set map with the extra requirement that it must preserves identity and multiplication. Forgetting this data defines a functor:

$$\begin{aligned} \mathsf{Mon} &\longrightarrow \mathsf{Set} \\ (M, *, e_M) &\longmapsto M \\ \left((M, *, e_m) \overset{f}{\to} (N, \odot, e_N)\right) &\longmapsto \left(M \overset{f}{\to} N\right) \end{aligned}$$

Such functor is called a *forgetful functor*, or *underlying functor*, and is often denoted U. It occurs in many instances: a ring is an Abelian group with a multiplication, an Abelian group is a group in which the multiplication is commutative, a group is a monoid for which every element has an inverse, and we just saw that a monoid is a set with extra structure. So we have all these forgetful functors:

$$\mathsf{Ring} \to \mathsf{Ab} \to \mathsf{Grp} \to \mathsf{Mon} \to \mathsf{Set}.$$

Generally they are all denoted U and there is usually no ambiguity. In practice, we often omit U. For instance we might prefer to say "the set M", or "M regarded as a set" instead of writing $U(M)$, for M a monoid and $U : \mathsf{Mon} \to \mathsf{Set}$.

Example 8.3.11. Given a monoid $(M, *, e_M)$ we can forget the multiplication but still keep track of the choice of unity. The assignment $(M, *, e_M) \to (M, e_M)$ defines a forgetful functor $\mathsf{Mon} \to \mathsf{Set}_*$.

Example 8.3.12. Given a set X, one can define the free group on X, denoted $F(X)$ formed of all possible words in X together with concatenation as multiplication and the empty word as unity. The definition can be extended to map of sets and thus we obtain a functor $F : \mathsf{Set} \to \mathsf{Grp}$. There exists a nice connection between F and

the forgetful U from previous example: given any set X and group G, there is an isomorphism of sets (i.e. bijection):

$$\mathrm{Hom}_{\mathsf{Grp}}(F(X), G) \cong \mathrm{Hom}_{\mathsf{Set}}(X, U(G)).$$

This is precisely a reformulation of the universal property of free groups. In fact, given a forgetful functor $U : \mathcal{C} \to \mathsf{Set}$, the free functor $F : \mathsf{Set} \to \mathcal{C}$ will be defined as the unique functor such that we obtain a "univesal" isomorphism:

$$\mathrm{Hom}_{\mathcal{C}}(F(X), C) \cong \mathrm{Hom}_{\mathsf{Set}}(X, U(C)),$$

given any set X and object $C \in \mathcal{C}$. We will make this idea precise in the next sections, it is important to notice now that there seems to be a general pattern with free objects and their universal properties.

Exercise 8.3.13. Let $U : \mathsf{Set}_* \to \mathsf{Set}$ be the forgetful functor on pointed sets. Given a set X, denote $X_+ = (X \coprod \{*\}, *)$ the set where we have added a point $*$ to X and chose it as a basepoint. Given a set map $f : X \to Y$, extend it to a set map $f_+ : X_+ \to Y_+$ by $f(*) = *$. Show this defines a functor $(-)_+ : \mathsf{Set} \to \mathsf{Set}_*$ such that we obtain an isomorphism of sets (i.e. bijection):

$$\mathrm{Hom}_{\mathsf{Set}_*}(X_+, (Y, y_0)) \cong \mathrm{Hom}_{\mathsf{Set}}(X, Y),$$

for any set X and pointed set (Y, y_0), where we denoted $Y = U(Y, y_0)$.

Example 8.3.14. Recall that for any set S, there are unique set maps $\emptyset \to S$ and $S \to \{*\}$. In a similar fashion, given any category $\mathcal{C}$, there are unique functors $\mathbb{0} \to \mathcal{C}$ and $\mathcal{C} \to \mathbb{1}$.

Example 8.3.15. Recall that choosing an element x in a set S defines a map $x : \{*\} \to S$. Similarly, choosing an object X in a category $\mathcal{C}$ amounts precisely to a functor $X : \mathbb{1} \to \mathcal{C}$.

Exercise 8.3.16. Show that choosing a morphism $f : X \to Y$ in $\mathcal{C}$ amounts precisely to a functor $f : \mathbb{2} \to \mathcal{C}$. Show that choosing an isomorphism $f : X \to Y$ in $\mathcal{C}$ amounts precisely to a functor $f : \mathbb{I} \to \mathcal{C}$.

One key importance of functors is that they detect non-isomorphic objects. Therefore given a functor $F : \mathcal{C} \to \mathcal{D}$, if you want to determine if objects in $\mathcal{C}$ are not isomorphic, you can use the functor F to travel in a different realm.

Theorem 8.3.17. Let $\mathcal{C}$ and $\mathcal{D}$ be categories. Let $F : \mathcal{C} \to \mathcal{D}$ be a functor. Suppose $f : X \to Y$ is an isomorphism in $\mathcal{C}$, then $F(f) : F(X) \to F(Y)$ is an isomorphism in $\mathcal{D}$ and $F(f)^{-1} = F(f^{-1})$. In particular, a functor *preserves* isomorphisms: if $X \cong Y$ in $\mathcal{C}$, then $F(X) \cong F(Y)$ in $\mathcal{D}$.

Proof. Since $f : X \to Y$ is an isomorphism in $\mathcal{C}$, there exists $f^{-1} : Y \to X$ in $\mathcal{C}$ such that $f^{-1} \circ f = \mathrm{id}_X$ and $f \circ f^{-1} = \mathrm{id}_Y$. Therefore:

$$\begin{aligned} F(f^{-1}) \circ F(f) &= F(f^{-1} \circ f), \text{ by composition preserving,} \\ &= F(\mathrm{id}_X) \\ &= \mathrm{id}_{F(X)}, \text{ by identities preserving.} \end{aligned}$$

Similarly, we obtain $F(f) \circ F(f^{-1}) = \mathrm{id}_{F(Y)}$. Thus $F(f) : F(X) \to F(Y)$ is an isomorphism in $\mathcal{D}$ with inverse $F(f^{-1}) : F(Y) \to F(X)$. □

Corollary 8.3.18. Let $\mathcal{C}$ and $\mathcal{D}$ be categories. Let $F : \mathcal{C} \to \mathcal{D}$ be functors. Let X and Y be objects in $\mathcal{C}$. If $F(X) \ncong F(Y)$ in $\mathcal{D}$, then $X \ncong Y$ in $\mathcal{C}$.

Example 8.3.19. Two groups cannot be isomorphic if their cardinals were not equal. This follows from the forgetful functor $\mathsf{Ring} \to \mathsf{Set}$.

Example 8.3.20. The ring $\mathbb{Z}$ and $\mathbb{Q}$ cannot be isomorphic (even though they have the same cardinality) because $\mathbb{Z}^\times \cong C_2 \ncong \mathbb{Q} - \{0\} = \mathbb{Q}^\times$.

⚠ **Warning 8.3.21.** In general, given a functor $F : \mathcal{C} \to \mathcal{D}$, if $F(X) \cong F(Y)$ in $\mathcal{D}$, there is no reason to expect that $X \cong Y$ in $\mathcal{C}$.

Example 8.3.22. Given any category $\mathcal{C}$, the unique functor $F : \mathcal{C} \to \mathbb{1}$ forces that $F(X) \cong F(Y)$ for any two object X and Y in $\mathcal{C}$.

Example 8.3.23. Consider the functor $(-)^\times : \mathsf{Ring} \to \mathsf{Grp}$ and the rings $\mathbb{Z}$ and $\mathbb{Z}/3\mathbb{Z}$. They are not isomorphic but $\mathbb{Z}^\times \cong C_2 \cong (\mathbb{Z}/3\mathbb{Z})^\times$.

Definition 8.3.24. A functor $F : \mathcal{C} \to \mathcal{D}$ is said to be *conservative* if it *reflects* isomorphisms: if $f : X \to Y$ is in $\mathcal{C}$ such that $F(f) : F(X) \to F(Y)$ is an isomorphism in $\mathcal{D}$, then f is an isomorphism in $\mathcal{C}$.

Example 8.3.25. The forgetful functor $\mathsf{Grp} \to \mathsf{Set}$ is conservative: a homomorphism of groups that is a bijection is an isomorphism. Many of the forgetful functors we have seen are conservative. However, not all forgetful functors are conservative (example in topology).

Example 8.3.26. Even though the forgetful functor $\mathsf{Ring} \to \mathsf{Grp}$ is conservative, the functor $(-)^\times : \mathsf{Ring} \to \mathsf{Grp}$ is not. For instance, consider the quotient map $\gamma : \mathbb{Z} \to \mathbb{Z}/3\mathbb{Z}$. Then $\gamma^\times : \mathbb{Z}^\times \to (\mathbb{Z}/3\mathbb{Z})^\times$ is an isomorphism but γ is not.

Example 8.3.27. Given a functor $F : \mathcal{C} \to \mathcal{D}$ and a functor $G : \mathcal{D} \to \mathcal{E}$, then just as we can compose functions or more generally morphisms in a category, we can compose functors and define $G \circ F : \mathcal{C} \to \mathcal{E}$ to be defined as follows:

$$\begin{array}{ccc} \mathcal{C} & \longrightarrow & \mathcal{E} \\ C & \longmapsto & G(F(C)) \\ \left(C \xrightarrow{f} C'\right) & \longmapsto & \left(G(F(C)) \xrightarrow{G(F(f))} G(F(C'))\right). \end{array}$$

Exercise 8.3.28. Verify that $G \circ F$ of Example 8.3.27 is indeed a functor.

It is tempting to consider a category of categories in which objects are categories and morphisms are functors. Unfortunately we encounter size issues in the same way that there cannot be a set of all sets. To palliate this issue we can consider a large category of all categories. A large category is not a category with our terminology as its hom can be a class.

Definition 8.3.29. Define CAT the large category of all categories as follows.

- **(i)** Its objects are categories.
- **(ii)** Given categories $\mathcal{C}$ and $\mathcal{D}$, the class $\mathrm{Hom}_{\mathsf{CAT}}(\mathcal{C}, \mathcal{D})$ comprises of all functors $\mathcal{C} \to \mathcal{D}$.
- **(iii)** Composition of functors is defined as in Example 8.3.27.

One can check this structure gives indeed a large category, in which the identities are given as in Example 8.3.2.

We shall see soon that the class $\mathrm{Hom}_{\mathsf{CAT}}(\mathcal{C}, \mathcal{D})$ can itself be endowed with a category structure that we will denote $\mathrm{Fun}(\mathcal{C}, \mathcal{D})$.

To avoid size issues, we may consider instead a category of small categories.

Definition 8.3.30. Define Cat as the full subcategory of CAT spanned by small categories. In this instance, the morphisms assemble to a set and not a class.

Example 8.3.31. We have seen that for any monoid M we can associate a category BM with one object, this defines a functor $B : \mathsf{Mon} \to \mathsf{Cat}$.

Exercise 8.3.32. Define Grpd as the full subcategory of Cat spanned by groupoids. Show we obtain a functor $B : \mathsf{Grp} \to \mathsf{Grpd}$.

8.4. Embedding categories

Definition 8.4.1. A functor $F : \mathcal{C} \to \mathcal{D}$ is said to be *faithful* if for all objects X and Y in $\mathcal{C}$, the set map $F : \mathrm{Hom}_{\mathcal{C}}(X,Y) \to \mathrm{Hom}_{\mathcal{D}}(F(X),F(Y))$ is injective. In other words, if $f : X \to Y$ and $g : X \to Y$ are maps in $\mathcal{C}$, then if $F(f) = F(g)$ as maps $F(X) \to F(Y)$ in $\mathcal{D}$, then $f = g$.

Example 8.4.2. Typically, forgetful functors are faithful functors. For instance two homomorphisms of groups are equal if they are equal as set maps.

Example 8.4.3. Given a category $\mathcal{C}$, there exists a unique functor $\mathcal{C} \to \mathbb{1}$. This functor is never faithful unless $\mathcal{C}$ is $\mathbb{0}$ or $\mathbb{1}$.

Example 8.4.4. Let $f : M \to N$ be a homomorphism of monoids. Then the induced functor $Bf : BM \to BN$ is faithful if and only if f is injective.

Definition 8.4.5. A *concrete category* $\mathcal{C}$ is a category together with a faithful functor $U : \mathcal{C} \to \mathsf{Set}$.

Example 8.4.6. The categories Mon, Grp, Ab, Set_*, Ring etc are all concrete categories when considering their forgetful functors onto Set.

Exercise 8.4.7. Suppose $F : \mathcal{C} \to \mathcal{D}$ is a faithful functor. Show that if a diagram in $\mathcal{C}$ commutes in $\mathcal{D}$ after applying F, then it commutes in $\mathcal{C}$.

Definition 8.4.8. A functor $F : \mathcal{C} \to \mathcal{D}$ is said to be *full* if for all objects X and Y in $\mathcal{C}$ the set map $F : \mathrm{Hom}_{\mathcal{C}}(X,Y) \to \mathrm{Hom}_{\mathcal{D}}((F(X),F(Y))$ is surjective. In other words, given a map $g : F(X) \to F(Y)$ in $\mathcal{D}$, there exists a map $f : X \to Y$ in $\mathcal{C}$ such that $F(f) = g$.

Example 8.4.9. Let $f : M \to N$ be a homomorphism of monoids. Then the induced functor $Bf : BM \to BN$ is full if and only if f is surjective.

Definition 8.4.10. A functor $F : \mathcal{C} \to \mathcal{D}$ is said to be:

- *essentially injective* if: given objects $X, Y \in \mathcal{C}$, if $F(X) \cong F(Y)$ in $\mathcal{D}$, then $X \cong Y$ in $\mathcal{C}$;
- *essentially surjective* if: for all $D \in \mathcal{D}$, there exists $C \in \mathcal{C}$ such that $F(C) \cong D$ in $\mathcal{D}$.

⚠ **Warning 8.4.11.** It is important to not confuse essentially injective with conservative (Definition 8.3.24). A functor can be essentially injective without being conservative.

Example 8.4.12. Example of a non essentially injective but conservative functor. Example of a conservative functor but non essentially injective.

Proposition 8.4.13. Let $F : \mathcal{C} \to \mathcal{D}$ be a conservative and full functor. Then F is essentially injective.

Proof. Let X and Y be objects in $\mathcal{C}$ and suppose $F(X) \cong F(Y)$ in $\mathcal{D}$. This means there exists a map $g : F(X) \to F(Y)$ in $\mathcal{D}$ that is an isomorphism. Since F is full, there exists a map $f : X \to Y$ in $\mathcal{C}$ such that $F(f) = g$. Since F is conservative, then f must be an isomorphism. Thus $X \cong Y$ in $\mathcal{C}$. □

Example 8.4.14. Choosing an isomorphism $f : X \to Y$ in a category $\mathcal{C}$ defines a functor $f : \mathbb{I} \to \mathcal{C}$ that is conservative and essentially injective but neither full nor faithful in general.

Definition 8.4.15. An *embedding of categories* is a functor $F : \mathcal{C} \to \mathcal{D}$ that is essentially injective and faithful. In this case we say $\mathcal{C}$ is *embedded* in $\mathcal{D}$. If the functor $F : \mathcal{C} \to \mathcal{D}$ is also full, then we say F is a *full embedding of categories* and $\mathcal{C}$ is *fully embedded* in $\mathcal{D}$.

Proposition 8.4.16. Given an embedding of categories $F : \mathcal{C} \to \mathcal{D}$ and objects X, Y in $\mathcal{C}$. Then $X \cong Y$ in $\mathcal{C}$ if and only if $F(X) \cong F(Y)$ in $\mathcal{D}$.

In a non-full embedding of categories $\mathcal{C}$ in $\mathcal{D}$ it is possible that one added more morphisms between objects in $\mathcal{D}$.

Example 8.4.17. A subcategory $\mathcal{C}$ of a category $\mathcal{D}$ defines an embedding of categories $\mathcal{C} \to \mathcal{D}$. A full subcategory $\mathcal{C}$ of $\mathcal{D}$ defines a full embedding of categories. Informally, a (full) embedding of categories $F : \mathcal{C} \to \mathcal{D}$ defines a (full) subcategory in $\mathcal{D}$ comprised of the *essential image* of F: the smallest subcategory of $\mathcal{D}$ containing all objects in $\mathcal{D}$ that are isomorphic to $F(X)$ for some $X \in \mathcal{C}$.

Definition 8.4.18. A functor $F : \mathcal{C} \to \mathcal{D}$ that is both full and faithful is called *fully faithful*.

Proposition 8.4.19. A fully faithful functor is conservative, and thus in particular essentially injective.

Proof. Let $F : \mathcal{C} \to \mathcal{D}$ be a fully faithful functor. Let $f : X \to Y$ be a map in $\mathcal{C}$. Suppose $F(f) : F(X) \to F(Y)$ is an isomorphism in $\mathcal{D}$. This means there exists $g : F(Y) \to F(X)$ in $\mathcal{D}$ such that $g \circ F(f) = \mathrm{id}_{F(X)}$ and $F(f) \circ g = \mathrm{id}_{F(Y)}$. Since F is full, there exists $h : Y \to X$ in $\mathcal{C}$ such that $F(h) = g$. We obtain the equalities:

$$F(h \circ f) = F(h) \circ F(f) = g \circ F(f) = \mathrm{id}_{F(X)} = F(\mathrm{id}_X).$$

As F is faithful, we obtain $h \circ f = \mathrm{id}_X$. We argue similarly to show $f \circ h = \mathrm{id}_Y$. Thus f is an isomorphism in $\mathcal{C}$ with inverse h. Thus F is conservative. □

Corollary 8.4.20. A functor is fully faithful if and only if it is a full embedding of categories.

8.5. Equivalences of categories and natural transformations

In previous section we saw that we can define a morphism between categories called functors. Omitting size issues, categories together with their functors assemble themselves into a category CAT. Therefore this defines a notion of an isomorphism of categories as in Definition 8.1.4. Two categories $\mathcal{C}$ and $\mathcal{D}$ are *isomorphic*, and we write $\mathcal{C} \cong \mathcal{D}$, if there exist functors $F : \mathcal{C} \to \mathcal{D}$ and $G : \mathcal{D} \to \mathcal{C}$ such that $G \circ F = \mathrm{id}_{\mathcal{C}}$ and $F \circ G = \mathrm{id}_{\mathcal{D}}$.

However, the definition is often extremely rigid. Indeed, we require that for any object X in $\mathcal{C}$ that $G(F(X))$ is *equal* to X, and given any morphism $f : X \to Y$, we have $G(F(f))$ is equal to f. But in practice, it is more likely that $G(F(X))$ is only isomorphic to X. In that case, we cannot have $G(F(f)) = f$ as they don't have same domains and codomains, but we can ask that they remain compatible. In other words, we would like the following diagram to commute:

$$\begin{array}{ccc} G(F(X)) & \xrightarrow{\cong} & X \\ {\scriptstyle GF(f)}\downarrow & & \downarrow{\scriptstyle f} \\ G(F(Y)) & \xrightarrow{\cong} & Y, \end{array}$$

for any morphism $f : X \to Y$ in $\mathcal{C}$. If the diagrams above was not commutative then if means we have lost information on the morphisms in $\mathcal{C}$.

We make precise the notion here. We begin by two motivating examples.

Example 8.5.1. Let $\mathbb{F}$ be a field. Our first example is motivated from the fact that every finite dimensional $\mathbb{F}$-vector space is isomorphic (but not equal) to $\mathbb{F}^n$ for some $n \geq 0$. First recall we have defined the category $\mathsf{Mat}(\mathbb{F})$ of matrices with coefficients in $\mathbb{F}$. Denote $\mathsf{Vect}_{\mathbb{F}}^{\mathsf{fd}}$ the category for which objects are pairs $(V, \mathcal{B})$ where V is a finite dimensional $\mathbb{F}$-vector space and $\mathcal{B}$ is an ordered (finite) basis of V. Morphisms $(V, \mathcal{B}) \to (V', \mathcal{B}')$ are linear transformations $V \to V'$ with no added requirements. Therefore $\mathsf{Vect}_{\mathbb{F}}^{\mathsf{fd}}$ is a full subcategory of $\mathsf{Vect}_{\mathbb{F}}$. Recall that a matrix $A \in \mathcal{M}_{m\times n}(\mathbb{F})$ can be regarded as a linear transformation:

$$\begin{array}{rcl} A : (\mathbb{F}^n, \mathcal{S}) & \longrightarrow & (\mathbb{F}^m, \mathcal{S}) \\ v & \longmapsto & Av. \end{array}$$

Here we denote $\mathcal{S}$ the standard basis of $\mathbb{F}^n$. This perspective defines a functor:

$$\begin{array}{rcl} \mathsf{Mat}(\mathbb{F}) & \longrightarrow & \mathsf{Vect}_{\mathbb{F}}^{\mathsf{fd}} \\ n & \longmapsto & (\mathbb{F}^n, \mathcal{S}) \\ A \in \mathcal{M}_{m\times n}(\mathbb{F}) & \longmapsto & \Big(A : (\mathbb{F}^n, \mathcal{S}) \to (\mathbb{F}^m, \mathcal{S})\Big). \end{array}$$

Let $\mathcal{B} = (b_1, \ldots, b_n)$ be a basis of a vector space V. Then for any $v \in V$, there exist unique scalars $\lambda_1, \ldots, \lambda_n \in \mathbb{F}^n$ such that $v = \lambda_1 b_1 + \cdots \lambda_n b_n$. This defines the vector $[v]_{\mathcal{B}} = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix}$ in $\mathbb{F}^n$. Given a linear transformation $T : (V, \mathcal{B}) \to (V', \mathcal{B}')$, we denote the matrix $[T]_{\mathcal{B}}^{\mathcal{B}'}$ comprised of the vectors $[T(b_1)]_{\mathcal{B}'}, \ldots, [T(b_n)]_{\mathcal{B}'}$ in its columns.

We obtain a functor:

$$\begin{array}{rcl} \mathsf{Vect}_{\mathbb{F}}^{\mathsf{fd}} & \longrightarrow & \mathsf{Mat}(\mathbb{F}) \\ (V,\mathcal{B}) & \longmapsto & \dim(V) \\ \Big((V,\mathcal{B}) \xrightarrow{T} (V',\mathcal{B}')\Big) & \longmapsto & [T]_{\mathcal{B}}^{\mathcal{B}'} \end{array}$$

One one hand the composition:

$$n \longmapsto (\mathbb{F}^n, \mathcal{S}) \longmapsto \dim(\mathbb{F}^n)$$

is the identity on objects and on morphisms. On the other hand, the composition:

$$(V,\mathcal{B}) \longmapsto \dim(V) \longmapsto (\mathbb{F}^{\dim(V)}, \mathcal{S})$$

is in general only an isomorphism given by:

$$\begin{array}{rcl} (V,\mathcal{B}) & \xrightarrow{\cong} & (\mathbb{F}^{\dim(V)}, \mathcal{S}) \\ v & \longmapsto & [v]_{\mathcal{B}}. \end{array}$$

Notice that these isomorphisms are compatible: given any linear transformation $T : (V,\mathcal{B}) \to (V',\mathcal{B}')$, we have the commutative diagram:

$$\begin{array}{ccc} (V,\mathcal{B}) & \xrightarrow{\cong} & (\mathbb{F}^{\dim(V)}, \mathcal{S}) \\ {\scriptstyle T}\downarrow & & \downarrow{\scriptstyle [T]_{\mathcal{B}}^{\mathcal{B}'}} \\ (V',\mathcal{B}') & \xrightarrow{\cong} & (\mathbb{F}^{\dim(V')}, \mathcal{S}) \end{array}$$

This expresses the familiar equation $[T(v)]_{\mathcal{B}'} = [T]_{\mathcal{B}}^{\mathcal{B}'}[v]_{\mathcal{B}}$. Therefore, although $\mathsf{Vect}_{\mathbb{F}}^{\mathsf{fd}}$ and $\mathsf{Mat}(\mathbb{F})$ are not isomorphic, they seem to be equivalent in many regards.

Example 8.5.2. From our perspective, a ring is always unital. We shall explain why here.

A ring R is said to be be augmented if there is a ring homomorphism $\varepsilon : R \to \mathbb{Z}$. In fact we can define the category of augmented rings as $\mathsf{Ring}_{/\mathbb{Z}}$. A non-unital ring R is a ring without the axiom of unity, and a non-unital ring homomorphism is a ring that doesn't preserve unity. This defines a category $\mathsf{Ring}_{\circ}$.

Given a non-unital ring R, let $R_+ = R \oplus \mathbb{Z}$. Then one can check that R_+ is a (unital) ring with unity $(0_R, 1)$ and is augmented via the quotient map $R_+ \to R_+/R \cong \mathbb{Z}$. Define a functor:

$$\begin{array}{c} (-)_+ : \mathsf{Ring}_{\circ} \longrightarrow \mathsf{Ring}_{/\mathbb{Z}} \\ R \to R_+ \\ \Big(R \xrightarrow{f} S\Big) \longmapsto \Big(R_+ \xrightarrow{f\oplus \mathrm{id}} S_+\Big). \end{array}$$

Given an augmented ring (R,ε), we can denote $R_- = \ker(\varepsilon)$, it is a non-unital ring.

$$\begin{array}{c} (-)_- : \mathsf{Ring}_{/Z} \longrightarrow \mathsf{Ring}_{\circ} \\ (R,\varepsilon) \longmapsto R_- \\ \Big(R \to S\Big) \longmapsto \Big(R_- \to S_-\Big) \end{array}$$

Suppose (R, ε) is an augmented ring. Then $((R,\varepsilon)_-)_+ = \ker(\varepsilon) \oplus \mathbb{Z}$. Since we have an isomorphism:

$$\begin{aligned} \ker(\varepsilon) \oplus \mathbb{Z} &\longrightarrow R \\ (r, 1) &\longmapsto r + 1_R \end{aligned}$$

Similarly, given a non unital ring R, then $(R_+)_- = \ker(R \oplus \mathbb{Z} \to \mathbb{Z}) \cong R$. These are not equal but really isomorphisms. One can view the category of non-unital rings $\mathsf{Ring}_\circ$ as living inside the category of unital rings Ring, as $\mathsf{Ring}_{/\mathbb{Z}}$ is a (non full) subcategory of Ring.

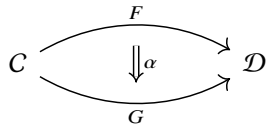
The above example hopefully served as motivation for the need of this coherence data on morphism.

Definition 8.5.3. Let $F : \mathcal{C} \to \mathcal{D}$ and $G : \mathcal{C} \to \mathcal{D}$ be functors. A *natural transformation* from F to G, denoted $\alpha : F \Rightarrow G$ is a collection of morphisms $\{\alpha_X : F(X) \to G(X) \mid X \in \mathcal{C}\}$ in $\mathcal{D}$ subject to the following requirement. For any morphism $f : X \to Y$ in $\mathcal{C}$, the following diagram commutes in $\mathcal{D}$:

$$\begin{array}{ccc} F(X) & \xrightarrow{\alpha_X} & G(X) \\ {\scriptstyle F(f)}\downarrow & & \downarrow{\scriptstyle G(f)} \\ F(Y) & \xrightarrow[\alpha_Y]{} & G(Y) \end{array}$$

If $\alpha_X : F(X) \to G(X)$ is an isomorphism in $\mathcal{D}$ for each $X \in \mathcal{C}$, then we say α is a *natural isomorphism* and we write it as $\alpha : F \tilde{\Rightarrow} G$.

Notation 8.5.4. Instead of saying "let $\alpha : F \Rightarrow G$ be a natural transformation between functors $F, G : \mathcal{C} \to \mathcal{D}$", we may compactly refer to the data as a diagram:



Bibliography

[Stacks] The Stacks Project Authors. *Stacks Project*. https://stacks.math.columbia.edu. 2018.